Published in final edited form as:
J Neurosci. 2011 April 6; 31(14): 5526–5539. doi:10.1523/JNEUROSCI.4647-10.2011.

Neural correlates of forward planning in a spatial decision task in humans

Dylan Alexander Simon[1] and **Nathaniel D. Daw**[1,2]

[1] Department of Psychology, New York University, 6 Washington Pl., New York, NY 10003

[2] Center for Neural Science, New York University, 6 Washington Pl., New York, NY 10003

Abstract

Although reinforcement learning (RL) theories have been influential in characterizing the brain's mechanisms for reward-guided choice, the predominant temporal difference (TD) algorithm cannot explain many flexible or goal-directed actions that have been demonstrated behaviorally. We investigate such actions by contrasting an RL algorithm that is model-based, in that it relies on learning a map or model of the task and planning within it, to traditional model-free TD learning. To distinguish these approaches in humans, we used fMRI in a continuous spatial navigation task, in which frequent changes to the layout of the maze forced subjects continually to relearn their favored routes, thereby exposing the RL mechanisms employed. We sought evidence for the neural substrates of such mechanisms by comparing choice behavior and BOLD signals to decision variables extracted from simulations of either algorithm. Both choices and value-related BOLD signals in striatum, though most often associated with TD learning, were better explained by the model-based theory. Further, predecessor quantities for the model-based value computation were correlated with BOLD signals in the medial temporal lobe and frontal cortex. These results point to a significant extension of both the computational and anatomical substrates for RL in the brain.

Keywords

learning; spatial; reinforcement; fMRI; striatum; decision

2 Introduction

Employing past experience to guide future decisions is critical for survival, but a longstanding question is how the brain represents this experience. A predominant theory is temporal difference (TD) reinforcement learning (RL), which learns from reinforcement the future reward value expected following an action (Sutton, 1988; Sutton and Barto, 1998). Much evidence links such learning to spiking and BOLD signals in the nigrostriatal dopamine system (Houk et al., 1994; Schultz et al., 1997; Berns et al., 2001; O'Doherty et al., 2002; Pagnoni et al., 2002).

However, such mechanisms, which rely on repeating successful actions (Thorndike, 1911), cannot explain flexible or novel action planning seen in tasks such as latent learning or reinforcer devaluation (Tolman, 1948; Balleine and Dickinson, 1998). There are many suggestions of such sophistication across species (Maguire et al., 1998; Hampton et al., 2006; Pan et al., 2007), notably lesion results in rodent conditioning (Balleine et al., 2008)

Correspondence should be addressed to Dylan A. Simon, dylex@nyu.edu.

and navigation (Packard and McGaugh, 1996) suggesting that it might coexist in the brain with simpler reinforcement mechanisms. Such behaviors are envisioned to arise from considering the future consequences of an action, drawing on a learned cognitive map or model of the environment (Thistlethwaite, 1951; Gallistel and Cramer, 1996). One candidate computational formalization of these processes is model-based RL (Doya, 1999; Daw et al., 2005; Johnson et al., 2007), which constructs the values of possible action trajectories indirectly by simulating a learned model of the environment. This planning process contrasts with model-free TD algorithms, which learn future values directly.

However, while there has been much work quantitatively investigating TD characterizations of learning (O'Doherty et al., 2003; Seymour et al., 2004; O'Doherty et al., 2006; Lee et al., 2004), much less research has analogously investigated the neural and computational substrates for model-based learning and planning. One promising domain for such an investigation is spatial navigation, which sparked early cognitive map work (Tolman, 1948) and in which a distinction has been made between deliberate "place" learning and habitual "response" behaviors (Blodgett and McCutchan, 1947) that may parallel the model-based vs. TD distinction.

We thus used fMRI to investigate the neural substrates for model-based learning and planning in humans navigating a virtual maze for money. This task had two key features that we expected would encourage a model-based strategy: first, basic structure of a spatial model is known a priori and need not have been learned; second, ongoing reconfiguration of the maze promoted continuous learning and on-line planning of new routes (Daw et al., 2005). These dynamic reconfigurations also generated discrepancies between hypothesized model-based and model-free update mechanisms, allowing us to distinguish these strategies over many trials and verify our hypothesis that behavior and value-related BOLD signals were driven by model-based rather than TD mechanisms. Having done so, we employed this computational characterization of the learning to begin to map the network supporting model-based values, much as has been done for TD, by seeking neural correlates of learning about the more elementary quantities from which model-based values are constructed.

3 Methods

Participants

Eighteen healthy, right-handed adults (10 female), 18 to 36 years of age performed the task for payment while undergoing functional magnetic resonance imaging. All participants gave informed consent and the study was approved by the New York University Committee on Activities Involving Human Subjects.

Task

Subjects navigated a virtual 4 × 4 grid of rooms (designated as *states* $s \in \mathcal{S}$) by making choices between the available rooms adjoining the current location (Fig. 1A). Subjects continuously viewed rendered images of a 3D representation of these rooms with a first-person perspective from their current position. The display included boundary cues and distal direction cues so that subjects could identify their position within the grid, as well as any rooms ahead of them (within a 100° viewing angle). Each of the 24 pairs of adjoining rooms was connected by a one-way door, which at any time was available for use in exactly one direction between the rooms.

At each room, subjects chose between the available doors by pressing one of three keys with their right hand so as either to move forward or to turn 90 degrees and move through the left or right door. It was not possible to backtrack (i.e., to exit a room via the door from which it was entered). We denote the cardinal directions of movement (N, E, S, W) as *actions* $a \in \mathcal{A}$,

where (due to one-way doors and the no-backtracking rule) on each particular trial, only a subset $A \subset \mathcal{A}$ of 1–3 directions can be selected. Once an acceptable choice was made, subjects viewed an animation moving to the selected adjoining room. In order to encourage planning of new routes, with a 10% probability at each step, but no more often than every 4 steps, a *jump* occurred in which a new room was selected at random from all 16 and instead of arriving in their chosen room, subjects viewed an animation rising above the maze and dropping into the new location.

Four rooms were designated as reward rooms, each with a corresponding fixed reward value of 2 or 3 units, such that each time a reward room was visited the stated reward was received. The locations and values of these rooms were instructed to the subjects and also represented in the visual display by flags above the rooms, visible from a distance. At the end of the study, subjects were paid proportional to final reward count (at $0.04 per unit).

The critical dynamic element of the task, designed to drive learning, was ongoing, random reconfiguration of the available transitions between adjoining rooms (Fig. 1B). Following each decision step, the doors between rooms could reverse their direction; this would happen independently at each door with probability $\frac{1}{24}$. This change process was additionally subject to the constraint that each room would always have at least one available exit. Only the state of the doors leading to or from the current room was visible on any particular trial (represented with colored signs at each door, with those visible in other, distant rooms colored gray), so subjects did not know when changes in the doors occurred until they encountered them.

Subjects were fully instructed on the dynamics of the task, including specific instruction of the independence of the random processes associated with jumps and doors (supplemental Fig. 1). Before scanning, subjects trained and practiced the task for 10 minutes on a different layout than would be used for the main experiment (reward locations and door directions). After entering the MRI they performed 25 trials to familiarize them with the scanner interface and reward locations, and then performed 1000 decision steps during functional image acquisition, with breaks every 250 steps.

Behavioral

We analyzed the sequences of subjects' choices (a_t) by comparing them step by step to those predicted by different learning algorithms modeled as having encountered the same state (s_t), action (a_t), reward (r_t), and jump (j_t) sequence up to each step. In particular, we compared different algorithms for evaluating actions, each formalized as a method for estimating an action value function ($Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$) based on earlier observations (e.g., of rewards received and available doors). The action value function maps each potential action at time t to a predicted value sum of expected future rewards (r) for each available option, discounted for delay according to the free discount parameter γ:

$$Q_t(s, a) = \mathrm{E}\left[\sum_{i=1} \gamma^{i-1} r_{t+i} \middle| s_t = s, a_t = a\right] \quad (1)$$

where each algorithm specifies a particular method for estimating this expectation. For each algorithm, we assumed a softmax decision rule to produce a probability of a choice (p) given the predicted values of all the available choices:

$$p_t(a) = \frac{\exp \beta Q_t(s_t, a)}{\sum_{a' \in A} \exp \beta Q_t(s_t, a')} \quad (2)$$

where β is a free "temperature" parameter controlling the degree of randomness in action selection.

For each algorithm we estimated the set of free parameters (θ, including γ and β), separately for each subject so as to minimize the negative log likelihood of all observed choices (i.e., the sum over the log of equation (2), for the action chosen on each of n trials):

$$l(\theta) = -\sum_{t=1}^{n} \log p_t(a_t|\theta)$$

To compare the quality of model fit correcting for the number of free parameters optimized, we estimated Bayes factors (Kass and Raftery, 1995), the ratio of the model evidences, i.e., the probabilities of the models given the data. To approximate the model evidence we computed BIC (Schwarz, 1978):

$$l(\hat{\theta}) + \frac{m}{2}\log n$$

where $l(\hat{\theta})$ is the negative log likelihood of data at the maxmimum likelihood parameters, $\hat{\theta}$; m is the number of free parameters optimized; and n is the number of observations or (non-trivial) choices the subject made (note that BIC as we define it is $\frac{1}{2}$ the standard definition, to put it in the same scale as likelihood and evidence measures; all statistical tests are corrected appropriately). As a standardized measure of model fit, we also report ρ^2, a pseudo-r^2 statistic which is analogous to a measure of variance accounted for and is computed as $1 - \frac{l(\hat{\theta})}{l_{\text{random}}}$ (Camerer and Ho, 1999; Daw et al., 2006). Also, allowing that the algorithm used might differ across subjects in the population as a random effect, we report statistical tests on the Bayes factors across subjects, along with the "exceedance probability" or posterior probability that one algorithm is the most common of a set across the population (Stephan et al., 2009), as computed using the `spm_BMS` function in SPM8.

To generate regressors reffecting predicted quantities from the models for fMRI analysis (below) we simulated the models for all subjects using a single set of parameters taken as the median of the best-fitting parameters over the individuals. The group median can be viewed as an estimator for the group-level parameters in a random effects model of the population (Holmes and Friston, 1998). We took this approach because we have repeatedly observed, in this and other data sets (Daw et al., 2006; Gläscher et al., 2010), that neural regressors generated using separate maximum likelihood estimates of the parameters produce poorer fMRI results (i.e., noisier neural effect size estimates and diminished sensitivity). This is likely because parameters are not always well identified at the individual level, and variability in the point estimates effectively results in noisy rescaling of regressors between subjects, which in turn suppresses population level significance in fMRI (see Daw (in press) for further discussion).

To compare subjects' performance in terms of payoffs earned, we determined two reference point payoffs for each subject: expected random payoff, and maximum possible payoff. Expected random payoff was determined simply by calculating the expected state occupancy under a uniform random policy, and weighting the rewards by their location's expected occupancy (note that this is slightly different that uniform occupancy due to the heterogeneous connectivity: more central rooms are more likely to be visited). The maximum possible payoff for a subject was defined as the largest payoff possible over all possible choice sequences for the particular sequence of door configurations that subject encountered. Note that actually taking advantage of such a policy would require the subject

to be omniscient or "psychic" about all current and future unobservable door changes. Because perfect play using only the available information is computationally intractable due to the partially observable nature of the task, we neither determined nor compared this value, but it is guaranteed to be somewhere between the best average play from any of our formalized algorithms and "psychic" play.

The timing of the task was such that the choices were first allowed to be entered 500 ms after all the information relevant to that choice was presented (Fig. 1A). As such, the task was not well-suited for analyzing reaction times, since subjects were presumably able to pre-plan their responses and time them to the appropriate moment. In order to examine reaction time effects given these limitations, we discarded all trials with reaction times under 50 ms, and analyzed the remainder using the same regressors as with fMRI (see Model-based Analysis below) as explanatory variables in linear regressions in which the dependent variable was taken as the log reaction time. Regression coefficients were computed per-subject, then tested across subjects to assess their significance as random effects (Holmes and Friston, 1998).

Algorithms

Although the task was simple to understand, an optimal solution is computationally intractable. This design allowed for a wide range of possible (suboptimal) strategies that could be employed. Thus, in analyzing the behavioral data, we are faced with (and did explore) a wide variety of algorithms employing different representations and learning methods based on both TD and planning processes.

The main questions of the study concern valuation by model-based planning. Such a strategy is categorically distinguished from more common "model-free" approaches to RL by two key features: the use of a model representing the environment, and on-line evaluation based on recently learned changes to this model. For specificity and efficiency, for the bulk of the analyses we report, we used a canonical model-based algorithm (value iteration) that exhibits these features. It is canonical in the sense of being derived directly from a formal definition of the decision problem (see e.g., Sutton and Barto (1998)); it is also, in the particular details and approximations of this derivation, the best fitting algorithm we discovered from the model-based class. To verify that behavior and BOLD signals are best explained by an approach of this sort, we compare its predictions to a canonical model-free algorithm (Q learning), which was also the best-fitting representative of that class we discovered. We additionally compare both algorithms (see supplemental material) to reduced or extended variants that isolate particular distinguishing features of the model-based and model-free approaches. However, these best fitting models, by virtue of being derived from a decision-theoretic definitions, are also computationally complex. Accordingly, we do not suggest that these algorithms are direct process-level accounts of the steps of computation, but rather that they are representative of the overall form of the relationships between experience, representation, and choices or BOLD activity. Additionally, as discussed further below, the quantities that these algorithms defined also help us to examine some process-level questions.

In the following descriptions, we take as data the experience of each subject over steps, t: visited states, $s_t \in \mathcal{S}$; rewards, $r_t \in \mathbb{N}$; available actions, $A_t \subset \mathcal{A}$; choices, $a_t \in \mathcal{A}$; and jumps, $j_t \in \{0, 1\}$. Each state action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ represents one side of a particular door within the maze, where only valid doors are considered. We use the fixed transition function $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ such that $T(s, a) = s'$ if behind the door (s, a) is the room s' (regardless of whether it is currently open), along with the reward map $R : \mathcal{S} \to \{0, 2, 3\}$ to represent the fixed reward locations. We also use the symbol A_t to indicate the set of available outgoing doors from the room s_t at time t. For clarity, we thus have the following invariants:

$$\begin{array}{rr} r_t = R(s_t) & \forall t \\ a_t \in A_t & \forall t \\ T(s_t, a_t) = s_{t+1} & \forall t : j_t = 0 \\ \exists a . T(s,a) = s' \iff \exists a' . T(s', a') = s & \forall s, s' \end{array}$$

Planning—Rather than estimating action values directly, a model-based approach learns a "model" of the structure of the task — here the current configuration of the maze — and computes action values by searching across possible future trajectories, accumulating the rewards in expectation according to the definition of these values (equation (1)).

To learn the model, our implementation represents the subject's estimate of the direction of each one-way door as a *probability* of it being open, $p_t(s, a)$, which is updated when a door is observed, and also decayed at each step by a free factor η, in order to capture subjects' knowledge of the chance that doors may have changed since last observed, as well as any other processes by which observed door knowledge plays a declining role in valuation (e.g., forgetting or search pruning). The two sides of each door may be learned independently, even though this may create a model inconsistent with the one-way dynamics. The probabilities are initialized to 0.5 and updated at each step in which a set of open doors A_t is observed in room s_t, according to:

$$\begin{array}{lll} p_t(s,a) \leftarrow p_{t-1}(s,a) + \eta(0.5 - p_{t-1}(s,a)) & \forall s, a & \text{(decay)} \\ p_t(s_t, a) \leftarrow \begin{cases} 1 & \text{if } a \in A_t \\ 0 & \text{if } a \notin A_t \end{cases} & \forall a & \text{(model update)} \end{array}$$

The other part of the task model — the reward value for each room, $R(s)$ — is assumed to be known. We believe this assumption to be innocuous, since this information was fixed, instructed, and signaled in the visual display.

Using the learned maze configuration, we compute state-action values based on a tree-search planning process terminating at reward states. For computational efficiency (e.g., in fitting free parameters to choice data), we implemented this planning process using value iteration, which simply unrolls a breadth-first search tree over the states from leaves (horizon 1 values) to roots (end horizon values). Specifically, at each step, in room s_t, initialize all $Q(s, a) \leftarrow 0$ and, for all (s, a) pairs in parallel, repeatedly perform:

$$Q(s,a) \leftarrow \begin{cases} R(s') & \text{if } R(s') \neq 0 \\ \gamma \sum_{A' \subseteq \mathcal{A}} \mathrm{P}[A' | \vec{p}_t, A' \neq \varnothing] \max_{a' \in A'} Q(s', a') & \text{otherwise} \end{cases} \quad \forall s, a, s' = T(s,a) \tag{3}$$

We took $Q_{\text{plan}}(s_t, a)$ to be the value resulting after 16 iterations of this update.

Here the sum takes an expectation over possible sets of open doors A' in state s' according to the current beliefs about the probability of each door being open individually, the no-backtracking constraint (here, that the door $T(s', a') = s$, from which s' is entered, must be closed), and the constraint that at least one door must be open:

$$\mathrm{P}[A' | \vec{p}_t, A' \neq \varnothing] = \frac{\prod_{a' \in A'} p_t(s', a') \prod_{a' \notin A'} (1 - p_t(s', a'))}{1 - \prod_{a'} 1 (1 - p_t(s', a'))}$$

The algorithm thus has 3 free parameters: η, γ, β.

Although this algorithm is derived directly from the definition of action value from equation (1), it does incorporate a number of simplifications or approximations, all of which accorded well with the data. First, we terminate each search path at reward. In terms of the definition of the decision variable, this is equivalent to treating the reward states as terminal in an episodic view of the problem (Sutton and Barto, 1998). Similarly, we terminate each search path if no reward has been found along it by a depth of 16. This is an innocuous assumption since the value of a reward converges to zero as its distance increases, given $\gamma < 1$ or $\eta > 0$. (Sixteen steps is above the maximum distance between any two points in the maze, and was well beyond the point at which relevant fit quantities changed meaningfully given the data; see supplemental material.) In terms of the model, the evaluation of the expectation treats the model as frozen throughout the iteration process, i.e., it does not take into account the effect of potential future observations and updates on the model (as a full Bayesian/POMDP approach would do). Finally, and closely related to this, it approximates the expectation over maze configurations as a factored tree of states, by treating the probability of a particular door set being open as independent between states within each iteration and, for each state, also independent between each iteration of the value update. These last assumptions allowed the algorithm to execute in reasonable time.

Finally, at the process level, there are many different approaches to evaluating the multi-step value expectation from equation (1). For instance, it seems most plausible that subjects search forward from the current state (or perhaps backward from a goal state), rather than from all states in parallel as in value iteration. However, for the current state, the total value and also the intermediate values (each step's *n*th horizon partial sums) from value iteration correspond to those that would be computed at each step by a breadth-first search. Other search processes, such as depth-first, visit the states in different order, and perhaps (e.g., due to stochastic pruning or early termination) only visit a subset of them on any particular trial. However, since a very wide family of such approaches can be viewed as different ways of evaluating the expectation defined by equation (1), their end values should coincide either exactly or (particularly in the average over trials) approximately with those we compute here. For instance, the end values we compute correspond, in the average, to those that would be computed if discounting is eliminated but paths are instead terminated stochastically with probability γ; or to values accumulated over a trajectory where door traversals are not weighted according to p_t but instead sampled with this probability (see Sutton and Pinette (1985); Suri and Schultz (2001); Smith et al. (2004) for related models).

TD—We use a model-free Q-learning algorithm (Watkins, 1989), augmented with eligibility traces. Such an algorithm maintains a representation of the state-action value function Q directly, and updates it locally following experience with particular state-action pairs and rewards. The inclusion of eligibility traces, for $\lambda > 0$, allows the algorithm to update the values for states and actions other than the pair most recently observed, but only backward along the recently encountered trajectory. In this implementation (unlike Watkins'), eligibility traces are truncated on "jump" events but not for exploratory actions.

The model has five free parameters: $Q_0, \alpha, \gamma, \lambda, \beta$. Specifically, each door within the maze, (s, a), is associated with a value, $Q_t(s, a)$, all initially set to Q_0. Each also has an associated trace $e_t(s, a)$, all initially 0. At each step, if door a_t is chosen in room s_t, arriving in room s_{t+1} (either via a jump, $j_t = 1$, or not, $j_t = 0$) with reward $r_{t+1} = R(s_{t+1})$ the variables are updated according to:

$$
\begin{aligned}
e_{t+1}(s,a) &\leftarrow \gamma\lambda e_t(s,a) && \forall s,a && \text{(decay traces)}\\
e_{t+1}(s_t,a_t) &\leftarrow e_{t+1}(s_t,a_t)+1 && && \text{(accumulating traces)}\\
e_{t+1}(s,a) &\leftarrow (1-j_t)e_t(s,a) && \forall s,a && \text{(truncate traces on jump)}\\
v &= r_{t+1}+\gamma \max_{a' \in A_{t+1}} Q_t(s_{t+1},a') && && \text{(value prediction)}\\
\delta_t &= v - Q_t(s_t,a_t) && && \text{(prediction error)}\\
Q_{t+1}(s,a) &\leftarrow Q_t(s,a)+\alpha e_{t+1}(s,a)\delta_t && \forall s,a && \qquad (4)
\end{aligned}
$$

$Q_{TD}(s; a)$ is simply the learned value function $Q_t(s; a)$.

Imaging

Functional imaging was performed on a 3T Siemens Allegra head-only scanner with a custom head coil (NM-011, Nova Medical, Wakefield, MA) located at the Center for Brain Imaging at New York University. Thirty-three contiguous oblique-axial EPI images (3 × 3 × 3 mm voxels) were obtained each 2000 ms TR, oriented 23° off the AC–PC axis so as to improve functional sensitivity in orbital frontal areas (Deichmann et al., 2003). Slices were positioned to obtain full coverage from the base of the orbitofrontal cortex and medial temporal lobes ventrally; coverage extended dorsally/caudally into the superior parietal lobule and above the dorsal anterior cingulate cortex but omitted some occipital and parietal regions, and in a few cases, some posterior-superior frontal regions. A high-resolution T1-weighted anatomical image (MPRAGE sequence, 1 × 1 × 1 mm) was also acquired for each subject.

Images were preprocessed and analyzed using the SPM5 software (Wellcome Department of Cognitive Neurology, London, UK), and final results were corrected for multiple comparisons using SPM8 (Wellcome Trust Centre for Neuroimaging, London, UK). Functional images were realigned for head motion, coregistered between runs and to the structural image, spatially normalized to MNI coordinates (SPM5 “segment and normalize”), and finally resampled to 2 × 2 × 2 mm voxels and smoothed with an 8 mm FWHM Gaussian kernel. Due to the short TR, interleaved acquisition, and fast events, we did not additonally resample temporally to correct for slice timing.

Neural models were analyzed using general linear models to obtain single-subject beta images. Regressors were convolved with SPM5’s canonical hemodynamic response function. To control for nuisance effects, all designs included: the six rigid-body motion parameters that were inferred by realignment; four event regressors covering times in which the subject was viewing animations of left turns, right turns, forward movement, and jump movement respectively; and a “no-choice” impulse event regressor at the time of choices in which the choice set size was one.

Separate coefficients were computed for each regressor for each of the four runs, and contrasts were computed by adding up these coefficients. Contrast values were then brought to the group level using one- or paired-sample t-tests for random effects. Unless otherwise noted, we produced whole brain effect maps using a $p < 0.001$ uncorrected threshold, and then assessed significance correcting for whole-brain multiple comparisons using topological cluster-size FDR, $p < 0.05$ as implemented in SPM8. (Note that cluster-level FDR is distinct from voxel-wise FDR, which has recently been argued to be invalid; Chumbley and Friston (2009).) Accordingly, reported peak t values are uncorrected, and significance is in relation to the containing cluster. SPMs have been displayed graphically by including all uncorrected activations, with clusters that did not reach significance, where assessed, depicted in a lighter, translucent color.

Model-based analysis

Each GLM included an event regressor containing an impulse at each choice (response) time, along with some number of parametric regressors on these events, depending on the particular computational algorithm being analyzed. These regressors were mean-corrected separately within choice and "no-choice" trials (according to the corresponding nuisance regressor), but, except where stated, when multiple parametric regressors were entered in a design, these were not orthogonalized against one another. Parametric regressors were derived from the sequence of predicted values or other latent variables produced by each algorithm, according to the learning algorithm exposed to the subject's actual experience up to the current trial. Because we were interested in many different, often highly correlated, properties of the neural signals, such as different deconstructions of the value signal, we ran separate GLMs to ask different questions, primarily focusing on distinct brain regions.

In order initially and qualitatively to identify basic activation patterns related to the predictions of either algorithm separately (which are correlated), the first two analyses entered the predicted $Q(s, a)$ values for the current state and chosen action (from equations (4) and (3)) as parametric regressors, with GLM1 containing only Q_{TD} and GLM2 only Q_{plan}. (We refer to these as the "chosen values.") To identify peak value-responsive voxels in an unbiased manner, and to directly compare the fit of these regressors, GLM3 contained both of these values as separate regressors, and a contrast summing the coefficients from both was used on the second level. Confining the analysis to an anatomically-defined striatal ROI (Maldjian et al., 2003) — which was of specific interest because it is often associated with TD (O'Doherty et al., 2006; Lohrenz et al., 2007) — we found all peak (locally maximally responsive) value voxels from GLM3 in caudate and putamen that exceeded a $p < 0.001$ uncorrected threshold and defined these as our voxels of interest (VOIs). In order to then compare between these two predictions with an independent test, the orthogonal contrast[1], taking the difference between the planning and TD coefficients, was used on these VOIs.

In order further to decompose and explore differences between the algorithms' predicted value signals, we performed additional analyses using a series representing the chosen values $Q(s, a)$, as they would be computed either by the TD or planning algorithms. Both algorithms can be viewed as representing cues as exponentially decaying sums of terms, either of expected rewards (for planning), or of prediction errors previously encountered at a state (for TD). Since the GLM is additive, we can use separate regressors to express the BOLD signal by their weighted sum, and estimate the relative weights.

For planning, the state-action values are explicitly computed as a sum of exponentially-discounted expected rewards expected at each future step (from equation (1)):

$$Q_{\mathrm{plan}}(s_t, a_t) = \sum_{i=1} \gamma^{i-1} \mathrm{E}\left[R(s_{t+i})\right]$$

Here the expectation is as in equation (3), and simply unrolls the independent reward terms from that computation. When making predictions from a forward breadth-first search, this is exactly how the values are computed: by considering reward at the next step, then potential rewards at the subsequent step, and so on. (Other types of search, e.g., depth first, do not

[1]Because there is correlation between TD and planning predictions, these two contrasts are not perfectly orthogonal in the space of the (temporally whitened) design matrix (Kriegeskorte et al., 2009). In fact, they are slightly anticorrelated ($r = -0.142$). This equates to a bias towards finding a more negative difference, i.e., TD coefficients being larger, when having first selected on their sum being large and positive; we neglect this bias since it works against the results reported here.

visit the states in the same order, but insofar as they compute the same end values they may still be decomposed this way.)

For the TD algorithm, the rewards expected at individual states are never explicitly represented, but instead values are produced (i.e., learned through the action of the learning rule) by accumulating them over time with each prediction error update, weighted by the learning rate. That is, we can unroll the effects of the iterated updates in equation (4) in a form similar to equation (1), expressing the learned value at a particular time as the exponentially weighted sum of previous prediction errors. For $\lambda = 0$, this sum is over the prediction errors encountered on previous state-action pair choices:

$$Q_{\mathrm{TD}}(s_t, a_t) = \sum_{i=1} \alpha(1-\alpha)^{i-1} \delta_{u_i(s_t,a_t)} \tag{5}$$

Where $\vec{u}(s_t, a_t)$ is the sequence of times at which that action was chosen in that state prior to time t, so that $u_0(s_t, a_t) = t$, $u_1(s_t, a_t)$ is the time of the first such preceding visit, and so on. When $\lambda > 0$ (as in our behavioral fits), a state-action is updated not only following visits to it, but also by prediction errors subsequently encountered at other states, weighted by the decaying eligibility trace. We can modify equation (5) to account for this effect by taking the terms δ in the sum to be themselves accumulated series of single-step prediction errors encountered subsequently to the state visit:

$$\delta_i = \sum_{k=0} (\gamma\lambda)^k \delta_{u_i(s_t,a_t)+k}$$

Here k ranges up until whichever is first: $u_i(s_t, a_t) + k = t - 1$ (the present) or $j_{u_i(s_t,a_t)+k+1} = 1$ (the first subsequent "jump") at which point eligibility is cleared.

For both algorithms, if the BOLD signal is representing the corresponding value, it should reffect the sum of all these terms, with the appropriate coefficients: the sums essentially unroll the computation or learning of the values as predicted by either algorithm. To investigate these predictions, we created two more designs that decompose the two chosen values using the first two terms of either sum. (Since the weights are exponentially decaying, the earliest terms should dominate.) GLM4 included parametric regressors for $\delta_{u1(s_t,a_t)}$ and $\delta_{u2(s_t,a_t)}$ from the TD algorithm, and GLM5 included E $[R(s_{t+1})]$ and E $[R(s_{t+2})]$ from planning (referred to as δ_1, δ_2, r_1, and r_2 respectively). For comparison, we also inferred what the expected coefficients from this analysis would be based on the Q coefficients from GLM1 and GLM2 and the (behaviorally fit) values of γ and α. These two GLMs were initially applied only to the identified VOIs. This analysis is similar to a number of techniques used to analyze neural data in terms of value subcomponents (Bayer and Glimcher, 2005; Montague et al., 2006; Samejima and Doya, 2008).

We additionally used GLM5 to seek areas better correlated with only the expected next reward r_1, viewed here as an intermediate quantity in the value computation as opposed to a portion of the full value. (As noted, r_1 is indeed the first partial sum computed during a breadth-first search; for another approach like depth-first, it might be viewed as the expectation over trials of the value of the first state visited.) We thus sought activity related specifically to r_1 rather than the cumulative future reward $Q \approx r_1 + \gamma r_2$, using the contrast $\sqrt{1+\gamma^2} r_1 > r_1 + \gamma r_2$. (Note that this contrast equates the length of the two contrast vectors to avoid confounding the test of the direction of the neural effect.)

Next, in looking for effects related to the iterative computation of future values, we first considered the total number of choices available from the current state, n_0. This information is clearly relevant for any decision-making system that considers all the options, and in particular, an algorithm that searches forward through possible routes will have this many starting points. We then considered the next-step expectation of this quantity: the *expected* number of total choices in all reachable rooms given the model of the doors specified by the best-fitting planning algorithm, using the behaviorally fit value for η and the subject's observations up to the current point:

$$n_0 = |A_t|$$
$$n_1 = \sum_{a \in A_t} \sum_{A'} \mathrm{P}\left[A_{t+1} = A' | s_{t+1} = T(s_t, a), \vec{p}\right] |A'|$$

where the conditional expectation is the same as in equation (3). Although the normative planning algorithm as we actually implement it examines all state-action-state pairs regardless of how likely it is that a door exists, a more realistic process-level search implementation would likely "prune" or examine the most likely transitions, thus requiring expected computation proportional to n_1. We constructed GLM6 with a regressor for n_0 and a regressor for n_1 orthogonalized against the n_0 regressor[2]. We also included a regressor of no interest containing reaction time for each trial, against which the other two regressors were orthogonalized. We then identified all voxels significantly responsive to n_0 ($p < 0.05$ cluster-size FDR on $p < 0.001$), and used this as a mask to identify regions responsive to n_1 using $p < 0.001$ and assessing significance with small-volume FWE correction. Unfortunately, because of our slice prescription, three subjects ended up with reduced coverage of superior frontal regions, resulting in these areas being masked out of our analysis due to missing data. Thus, to study the extent of activity identified in pre-motor regions, these three subjects were left out and the GLM6 analysis repeated using the remaining 15 subjects.

Finally, in order to investigate whether obtained results were specifically related to model-based planning processes, we studied how neural effects covaried with the degree to which the planning or TD models fit their data (measured by the per-subject log likelihood of the choice data under either model, or the difference between the two). In particular, we selected the per-subject β values from the peaks of the relevant contrast and correlated these with the log likelihood measures from the per-subject behavioral fits, assessing one-tailed significance for the correlation coefficient. Since the contrasts used to define the peak voxels are main effects over all subjects, and since, further, they are extrema of contrasts unrelated to the likelihood measures, the resulting correlations will be unbiased and not subject to corrections for the whole-brain multiple comparisons involved in seeking the peak voxel.

4 Results

Behavioral

On average over 1000 steps, subjects earned \$23.78 ± \$1.91 (mean ± 1 standard deviation). These earnings exceeded what would have been expected under chance performance by 12.5% ± 8.8% on average, which was significantly different from zero across subjects ($t_{17} = 6.82$) and numerically greater than zero for 16/18 subjects individually. While it is computationally intractable to define the earnings of an optimal decision maker in this task, an upper bound on this quantity is the earnings of a "psychic" subject who was fully

[2]Again, in order to be sure that these tests are independent, we need to consider whether these are truly orthogonal contrasts given temporal autocorrelation. Post-whitening, we find that they are very slightly anticorrelated ($r = -0.034$), so since we are only looking for where their signs agree, this can only make the test more conservative.

informed about the maze state at each step, and behaved optimally according to this knowledge. On average, earnings were 10.4% ± 5.4% worse than this benchmark. Together, these results suggest that subjects were reasonably successful at harvesting rewards.

Learning models

We attempted to characterize subjects' learning — that is, how their choices depended on previous feedback — by fitting two alternative algorithms to explain their trial-by-trial choices. These exemplify two representational strategies for reinforcement learning: a model-based planning approach, which learns a representation of the maze layout and evaluates actions using it, and a model-free TD approach that learns estimates of actions' values directly and locally. These approaches have been argued to formalize a longstanding distinction in psychology between response-based approaches and more cognitive, map-based or goal-directed approaches (Doya, 1999; Dickinson and Balleine, 2002; Valentin et al., 2007; Gläscher et al., 2010). We hypothesized that the task would favor a model-based strategy instead of the model-free strategy quantified in many previous fMRI studies of decision making (Daw et al., 2005), allowing us to examine the neural implementation of such learning. In this task, the ongoing maze reconfigurations play a similar role to an outcome revaluation manipulation (Balleine and Dickinson, 1998), allowing the strategies to be distinguished by their distinct predictions about how behavior should adjust following observed changes. In particular, although the strategies are related in that they are pursuing the same ends, they make different trial-by-trial predictions about choices due to drawing on past experience to evaluate options using different strategies and representations. Notably, the TD approach updates actions' predicted values only locally after they are encountered (via a so-called bootstrapping process in which value estimates are updated based on adjacent ones), whereas a model-based approach incorporates all learned information into a map of the environment resulting in a global update of the derived action value estimates. This delay in the propagation of learning in a TD model predicts that choices should sometimes not respect recently learned information (Daw et al., 2005).

We fit each subject's trial-by-trial choice behavior individually with each model, and assessed the relative goodness of fit. Aggregating the data likelihoods across subjects (which is equivalent to assuming that all subjects used the same one of the models), the group's behavior was best explained by planning (BIC 3098), and worse by TD (BIC 3397; random was 4085).

We may instead consider that the identity of the best fitting model might have varied from subject to subject, and characterize the population's tendencies by the summary statistics on their individual fits, analogous to a random-effect analysis in fMRI (Stephan et al., 2009). Thus, the average Bayes Factor (the difference in BIC scores or approximate log odds in favor of one model vs. another) was 16.61 in favor of model based planning over TD. This was significantly different from zero across subjects ($t_{17} = 4.92$, $p = 0.0001$), indicating that a subject drawn randomly from the population will, on average, exhibit behavior better fit by model-based RL compared to TD. An alternative way to characterize the predominance of the strategies in the population is to fit the entire behavioral dataset with a mixture model in which each subject exhibits exactly one of the candidate algorithms, the identity of which is treated as a random variable (BMS, Stephan et al. (2009)). In such a fit, it was overwhelmingly likely that planning was the more common strategy (expected frequency E [p(plan)] = 0.947, "exceedance probability" P [p(plan) > p(TD)] > 0.999). These results suggest that subjects' learning about choices in this task was, at the population level, predominantly driven by model-based spatial planning. The comparison of the model-based and TD approaches suggests that values are determined prospectively by planning rather than by local bootstrap-based learning.

We may also break down the contributions of individual subjects to these group-wise results. Primarily, both of the learning models fit significantly better than chance for each subject (likelihood ratio tests; for TD all $\chi^2_5 \geq 13.32$, $p < 0.031$; for planning all $\chi^2_3 \geq 12.02$, $p < 0.008$). Comparing BIC scores for each individual, planning was favored over TD for 17/18 subjects. These results further indicate that while there is some evidence of individual variability among the subjects, the predominant strategy appears to be model-based RL (Fig. 2).

We additionally compared each algorithm to a reduced or augmented version, to isolate the necessity of the kind of learning each posited. In particular, we tested "dead reckoning" variants of the planning algorithm that did not involve on-line learning of the map of doors but instead evaluated actions only on the basis of the distance to reward, essentially relying only on the known spatial structure and reward locations. The full planning model explained choices better than these variants, supporting the interpretation that subjects plan using a learned transition map. Even so, dead reckoning models still fit the choices better than TD, providing further evidence that even simple planning processes dominate TD learning in this task. In fact, the dead reckoning fit was not improved by incorporating TD learning such that the fixed distanced-weighted values were updated based on experience using TD. For full details on these analyses and comparisons with other variants in each model class, see supplemental materials, supplemental Fig. 2, supplemental Fig. 3, and Table S1.

Reaction times

We reasoned that if subjects were planning trajectories by forward search, as our results suggest, then this might be reflected in their reaction times as well as their choices. In particular, we hypothesized that subjects' reaction times would be longer on steps when the search was more extensive. This would predict longer reaction times not only in rooms in which they were facing more open doors (a quantity we called n_0, see Methods), but that they would also be longer for searches in which they *expected* that more doors would be open in subsequent rooms, a measure unique to a forward planning model. We defined this quantity, n_1, in expectation at each step according to the model's learned beliefs about the maze.

One complication in assessing this hypothesis is that subjects were allowed to enter a decision one-half second after they first entered a room and observed the doors available there. This pause allowed subjects to decide and prepare their responses during this time, making reaction times a poor measure of planning. Accordingly, there were a high proportion of extremely fast responses: reaction times averaged 278 ± 251 ms, with 36.1% of responses under 150 ms and 11.8% under 50 ms. In order to focus on the subset of trials in which reaction time might reflect differential amounts of planning, we eliminated the fastest reaction times (those less than 50 ms) from analysis.

For the remaining subset of trials, we found weak but significant effects of the search complexity ($t_{17} = 4.50$, $p = 0.0003$ for n_0, $t_{17} = 2.49$, $p = 0.023$ for n_1) on log reaction time, such that more complex choices resulted in longer reaction times.

Imaging

Given that the behavioral analysis indicated that the predominant learning strategy among our candidates was model-based planning, we next exploited this model to interrogate related neural signals. Our overall strategy, based on previous work on TD learning, was to use simulations of the fit algorithm to define trial-by-trial timeseries of relevant variables such as predicted action values, in order to seek and tease apart neural correlates of these otherwise subjective quantities (O'Doherty et al., 2007). For this, we used on the model-

based algorithm (and, for initial analyses, also the TD one) along with the medians of the parameters that best fit the individual subjects' choices (Table 1; see also supplemental Table S1). The median was used because in our experience (Daw et al., 2006; Gläscher et al., 2010; Daw, in press), unregularized maximum likelihood parameter estimates from individuals tend to be too noisy to obtain reliable neural results. Since what distinguishes planning from model-free RL is that it constructs action values from more elementary information on-line at choice time rather than simply retrieving previously learned aggregate values as in TD, our primary questions concerned dissecting these computations. As a first step, we sought neural correlates of aggregate chosen values; we did this both for values predicted by planning and for those predicted by TD in order to verify our hypothesis that in this task neural value signals, like choices, were predominantly better explained by planning. To maximize power, we compared the algorithms on the basis of BOLD signal in value-responsive voxels selected in an unbiased manner. After similarly confirming that planning predominates in striatal value signals over the population, we proceeded to tease apart these activations and the computations that we hypothesized would give rise to them by examining neural correlates of the components of model-based value construction. In particular, model-based predictions of aggregate future value are based on two quantities: predicted single-step rewards, and predicted future states (based on knowledge of state transitions). We thus sought neural correlates related to both of these hypothesized representations. Finally, in order to verify the extent to which our results related specifically to planning processes as opposed to valuation more generally, we studied whether individual variation in the strength of our neural effects covaried, across subjects, with the extent to which planning vs. TD explained their choice behavior.

Unless otherwise stated, all *t*-statistics are uncorrected and refer to peak voxels of clusters that have been deemed significant $p < 0.05$ by cluster-size FDR correction.

Correlates of value

Following much previous work on RL, we began by generating the sequence of values, Q, that each algorithm would predict at each step of the task on the basis of previous experience (Tanaka et al., 2004; O'Doherty et al., 2006; Seymour et al., 2007; Wittmann et al., 2008). We first asked where the BOLD signal significantly correlated with the sequences of values for the chosen actions ("chosen values"), for TD and planning considered separately. For plan-predicted values within striatum, we found that clusters in bilateral posterior vetrolateral putamen/claustrum and bilateral dorsolateral prefrontal cortex extending ventrally into orbitofrontal regions correlated significantly (peaks: [−32,−8,−6] $t_{17} = 6.60$, [28, −2, −10] $t_{17} = 4.72$, [−48, 38, 18] $t_{17} = 6.79$, [52, 40, 8] $t_{17} = 5.34$). We also found a weaker correlation in the bilateral ventral caudate that did not survive cluster-level correction for multiple comparisons (peaks: [−24, 16, −8] $t_{17} = 4.97$, [12, 14, −10] $t_{17} = 4.53$, n.s.). Using the TD-based values, we found only one significant cluster in striatum (left vetrolateral putamen peak: [−30, −8, −6] $t_{17} = 5.34$). The ventral and dorsomedial regions of striatum have commonly been identified with general reward and value expectations (Delgado et al., 2000; Knutson et al., 2001; O'Doherty, 2004; Tricomi et al., 2004; Samejima et al., 2005), although the posterior putamen regions have been less commonly implicated (Delgado et al., 2003; O'Doherty et al., 2004). Outside of striatum, we found that distinct regions of bilateral inferior frontal, postcentral, and insular cortex showed significant correlations with these values as well, again apparently more robustly to the planning values (Fig. 3). However, cortical correlates of chosen value are found in ventral pre-frontal areas (Knutson et al., 2005; Kable and Glimcher, 2007; Hampton and O'Doherty, 2007; Gläscher et al., 2009; Wunderlich et al., 2009; Chib et al., 2009) more often than other areas (Breiter et al., 2001; Plassmann et al., 2007; Hare et al., 2008). On a targeted investigation of vmPFC, we found that there was a correlation with planning values in this

region that did not meet our reporting threshold for uncorrected significance (peak: [−6, 40, 8] $t_{17} = 3.03$, $p = 0.003$ uncorrected), while correlates with TD values here were much weaker ($p > 0.01$).

As expected given the findings that the behavior showed evidence predominantly for only a single sort of (model-based) learning, as well as the correlation between the two sets of values, the two maps show a good deal of similarity, though overall the planning responses are slightly stronger. This suggests, consistent with the choice analysis, that neural signals aligned with model-based valuation. However, a difference in thresholded statistical maps does not itself demonstrate a statistically significant difference ("the imager's fallacy", Henson (2005)); thus we next sought to confirm statistically the superiority of planning.

TD vs. model-based value

Since the algorithms made similar predictions for many situations, and thus the chosen value regressors themselves were substantially correlated between algorithms ($r = 0.791$; for related correlations see supplemental Fig. 3), we wished to maximize power for comparing them by minimizing multiple comparisons. To this end, we targeted our direct comparison by first identifying a small number of voxels of interest (VOIs) in an anatomically-constrained striatal ROI that showed value-selective activations. We identified these voxels in a manner that did not bias the subsequent test for differences between the two chosen value regressors, by using a summed contrast over both of them. Based on other studies, we focused on the caudate and putamen as regions relevant for value computations (see Methods for more information on voxel selection). The four selected VOIs were [−32, −8, −6] (left ventrolateral putamen), [28, −6, −12] (right ventrolateral putamen/pallidus border), [−20, 16, −6] (left anterior ventromedial putamen), and [20, 18, −10] (right anterior ventromedial putamen) (Fig. 4). We then asked whether these voxels were significantly more correlated with planning or TD-based chosen values using the orthogonal (difference) contrast between both regressors in the same GLM. Each of these voxels showed a significantly stronger response to the planning predictions ($p < 0.03$ FDR corrected for the four comparisons). These findings provide evidence that neural correlates of value in striatum are not strictly bound to a TD-based value computation, and may instead be informed by a model of the environment. This sharply contrasts with the common interpretation of the mesolimbic dopamine system as implementing TD learning (Seymour et al., 2004; Lee et al., 2004; O'Doherty et al., 2006).

We also repeated these neural analyses using a version of TD augmented with "dead-reckoning" planning values for initialization. That the results were largely the same (see supplemental material, supplemental Fig. 4, and supplemental Fig. 5) further suggests that value signals in ventral striatum reflect those generated from a learned cognitive map rather than from TD learning.

Decomposition of value responses

Having established that neural correlates of value prediction, like choices, were well explained by model-based values, we sought to dissect the computation of these values into the components from which, according to the theory, they are computed. We first attempted to tease apart the striatal BOLD response by separately investigating the effects of component quantities that should be combined together in the value computation, and in so doing to visualize the features of the response that gave rise to the previous finding (that value is better explained by model-based planning than TD). In particular, the values from both algorithms amount to weighted sums over a series of quantities (see Methods). TD updates values by accumulating prediction errors, resulting in a net learned value that at each step corresponds to the weighted sum over errors received following previous

experiences with an action. Model-based planning produces a value at each step by a (time-discount weighted) sum over rewards predicted by the model at each timestep into the future. We therefore may use the additivity of the GLM to seek to explain the net BOLD signal as a weighted sum of either sequence of subquantities. Thus we unrolled the first two steps of each of these computations and entered them as separate regressors: $Q_{\text{plan}} = r_1 + \gamma r_2 + \cdots$, $Q_{\text{TD}} = \alpha(\delta_1 + (1 - \alpha)\delta_2 + \cdots)$. (The early terms should dominate since in both algorithms the weights decline exponentially as the series progresses.)

We extracted the sizes of these effects in our voxels of interest (Fig. 5). As can be seen in the figure, the patterns of BOLD activation in all the voxels, in terms of the subcomponents of the value, more closely follows the pattern predicted by model-based planning than by TD, with the directions and relative sizes of effects consistently in line with the predictions. Having selected these voxels for having a large correlation with the chosen value Q (over both algorithms symmetrically), we of course bias the statistical test for correlations with components of both algorithms' value computation to be positive as well. Thus as expected, we found all the correlations with r_1 were significantly positive ($p < 0.035$, all tests FDR corrected for the four voxels but uncorrected for multiple comparisons in VOI selection), as well as one of the paired r_2 correlations ($p = 0.02$). On the other hand, only one of the δ_1 correlations were significant ($p = 0.033$) while none of the paired δ_2 correlations were ($p > 0.1$), despite the bias toward a positive finding expected from the selection of the voxels.

One-step reward

We next sought neural reflections for the components of the world model elsewhere in the brain, in order to begin to map the broader network supporting the computation. In particular, these analyses use subcomponents of the normatively defined model-based values to examine aspects of their process-level computations. In the theory, model-based values are computed at choice time by summing expected rewards over candidate trajectories. Specifically, model-based planning predicts action values by combining information from two representations: a map of where rewards are located, and a map of transitions between states (here, doors). We thus hypothesized that we should be able to find neural correlates related to both representations. To seek a representation of rewards, we considered the value r_1 discussed above, which has a clear interpretation: it is the expected immediate reward to be received in the next room. This is the first relevant value that a planning process would need to "look up" when searching forward paths, where the expectation over future states can either reflect the average over paths considered first on different trials (as in depth first search) or be explicitly computed in the brain on each particular trial (as in breadth-first search, where the expectation r_1 is the first intermediate partial sum in computing a full action value).

The present analysis thus seeks activity related to the next reward separately, rather than as portion of a net signal related to the chosen value Q, as in the previous section. Of course, r_1 and Q are strongly correlated, as the former is the first term in the sum defining the latter. To distinguish these possibilities and find activity related specifically to elemental rewards rather than aggregate future values, we searched for regions in which the correlation with r_1 was significantly greater than the correlation with the summed value, here approximated by the sum of both of the first two terms in the series $r_1 + \gamma r_2 \approx Q_{\text{plan}}$ (Fig. 6). This contrast revealed a pair of regions containing significant clusters: left superior frontal cortex (peak: [−18, 46, 46] $t_{17} = 5.58$), and right parahippocampal gyrus (peak: [18, −6, −20] $t_{17} = 4.64$; left [−34, −14, −18] $t_{17} = 4.66$, n.s.). Such activity might either represent associations of place with reward, as perhaps in the case of the MTL activations, or reflect the incorporation of these one-step rewards into planning computations (Foster et al., 2000; Hasselmo, 2005; Zilli and Hasselmo, 2008). Also, in addition to their spatial associations, areas in MTL have

been implicated more generally in drawing on memory to project future events (Buckner and Carroll, 2007; Hasselmo, 2009).

Transitions and planning

Next, we sought neural evidence for a representation of the other aspect of the hypothesized world model, the maze transition structure. We did so indirectly, by using hypothesized effects of search complexity on activity related to choice difficulty in a manner analogous to the analysis of reaction times discussed above. In most accounts of decisions, choice difficulty increases with the size of the choice set (here, the number of doors in the current room). However, if actions are evaluated online via some kind of realistic forward planning process, then choice difficulty should also depend on the complexity of the subsequent search: e.g., the number of choices expected to be available in subsequent rooms. (Here again, the expectation, computed from door probabilities in our full model, is meant to reflect the average search-related activity over trials in which different numbers of branches may be actually examined in a pruning search.) We thus looked throughout the brain for regions that correlated with the number of currently available choices, n_0, and tested whether these also depended on the number of choices expected to be available in the next room, in expectation over potential choices at the first step, n_1 (see Methods). Since reaction time was also previously shown to be correlated with these same quantities, in order to rule it out as a confound in neural activity it was included as a nuisance regressor in this analysis, and the variables of interest were orthogonalized against it.

Within the mask of regions significantly correlated with n_0 (totaling 10880 mm^3, 1360 voxels for positive), we found three relevant regions that also correlated positively with n_1 (Fig. 7) (while t-statistics are still uncorrected, reported peak voxels are all significant $p < 0.05$ corrected voxel-wise for FWE over multiple comparisons within the n_0 mask): bilateral precentral cortex (peaks: [−38, 4, 30] $t_{17} = 4.80$, [42, 8, 28] $t_{17} = 4.86$), anterior insula (peaks: [−34, 24, 0] $t_{17} = 6.86$, [38, 22, 6] $t_{17} = 5.79$), and also medial cingulate/SMA in the subset of 15 subjects with coverage in that region (peaks: [−12, 14, 50] $t_{14} = 5.14$, [0, 18, 48] $t_{14} = 5.48$). This indicates that these regions may be participating in a search-based planning process. We performed the same analysis looking for regions negatively correlated with both search difficulty regressors (mask 34656 mm^3, 4332 voxels for negative), and found medial prefrontal cortex (peak: [−2, 46, 0] $t_{17} = 4.93$) and bilateral amygdala/hippocampus (peaks: [−18, −8, −20] $t_{17} = 5.06$, [22, −4, −18] $t_{17} = 5.44$). This region of mPFC is often associated with future value (though the trend toward this correlation did not reach significance in the present study). Nonetheless, BOLD correlations with value cannot explain the present activation, because n_1 has a slight positive correlation with action value, and therefore a value confound would predict a positive correlation with n_1 in BOLD. Instead, we speculate that activity negatively correlated with search difficulty there may relate to assessing the costs of the search process, e.g., for the purpose of deciding whether it is worthwhile to complete the computation (Daw et al. (2005); Rushworth and Behrens (2008); M. Keramati, personal communication). In particular, both this activity and the value responses observed in ventromedial PFC other studies may reflect a top-down assessment for strategy selection based on expected cost, benefit, as well as the uncertainty associated with each approach (Dickinson, 1985; Barraclough et al., 2004; Lee et al., 2004).

Together, these results suggest that using model-based planning to project rewards on the basis of a remembered “map” of a maze employs a broad temporal and frontal network, areas broadly associated with memory and control.

Individual differences

Finally, in order to investigate whether our neural effects were specifically related to planning, or instead to valuation more generally (e.g., in a way that might be common or generic to TD and planning strategies) or even incidentally (e.g., unrelated to choice), we tested whether individual differences in fit quality between the two algorithms to choice behavior covaried across individuals with the strength of the neural responses found in each of these analyses (Hampton et al., 2008). In this way, since subjects varied in the extent to which their behavior was explained by either strategy, we made use of this variability to investigate the behavior's relationship with the neural signaling. In each case, we select the β contrast values from each individual from the identified group peak voxel, and correlate these with either the log likelihoods of the fits to the two algorithms to choices (such that larger, less negative numbers indicate better fits), or the Bayes factor from choices, i.e., the difference between them (so that larger numbers indicate better fits of planning than TD to choices). Of the striatal VOIs responsive to value, the strength of the unbiased value effect and the difference in the right lateral voxel (i.e., the $Q_{plan} + Q_{TD}$ contrast) covaried significantly with the individual subject planning likelihoods ($r(16) = 0.498$, $p = 0.018$), and both this and contrast comparing value predictions ($Q_{plan} > Q_{TD}$) correlated with the analogous Bayes factor between the two in the right medial voxel ($r(16) = 0.552$ and $r(16) = 0.537$, $p < 0.011$), but none with the TD likelihoods ($p > 0.05$). This indicated that the value-related neural effects are also stronger in subjects whose behavior is better explained by planning, consistent with the identification of this activity with planning. Similar results were also seen for the correlates we associate with the representations of the world model. Specifically, we found that the parahippocampal responses to expected next-step reward covaried significantly with the Bayes factor ($r(16) = 0.520$ for right, $r(16) = 0.728$ for left, both $p < 0.014$), and also that the search complexity responses in the SMA covaried with choice likelihoods under planning (for the full set of subjects $r(16) = 0.504$, $p = 0.016$; $r(13) = 0.510$, $p = 0.026$ for the reduced set) but not with the likelihoods under TD ($p > 0.05$). All of these findings support the inference that these signals are related to the decision-making behaviors studied, and are consistent with this activity supporting planning. However, although the lack of a similar relationship with model-free valuations might be interpreted as supporting the specificity of these signals to model-based planning, negative results must be interpreted with caution. For instance, to the extent the task design was successful in precluding the use of TD, it may not elicit meaningful individual differences in the fit of the TD model.

5 Discussion

Computational theories have driven rapid progress quantifying neural signals in the mesostriatal system, primarily in terms of model-free RL (Bar-Gad et al., 2003; Tricomi et al., 2004; O'Doherty et al., 2004; Bayer and Glimcher, 2005; Morris et al., 2006; Schonberg et al., 2007; Hikosaka et al., 2008; Bromberg-Martin and Hikosaka, 2009). Yet it has long been argued that the brain also employs more sophisticated and categorically distinct mechanisms such as cognitive maps (Tolman, 1948; Thistlethwaite, 1951). We extended the theory-driven fMRI approach to model-based planning by leveraging a quantitative characterization of its decision variables to investigate their neural substrates (Daw et al., 2005; Johnson et al., 2007). Having first verified that choices, RTs, and striatal BOLD responses suggest a forward planning mode of valuation in this task | in contrast to broadly successful TD models and their theoretical applicability even to complex spatial tasks (Sutton, 1988; Foster et al., 2000; Stone et al., 2008) | we aimed to map the network implementing such planning by seeking correlates of the theorized construction of these values from reward and state predictions.

To distinguish learning strategies, we elicited ongoing learning via continuous maze reconfiguration. This echoes the logic of Tolman's (1948) demonstrations that rats could plan novel routes following maze changes, and of "place" responses, wherein rats approach previously rewarded locations from novel starting points (Packard and McGaugh, 1996). Rather than such a one-shot challenge (Gläscher et al., 2010; Valentin et al., 2007), we examined how subjects adjust their behavior following many small changes to the maze. This approach separates the valuation strategies partially (producing correlated predictions) but consistently over many trials, and is better suited to test accounts of trial-by-trial learning. We use a normative value iteration algorithm rather than a process-level account to define the quantities of interest, but these quantities should (and largely do according to an analysis of the covariation of regressors derived from different model variants; see supplemental Fig. 3) represent the model-based class more generally. Indeed, some aspects of our results (e.g., effects of search complexity and superior fits for reward-terminated searches) seem to resonate with a process-level tree search model, potentially one involving selective pruning. Nevertheless, more detailed studies will be required to investigate these fine algorithmic distinctions.

Although we adopt a spatial framing, our questions are more akin to previous studies of RL than to other work on navigation. For instance, our focus on behavioral adjustment precludes studying optimal or repeated routes; whereas Yoshida and Ishii (2006) used well-learned behavior in a similar task to study how subjects resolved uncertainty about their location, we used visual cues to minimize locational uncertainty while investigating learning. Also, while the distinction in spatial research between planned navigation to a place and executing a learned response resonates with the algorithmic distinction studied here, much navigational work focuses on another aspect: allocentric vs. egocentric representations (Maguire et al., 1999; Hartley et al., 2003; Burgess, 2006; Iglói et al., 2009). Therefore, although our logic parallels attempts to differentiate two navigational strategies (Doeller et al., 2008), since we do not manipulate location cues or viewpoints, our data only speak to the reference frame distinction insomuch as it coincides (plausibly but not unproblematically; Gallistel and Cramer (1996); Iglói et al. (2010); Weniger et al. (2010)) with the distinction between model-based and model-free learning.

In other areas of learning, a distinction is drawn between one network for habitual, overtrained responses associated with striatum (especially dorsolateral; Knowlton et al. (1996); Packard and Knowlton (2002)), and another, separate or competing, associated with MTL and PFC for planning "goal-directed" responses (Packard and McGaugh, 1996; Burgess et al., 2002; Poldrack and Packard, 2003; Doeller et al., 2008). FMRI studies have shown that value-related BOLD responses in PFC reflect knowledge about higher-order task contingencies, consistent with involvement in model-based reasoning, though not specifically its characteristic use in RL for sequential planning (Hampton et al., 2006; Hampton et al., 2008). In the current task, where behavior suggested valuations were primarily model-based, we found that the same was true even of responses in ventral striatum. These areas have also, in previous studies, been associated with the habit system and TD. Although we cannot directly test this interpretation since we did not detect neural or behavioral evidence for TD in our data, our results together with those others suggest that the two putative systems may be partly overlapping or convergent, with striatum potentially acting as a common value target and locus of action selection (Samejima et al., 2005). This may relate to unit recordings suggesting model-based knowledge in striatum's dopaminergic afferents (Bromberg-Martin and Hikosaka, 2009) and also lesion work in rodents implicating striatum in both place and goal-directed responding, albeit a different, dorsomedial, part (Devan and White, 1999; Yin and Knowlton, 2004; Balleine et al., 2007).

While our striatal results suggest that the substrates of model-based value are more convergent with the purported TD system than might have been suspected, our remaining results, locating antecedents of these values in MTL and frontal cortex, appear more specific to model-based planning. First, effects of anticipated choice set size both on RTs and on BOLD responses offer direct evidence for forward lookahead at choice time. The regions implicated here were primarily more posterior than might have been expected on the basis of work on prefrontal involvement in decision-making (Hampton et al., 2006; Kennerley et al., 2006; Pan et al., 2007), neuroeconomics (McClure et al., 2004; Mushiake et al., 2006), or memory (Poldrack et al., 2001; Poldrack and Packard, 2003): in particular, the contrast revealed posterior frontal regions along the motor cortex. In fact, these areas and SMA in particular have been associated with movement sequencing in other motor tasks (Tanji and Shima, 1994; Hoshi and Tanji, 2004; Lee and Quessy, 2003). The effect in SMA was stronger for subjects whose behavior was better fit by model-based planning (but not so for TD), further indicating that this activity is related to the computations we hypothesize.

Meanwhile, although in other tasks BOLD activity in ventromedial PFC is often found to correlate with expected value (Plassmann et al., 2007; Hare et al., 2008; Wunderlich et al., 2009; Chib et al., 2009), such a correlation did not reach significance in our study. While this may be due to technical limitations (Deichmann et al., 2003), this difference might also relate to the spatial or sequential framing of this task shifting activity to other areas, such as MTL. Although our study does not directly address this possibility, correlations with the next-step reward value in anterior MTL (hippocampus and hippocampal gyrus) are consistent with many other findings indicating that these areas may subserve cognitive maps or spatial associations (O'Keefe, 1990; Maguire et al., 1998; Burgess et al., 2002; Johnson et al., 2007).

Although we have interpreted value correlates in ventral striatum as similar to those seen in studies of putatively model-free RL, in those studies striatal BOLD is more commonly seen to covary with reward prediction error (Pagnoni et al., 2002; McClure et al., 2003; Yacubian et al., 2006; Hampton and O'Doherty, 2007) rather than value (though see Delgado et al. (2000); Tanaka et al. (2004); Tom et al. (2007); Kable and Glimcher (2007); on unit physiology, Arkadir et al. (2004); Samejima et al. (2005); Kim et al. (2009)). Here, we found less robust correlations with a TD prediction error in striatum (analyses not shown); however, the two variables are highly related, and our task was not aimed at distinguishing them (Hare et al., 2008). Notwithstanding that, another possibility is that our task recruits a distinct but anatomically overlapping model-based choice process, which would not be expected to use a TD error signal since model-based valuation constructs values by forward lookahead rather than error-driven learning. This interpretation by no means contradicts the substantial evidence for TD prediction error signals in other tasks.

Overall, we demonstrate dynamic, parametric correlates in various brain areas for a number of previously unstudied decision-related variables, such as one-step reward predictions and search complexities. Should these correlates prove generalizable to future studies, they present a new possibility for investigating specific hypotheses about the details of human valuation, for example using BOLD activity in SMA to track a search process. The model-based approach to understanding the details of value prediction by examining its neural correlates in striatum and vmPFC has had considerable success (e.g., Hampton et al. (2006); Hampton et al. (2008)), and, with these tools, it could be extended to new computational questions and brain areas in order to elucidate further the details of human decision making. Similarly, our demonstration that we can identify these behavioral and neural signatures for model-based, as opposed to more commonly identified model-free, valuation in humans lays the groundwork for future studies of how these two approaches trade off or are controlled. Such information would be relevant to a number of problems of self-control hypothesized to

relate to the compulsive nature of model-free habits, including overeating and drug abuse (Ainslie, 2001; Loewenstein and O'Donoghue, 2004; Everitt and Robbins, 2005).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The project described was supported by Award Number R01MH087882 from the National Institute Of Mental Health as part of the NSF/NIH CRCNS Program, and by a Scholar Award from the McKnight Foundation. The content is solely the responsibility of the authors and does not necessarily represent the offcial views of the National Institute Of Mental Health or the National Institutes of Health.

References

Ainslie, G. Breakdown of will. Cambridge Univ Pr; 2001.

Arkadir D, Morris G, Vaadia E, Bergman H. Independent coding of movement direction and reward prediction by single pallidal neurons. J Neurosci. 2004;24:10047–10056. [PubMed: 15537873]

Balleine, BW.; Daw, ND.; O'Doherty, JP. Multiple forms of value learning and the function of dopamine. In: Glimcher, PW.; Camerer, CF.; Fehr, E.; Poldrack, RA., editors. Neuroeconomics: Decision Making and the Brain. Vol. chapter 24. Academic Press; London: 2008. p. 367-387.

Balleine BW, Delgado MR, Hikosaka O. The role of the dorsal striatum in reward and decision-making. J Neurosci. 2007;27:8161–8165. [PubMed: 17670959]

Balleine BW, Dickinson A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology. 1998;37:407–419. [PubMed: 9704982]

Bar-Gad I, Morris G, Bergman H. Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. Prog Neurobiol. 2003;71:439–473. [PubMed: 15013228]

Barraclough DJ, Conroy ML, Lee D. Prefrontal cortex and decision making in a mixed-strategy game. Nat Neurosci. 2004;7:404–410. [PubMed: 15004564]

Bayer HM, Glimcher PW. Midbrain dopamine neurons encode a quantitative reward pre-diction error signal. Neuron. 2005;47:129–141. [PubMed: 15996553]

Berns GS, McClure SM, Pagnoni G, Montague PR. Predictability modulates human brain response to reward. J Neurosci. 2001;21:2793–2798. [PubMed: 11306631]

Blodgett HC, McCutchan K. Place versus response learning in the simple T-maze. Journal of experimental psychology. 1947;37:412–422. [PubMed: 20267832]

Breiter HC, Aharon I, Kahneman D, Dale A, Shizgal P. Functional imaging of neural responses to expectancy and experience of monetary gains and losses. Neuron. 2001;30:619–639. [PubMed: 11395019]

Bromberg-Martin ES, Hikosaka O. Midbrain dopamine neurons signal preference for advance information about upcoming rewards. Neuron. 2009;63:119–126. [PubMed: 19607797]

Buckner RL, Carroll DC. Self-projection and the brain. Trends Cogn Sci. 2007;11:49–57. [PubMed: 17188554]

Burgess N. Spatial memory: how egocentric and allocentric combine. Trends Cogn Sci. 2006;10:551–557. [PubMed: 17071127]

Burgess N, Maguire EA, O'Keefe J. The human hippocampus and spatial and episodic memory. Neuron. 2002;35:625–641. [PubMed: 12194864]

Camerer C, Ho TH. Experience-weighted attraction learning in normal form games. Econometrica. 1999;67:827–874.

Chib VS, Rangel A, Shimojo S, O'Doherty JP. Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. J Neurosci. 2009;29:12315–12320. [PubMed: 19793990]

Chumbley JR, Friston KJ. False discovery rate revisited: FDR and topological inference using Gaussian random fields. Neuroimage. 2009;44:62–70. [PubMed: 18603449]

Daw, ND. Trial-by-trial data analysis using computational models. In: Phelps, EA.; Robbins, TW.; Delgado, M., editors. Affect, Learning and Decision Making, Attention and Performance XXIII. Oxford University Press; (in press)

Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005;8:1704–1711. [PubMed: 16286932]

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. Nature. 2006;441:876–879. [PubMed: 16778890]

Deichmann R, Gottfried JA, Hutton C, Turner R. Optimized epi for fmri studies of the orbitofrontal cortex. Neuroimage. 2003;19:430–441. [PubMed: 12814592]

Delgado MR, Locke HM, Stenger VA, Fiez JA. Dorsal striatum responses to reward and punishment: Effects of valence and magnitude manipulations. Cogn Affect Behav Neurosci. 2003;3:27–38. [PubMed: 12822596]

Delgado MR, Nystrom LE, Fissell C, Noll DC, Fiez JA. Tracking the hemodynamic responses to reward and punishment in the striatum. J Neurophysiol. 2000;84:3072–3077. [PubMed: 11110834]

Devan BD, White NM. Parallel information processing in the dorsal striatum: Relation to hippocampal function. J Neurosci. 1999;19:2789–2798. [PubMed: 10087090]

Dickinson A. Actions and habits: The development of behavioural autonomy. Philos Trans R Soc Lond B Biol Sci. 1985;308:67–78.

Dickinson, A.; Balleine, B. The Role of Learning in the Operation of Motivational Systems. John Wiley & Sons, Inc; 2002.

Doeller CF, King JA, Burgess N. Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. Proc Natl Acad Sci. 2008;105:5915–5920. [PubMed: 18408152]

Doya K. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Networks. 1999;12:961–974. [PubMed: 12662639]

Everitt BJ, Robbins TW. Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nat Neurosci. 2005;8:1481–1489. [PubMed: 16251991]

Foster D, Morris R, Dayan P. A model of hippocampally dependent navigation, using the temporal difference learning rule. Hippocampus. 2000;10:1–16. [PubMed: 10706212]

Gallistel CR, Cramer AE. Computations on metric maps in mammals: getting oriented and choosing a multi-destination route. J Exp Biol. 1996;199:211–217. [PubMed: 8576692]

Gläscher J, Daw ND, Dayan P, O'Doherty JP. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron. 2010;66:585–595. [PubMed: 20510862]

Gläscher J, Hampton AN, O'Doherty JP. Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. Cereb Cortex. 2009;19:483–495. [PubMed: 18550593]

Hampton AN, Bossaerts P, O'Doherty JP. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J Neurosci. 2006;26:8360–8367. [PubMed: 16899731]

Hampton AN, Bossaerts P, O'Doherty JP. Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc Natl Acad Sci. 2008;105:6741–6746. [PubMed: 18427116]

Hampton AN, O'Doherty JP. Decoding the neural substrates of reward-related decision making with functional MRI. Proc Natl Acad Sci. 2007;104:1377–1382. [PubMed: 17227855]

Hare TA, O'Doherty JP, Camerer CF, Schultz W, Rangel A. Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. J Neurosci. 2008;28:5623–5630. [PubMed: 18509023]

Hartley T, Maguire EA, Spiers HJ, Burgess N. The well-worn route and the path less traveled: Distinct neural bases of route following and wayfinding in humans. Neuron. 2003;37:877–888. [PubMed: 12628177]

Hasselmo ME. A model of prefrontal cortical mechanisms for goal-directed behavior. J Cogn Neurosci. 2005;17:1115–1129. [PubMed: 16102240]

Hasselmo ME. A model of episodic memory: Mental time travel along encoded trajectories using grid cells. Neurobiol Learn Mem. 2009;92:559–573. [PubMed: 19615456]

Henson R. What can functional neuroimaging tell the experimental psychologist? The Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology. 2005;58:193–233.

Hikosaka O, Bromberg-Martin E, Hong S, Matsumoto M. New insights on the subcortical representation of reward. Curr Opin Neurobiol. 2008;18:203–208. [PubMed: 18674617]

Holmes, AP.; Friston, KJ. Generalisability, random effects & population inference. Neuroimage; 4th International Conference on Functional Mapping of the Human Brain; Montreal, Quebec, Canada. 1998. p. S754

Hoshi E, Tanji J. Differential roles of neuronal activity in the supplementary and pre-supplementary motor areas: From information retrieval to motor planning and execution. J Neurophysiol. 2004;92:3482–3499. [PubMed: 15269227]

Houk, JC.; Adams, JL.; Barto, AG. A model of how the basal ganglia generate and use neural signals that predict reinforcement. In: Houk, JC.; Davis, JL.; Beiser, DG., editors. Models of information processing in the basal ganglia. Vol. chapter 13. MIT Press; 1994. p. 249-270.

Iglói K, Zaoui M, Berthoz A, Rondi-Reig L. Sequential egocentric strategy is acquired as early as allocentric strategy: Parallel acquisition of these two navigation strategies. Hippocampus. 2009;19:1199–1211. [PubMed: 19360853]

Iglói K, Doeller CF, Berthoz A, Rondi-Reig L, Burgess N. Lateralized human hippocampal activity predicts navigation based on sequence or place memory. Proc Natl Acad Sci. 2010;107:14466–14471. [PubMed: 20660746]

Johnson A, van der Meer MAA, Redish AD. Integrating hippocampus and striatum in decision-making. Curr Opin Neurobiol. 2007;17:692–697. [PubMed: 18313289]

Kable JW, Glimcher PW. The neural correlates of subjective value during intertemporal choice. Nat Neurosci. 2007;10:1625–1633. [PubMed: 17982449]

Kass RE, Raftery AE. Bayes factors. J Amer Stat Assoc. 1995;90:773–795.

Kennerley S, Walton M, Behrens T, Buckley M, Rushworth M. Optimal decision making and the anterior cingulate cortex. Nat Neurosci. 2006;9:940–947. [PubMed: 16783368]

Kim H, Sul JH, Huh N, Lee D, Jung MW. Role of striatum in updating values of chosen actions. J Neurosci. 2009;29:14701–14712. [PubMed: 19940165]

Knowlton BJ, Mangels JA, Squire LR. A neostriatal habit learning system in humans. Science. 1996;273:1399–1402. [PubMed: 8703077]

Knutson B, Fong GW, Adams CM, Varner JL, Hommer D. Dissociation of reward anticipation and outcome with event-related fmri. Neuroreport. 2001;12:3683–3687. [PubMed: 11726774]

Knutson B, Taylor J, Kaufman M, Peterson R, Glover G. Distributed neural representation of expected value. J Neurosci. 2005;25:4806–4812. [PubMed: 15888656]

Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci. 2009;12:535–540. [PubMed: 19396166]

Lee D, Conroy ML, McGreevy BP, Barraclough DJ. Reinforcement learning and decision making in monkeys during a competitive game. Cogn Brain Res. 2004;22:45–58.

Lee D, Quessy S. Activity in the supplementary motor area related to learning and performance during a sequential visuomotor task. J Neurophysiol. 2003;89:1039–1056. [PubMed: 12574479]

Loewenstein, G.; O'Donoghue, T. Working Papers 04–14. Cornell University, Center for Analytic Economics; 2004. Animal spirits: Affective and deliberative processes in economic behavior.

Lohrenz T, McCabe K, Camerer CF, Montague PR. Neural signature of fictive learning signals in a sequential investment task. Proc Natl Acad Sci. 2007;104:9493–9498. [PubMed: 17519340]

Maguire EA, Burgess N, O'Keefe J. Human spatial navigation: cognitive maps, sexual dimorphism, and neural substrates. Curr Opin Neurobiol. 1999;9:171–177. [PubMed: 10322179]

Maguire EA, Burgess N, Donnett JG, Frackowiak RSJ, Frith CD, O'Keefe J. Knowing where and getting there: A human navigation network. Science. 1998;280:921–924. [PubMed: 9572740]

Maldjian JA, Laurienti PJ, Burdette JB, Kraft RA. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage. 2003;19:1233–1239. [PubMed: 12880848]

McClure SM, Berns GS, Montague PR. Temporal prediction errors in a passive learning task activate human striatum. Neuron. 2003;38:339–346. [PubMed: 12718866]

McClure SM, Laibson DI, Loewenstein G, Cohen JD. Separate neural systems value immediate and delayed monetary rewards. Science. 2004;306:503–507. [PubMed: 15486304]

Montague PR, King-Casas B, Cohen JD. Imaging valuation models in human choice. Annu Rev Neurosci. 2006;29:417–448. [PubMed: 16776592]

Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H. Midbrain dopamine neurons encode decisions for future action. Nat Neurosci. 2006;9:1057–1063. [PubMed: 16862149]

Mushiake H, Saito N, Sakamoto K, Itoyama Y, Tanji J. Activity in the lateral prefrontal cortex reflects multiple steps of future events in action plans. Neuron. 2006;50:631–641. [PubMed: 16701212]

O'Doherty JP, Hampton A, Kim H. Model-based fmri and its application to reward learning and decision making. Ann N Y Acad Sci. 2007;1104:35–53. [PubMed: 17416921]

O'Doherty JP. Reward representations and reward-related learning in the human brain: insights from neuroimaging. Curr Opin Neurobiol. 2004;14:769–776. [PubMed: 15582382]

O'Doherty JP, Buchanan TW, Seymour B, Dolan RJ. Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum. Neuron. 2006;49:157–166. [PubMed: 16387647]

O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. Temporal difference models and reward-related learning in the human brain. Neuron. 2003;38:329–337. [PubMed: 12718865]

O'Doherty JP, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science. 2004;304:452–454. [PubMed: 15087550]

O'Doherty JP, Deichmann R, Critchley HD, Dolan RJ. Neural responses during anticipation of a primary taste reward. Neuron. 2002;33:815–826. [PubMed: 11879657]

O'Keefe J. A computational theory of the hippocampal cognitive map. Prog Brain Res. 1990;83:301–312. [PubMed: 2203101]

Packard MG, Knowlton BJ. Learning and memory functions of the basal ganglia. Annu Rev Neurosci. 2002;25:563–593. [PubMed: 12052921]

Packard MG, McGaugh JL. Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. Neurobiol Learn Mem. 1996;65:65–72. [PubMed: 8673408]

Pagnoni G, Zink CF, Montague PR, Berns GS. Activity in human ventral striatum locked to errors of reward prediction. Nat Neurosci. 2002;5:97–98. [PubMed: 11802175]

Pan X, Sawa K, Sakagami M. Model-based reward prediction in the primate prefrontal cortex. Neurosci Res. 2007;58:S229–S229.

Plassmann H, O'Doherty J, Rangel A. Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. J Neurosci. 2007;27:9984–9988. [PubMed: 17855612]

Poldrack RA, Clark J, Par-Blagoev EJ, Shohamy D, Creso Moyano J, Myers C, Gluck MA. Interactive memory systems in the human brain. Nature. 2001;414:546–550.

Poldrack RA, Packard MG. Competition among multiple memory systems: converging evidence from animal and human brain studies. Neuropsychologia. 2003;41:245–251. [PubMed: 12457750]

Rushworth MFS, Behrens TEJ. Choice, uncertainty and value in prefrontal and cingulate cortex. Nat Neurosci. 2008;11:389–397. [PubMed: 18368045]

Samejima, K.; Doya, K. Neural Information Processing. Springer; Berlin: 2008. Estimating internal variables of a decision maker's brain: A model-based approach for neuroscience; p. 596-603.

Samejima K, Ueda Y, Doya K, Kimura M. Representation of action-specific reward values in the striatum. Science. 2005;310:1337–1340. [PubMed: 16311337]

Schonberg T, Daw ND, Joel D, O'Doherty JP. Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. J Neurosci. 2007;27:12860–12867. [PubMed: 18032658]

Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997;275:1593–1599. [PubMed: 9054347]

Schwarz G. Estimating the dimension of a model. The Annals of Statistics. 1978;6:461–464.

Seymour B, Daw N, Dayan P, Singer T, Dolan R. Differential encoding of losses and gains in the human striatum. J Neurosci. 2007;27:4826–4831. [PubMed: 17475790]

Seymour B, O'Doherty JP, Dayan P, Koltzenburg M, Jones AK, Dolan RJ, Friston KJ, Frack-owiak RS. Temporal difference models describe higher-order learning in humans. Nature. 2004;429:664–667.

Smith A, Li M, Becker S, Kapur S. A model of antipsychotic action in conditioned avoidance: A computational approach. Neuropsychopharmacology. 2004;29:1040–1049. [PubMed: 14997176]

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. Neuroimage. 2009;46:1004–1017. [PubMed: 19306932]

Stone, EE.; Skubic, M.; Keller, JM. Adaptive temporal difference learning of spatial memory in the water maze task. 7th IEEE International Conference on Development and Learning; 2008. p. 85-90.

Suri RE, Schultz W. Temporal difference model reproduces anticipatory neural activity. Neural Computation. 2001;13:841–862. [PubMed: 11255572]

Sutton RS. Learning to predict by the methods of temporal differences. Mach Learn. 1988;3:9–44.

Sutton, RS.; Barto, AG. Reinforcement Learning. MIT Press; Cambridge, MA: 1998.

Sutton, RS.; Pinette, B. The learning of world models by connectionist networks. Proceedings of the seventh annual conference of the cognitive science society; Citeseer. 1985. p. 54-64.

Tanaka SC, Doya K, Okada G, Ueda K, Okamoto Y, Yamawaki S. Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. Nat Neuosci. 2004;7:887–893.

Tanji J, Shima K. Role for supplementary motor area cells in planning several movements ahead. Nature. 1994;371:413–416. [PubMed: 8090219]

Thistlethwaite D. A critical review of latent learning and related experiments. Psychol Bull. 1951;48:97–129. [PubMed: 14834282]

Thorndike, EL. Animal intelligence: An Experimental Study of the Associate Processes in Animals. Macmillan; New York: 1911. p. 29-58.

Tolman E. Cognitive maps in rats and men. Psychol Rev. 1948;55:189–208. [PubMed: 18870876]

Tom SM, Fox CR, Trepel C, Poldrack RA. The neural basis of loss aversion in decision-making under risk. Science. 2007;315:515–518. [PubMed: 17255512]

Tricomi EM, Delgado MR, Fiez JA. Modulation of caudate activity by action contingency. Neuron. 2004;41:281–292. [PubMed: 14741108]

Valentin VV, Dickinson A, O'Doherty JP. Determining the neural substrates of goal-directed learning in the human brain. J Neurosci. 2007;27:4019–4026. [PubMed: 17428979]

Watkins, CJCH. PhD diss. Cambridge Univ; 1989. Learning from delayed rewards.

Weniger G, Siemerkus J, Schmidt-Samoa C, Mehlitz M, Baudewig J, Dechent P, Irle E. The human parahippocampal cortex subserves egocentric spatial learning during navigation in a virtual maze. Neurobiol Learn Mem. 2010;93:46–55. [PubMed: 19683063]

Wittmann BC, Daw ND, Seymour B, Dolan RJ. Striatal activity underlies novelty-based choice in humans. Neuron. 2008;58:967–973. [PubMed: 18579085]

Wunderlich K, Rangel A, O'Doherty JP. Neural computations underlying action-based decision making in the human brain. Proc Natl Acad Sci. 2009;106:17199–17204. [PubMed: 19805082]

Yacubian J, Gläscher J, Schroeder K, Sommer T, Braus DF, Buchel C. Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. J Neurosci. 2006;26:9530–9537. [PubMed: 16971537]

Yin HH, Knowlton BJ. Contributions of striatal subregions to place and response learning. Learning & Memory. 2004;11:459–463.

Yoshida W, Ishii S. Resolution of uncertainty in prefrontal cortex. Neuron. 2006;50:781–789. [PubMed: 16731515]

Zilli EA, Hasselmo ME. Modeling the role of working memory and episodic memory in behavioral tasks. Hippocampus. 2008;18:193–209. [PubMed: 17979198]

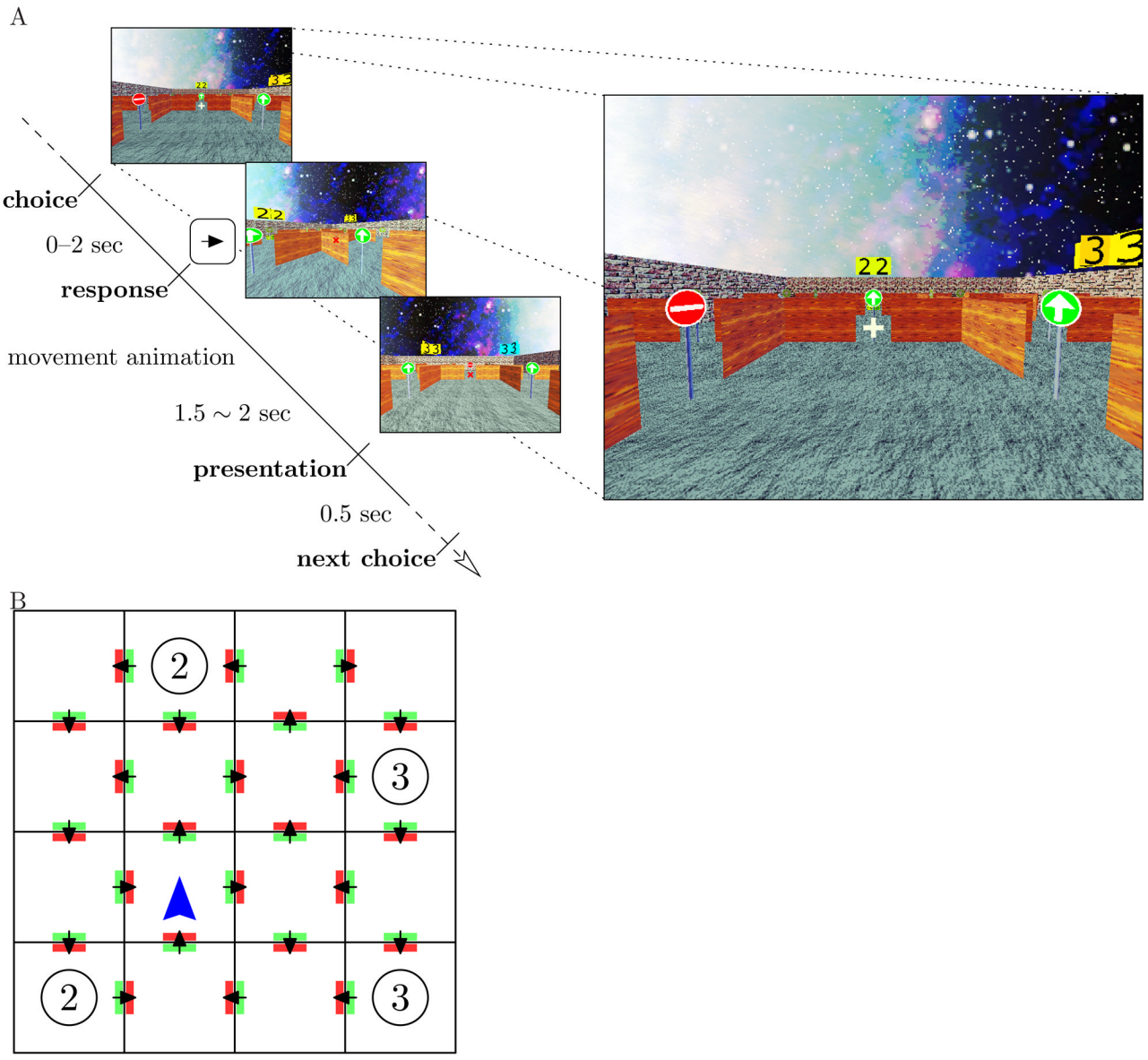

Figure 1. Task flow and example state

(A) Subjects were cued to choose a direction by pressing a key. If the subject did not respond within 2 seconds, she lost a turn and was again presented with the same choice (no movement). Otherwise, an animation was shown moving to the room in the selected direction (or to a random room for randomly occurring "jumps"); this movement lasted 1.5-2 seconds jittered uniformly. Then, the next room was presented, including the available transitions from that room and any received reward. Finally, after 0.5 seconds the subject was cued to make the next decision. Only the doors in the current room were visible to the subject. (B) A possible abstract layout of the task, where each square represents a room, and each arrow represents an available door direction the subject may choose from. Circles represent reward locations, where the subject would gain the indicated reward value each time the room was visited. At each step, each one-way door could flip direction independently with probability $\frac{1}{24}$.

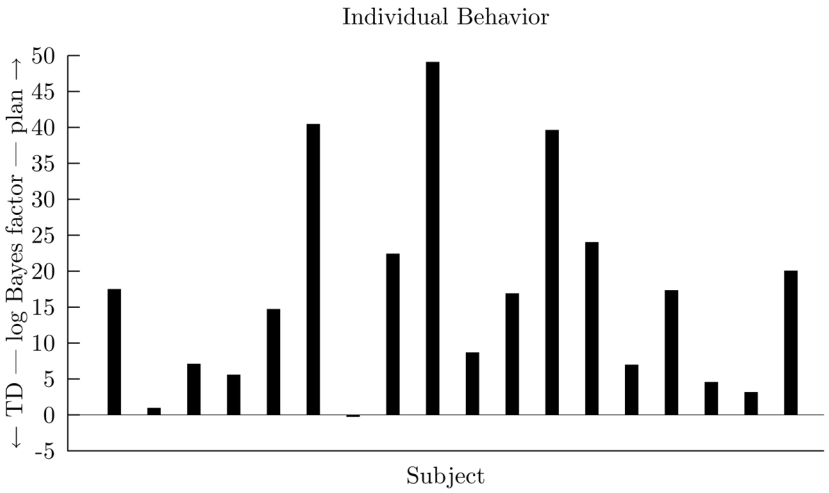

Figure 2. Behavioral model likelihood comparison

Negative log likelihood evidence values under BIC. Per-subject log Bayes factors comparing planning against TD.

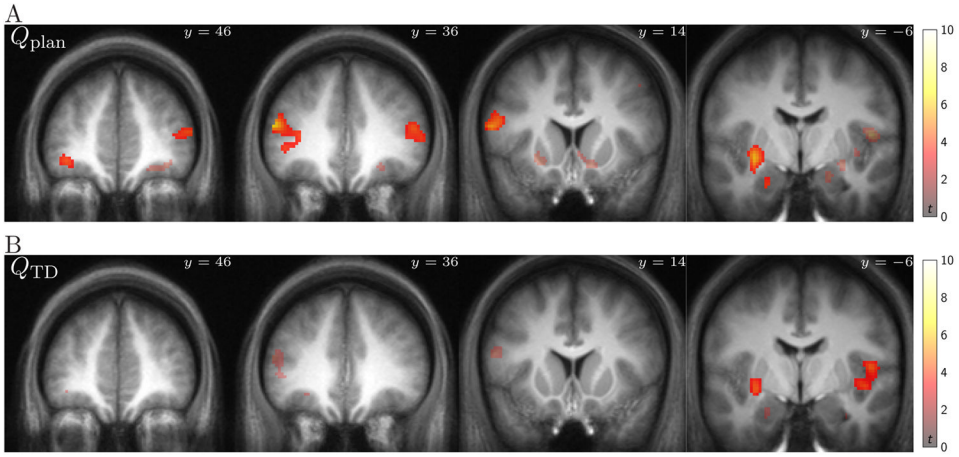

Figure 3. Value-responsive areas

T-statistic map of group response size to (A) planned and (B) TD-based value predictions from separate models (shown at $p < 0.001$ uncorrected, significant $p < 0.05$ FDR clusters highlighted).

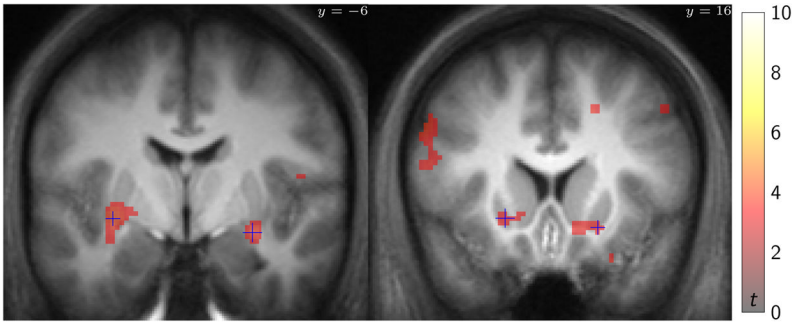

Figure 4. Identification of value-related voxels of interest

T-statistic map of group response size to either planned or TD-based value predictions (summed contrast, shown at $p < 0.001$ uncorrected, significance not assessed). The most responsive peak voxels of this map anatomically within striatum were identified for further analysis.

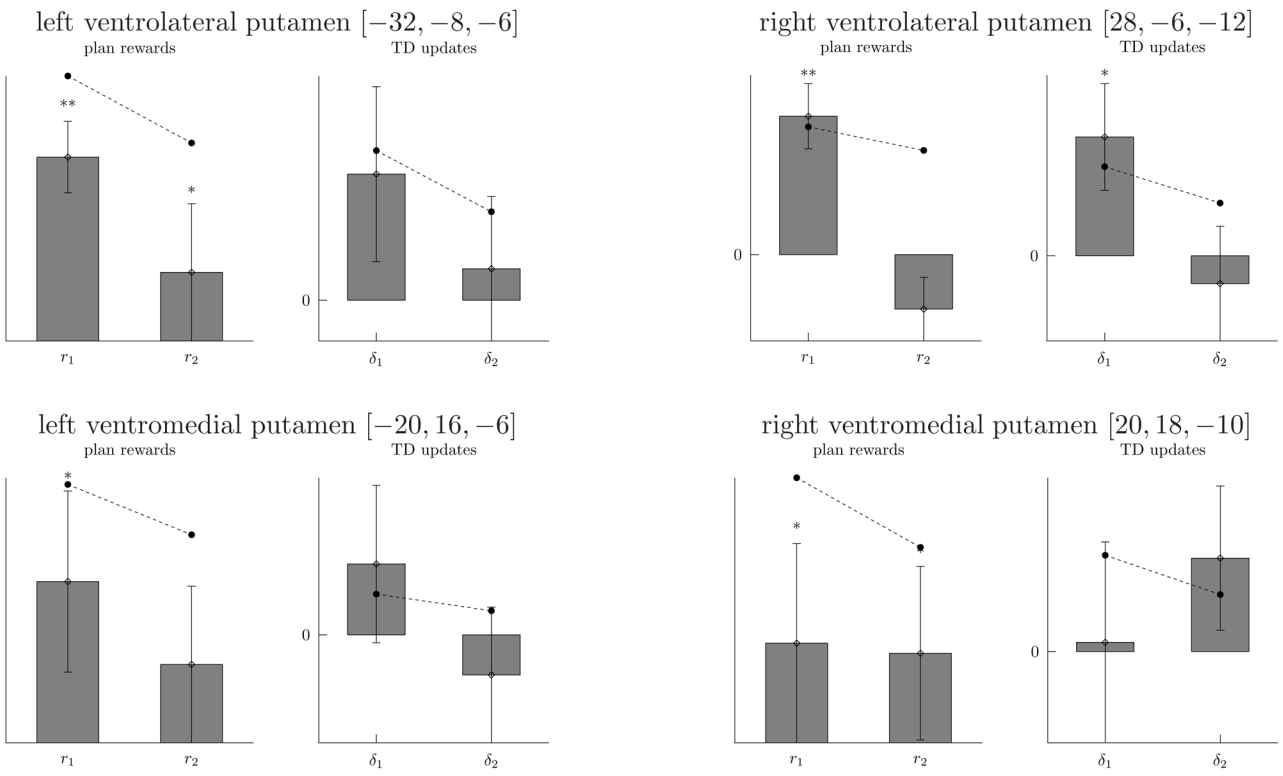

Figure 5. Striatal BOLD responses to partial value components

Responses to key components of the value predictions as predicted by the two algorithms in the previously- identified voxels of interest (VOIs). Also shown are the predicted responses from the overall value fit assuming exponential discounting and updating. Note that significance s are biased by voxel selection.

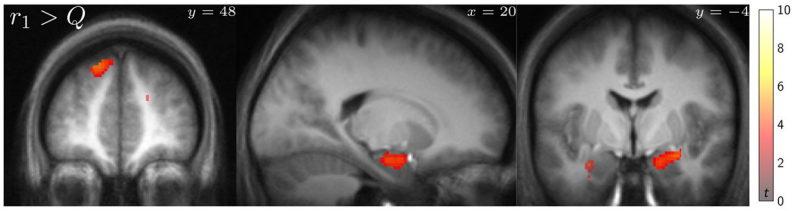

Figure 6. Responses to predicted next-step rewards beyond chosen values

T-statistic map of responsive regions to choices that are expected to lead to a reward room (r_1), greater than the first two terms of the value equation ($r_1 + \gamma r_2$, shown at $p < 0.001$ uncorrected; significant $p < 0.05$ FDR clusters highlighted).

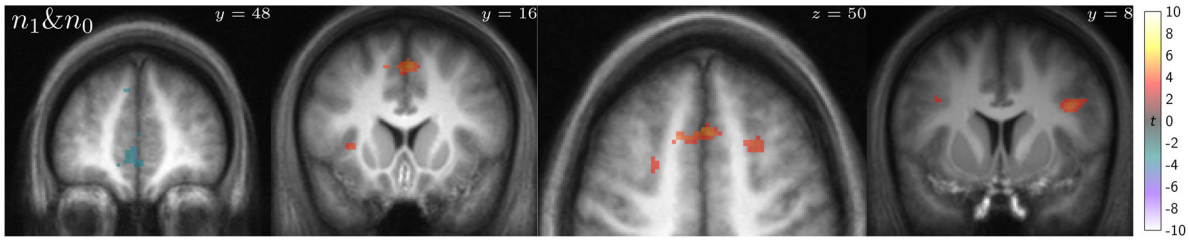

Figure 7. Response to both one-step predicted and immediate choice count

Masked T-statistic map of responses to expected next-step choice set size within regions responsive to current choice set size (all n_0 significant $p < 0.05$ FDR cluster-size; n_1 shown at $p < 0.001$ uncorrected; two-tailed).

Table 1
Distribution of subjects' individual maximum likelihoods and parameter estimates

Quartiles (**medians**) of best-fitting parameters for the two algorithms used to produce regressors for imaging analysis, along with negative log likelihood (NLL), BIC estimated evidence, and pseudo-r^2 measures of individual fit quality.

	Plan	TD	Random
β	4.081,**11.78**,17.14	3.405, **5.315**, 6.841	
γ	0.461,**0.816**,0.861	0.550, **0.861**, 0.936	
η	0.058,**0.142**,0.516		
Q_0		1.260, **2.883**, 6.774	
α		0.319, **0.408**, 0.579	
λ		0.565, **0.756**, 0.915	
NLL	131.1,**156.3**,195.8	152.7, **171.3**, 207.5	214.7, **224.5**, 237.0
BIC	139.6,**165.0**,204.5	167.1, **185.6**, 222.0	214.7, **224.5**, 237.0
ρ^2	.141,**.279**, .391	.130, **.226**, .321	(0)