EJN fens European Journal of Neuroscience

European Journal of Neuroscience, Vol. 35, pp. 1092–1104, 2012 doi:10.1111/j.1460-9568.2012.08017.x

Generalization of value in reinforcement learning by humans

G. Elliott Wimmer,[1] Nathaniel D. Daw[2,*] and Daphna Shohamy[1,*]
[1]Department of Psychology, Columbia University, 1190 Amsterdam Ave., 406 Schermerhorn Hall MC5501, New York, NY, USA
[2]Center for Neural Science & Department of Psychology, New York University, 4 Washington Place, Room 809, NY, USA

Keywords: computational model, hippocampus, memory, reward, ventral striatum

Abstract

Research in decision-making has focused on the role of dopamine and its striatal targets in guiding choices via learned stimulus–reward or stimulus–response associations, behavior that is well described by reinforcement learning theories. However, basic reinforcement learning is relatively limited in scope and does not explain how learning about stimulus regularities or relations may guide decision-making. A candidate mechanism for this type of learning comes from the domain of memory, which has highlighted a role for the hippocampus in learning of stimulus–stimulus relations, typically dissociated from the role of the striatum in stimulus–response learning. Here, we used functional magnetic resonance imaging and computational model-based analyses to examine the joint contributions of these mechanisms to reinforcement learning. Humans performed a reinforcement learning task with added relational structure, modeled after tasks used to isolate hippocampal contributions to memory. On each trial participants chose one of four options, but the reward probabilities for pairs of options were correlated across trials. This (uninstructed) relationship between pairs of options potentially enabled an observer to learn about option values based on experience with the other options and to generalize across them. We observed blood oxygen level-dependent (BOLD) activity related to learning in the striatum and also in the hippocampus. By comparing a basic reinforcement learning model to one augmented to allow feedback to generalize between correlated options, we tested whether choice behavior and BOLD activity were influenced by the opportunity to generalize across correlated options. Although such generalization goes beyond standard computational accounts of reinforcement learning and striatal BOLD, both choices and striatal BOLD activity were better explained by the augmented model. Consistent with the hypothesized role for the hippocampus in this generalization, functional connectivity between the ventral striatum and hippocampus was modulated, across participants, by the ability of the augmented model to capture participants' choice. Our results thus point toward an interactive model in which striatal reinforcement learning systems may employ relational representations typically associated with the hippocampus.

Introduction

Research in decision-making posits a computational role for the dopamine system and its striatal targets in guiding choices via learned stimulus–reward or stimulus–response associations (Houk *et al.*, 1995; Schultz *et al.*, 1997; Frank *et al.*, 2004; Everitt & Robbins, 2005; Daw & Doya, 2006; Schultz, 2006; Rangel *et al.*, 2008). However, there has been increasing recognition that this narrow mechanism for 'habit' learning cannot explain the full diversity of choice behavior, or even the contribution of striatum to it (Balleine *et al.*, 2008; Rangel *et al.*, 2008; Redish *et al.*, 2008). Still, it remains less precisely understood how other forms of learning, possibly involving distinct cognitive and neural systems, contribute to choice (Doya, 1999; Daw *et al.*, 2005; Daw & Shohamy, 2008).

One promising avenue for addressing this gap is the largely separate domain of memory research, where a finely detailed distinction between different forms of learning has long been established (Schacter, 1990; Squire, 1992; Knowlton *et al.*, 1996; Gabrieli, 1998; Eichenbaum & Cohen, 2001). Perhaps the best-characterized system is that for episodic memory, associated with the hippocampus and operationally distinguished from a striatal habit system (Schacter, 1990; Squire, 1992; Knowlton *et al.*, 1996; Gabrieli, 1998; Eichenbaum & Cohen, 2001; Poldrack *et al.*, 2001; Hartley & Burgess, 2005; Foerde *et al.*, 2006; Mattfeld & Stark, 2011). Echoing non-habitual accounts of decision-making, hippocampal memories represent the relation between multiple arbitrarily associated stimuli. Due to their relational nature, hippocampal memories are also flexible and can be generalized across stimuli and contexts (Cohen & Eichenbaum, 1993; Dusek & Eichenbaum, 1997; Eichenbaum & Cohen, 2001; Davachi, 2006; Shohamy *et al.*, 2008; Staresina & Davachi, 2009).

In memory tasks, the relational hallmark of hippocampal memories has been demonstrated using procedures that first embed relations among stimuli and then probe whether later choices reflect relational knowledge (Dusek & Eichenbaum, 1997; Myers *et al.*, 2003; Preston *et al.*, 2004; Greene *et al.*, 2006; Shohamy *et al.*, 2006; Shohamy & Wagner, 2008; Zeithamova & Preston, 2010). For example, in 'acquired equivalence', people first learn that stimulus A is associated

Correspondence: Dr D. Shohamy, as above.
E-mail: shohamy@psych.columbia.edu

*These authors contributed equally to this work and ordering was determined arbitrarily.

Received 14 October 2011, revised 21 December 2011, accepted 23 December 2011

with outcome X, and that stimulus B is also associated with outcome X. Having indirectly learned that A and B are related, in terms of their common outcome X, people later transfer additional knowledge about stimulus A to stimulus B, presumably based on the learned 'equivalence' between them. Converging evidence suggests that acquired equivalence depends on the hippocampus and surrounding medial-temporal lobe cortex (e.g. Coutureau *et al.*, 2002; Myers *et al.*, 2003; Shohamy & Wagner, 2008).

Here, we sought to use this approach in the context of a reinforcement learning task to determine whether relational encoding contributes to decision-making. Participants made repeated choices in a reward learning task, in which the probability of reward associated with each of four options diffused randomly. Structured relationships between option outcomes were embedded via correlated reward probabilities between pairs of options, creating an (uninstructed) equivalence between them (Fig. 1). Thus, this task incorporates one of the essential elements of 'acquired equivalence' (Honey & Hall, 1989; Myers *et al.*, 2003; Shohamy & Wagner, 2008), namely that pairs of options are related by virtue of sharing a common outcome, enabling (if this structure is detected and encoded) generalization of subsequent learning between them. However, in contrast to studies in the memory domain, the common outcome here is a correlated likelihood of reward, rather than a particular stimulus. Moreover, this correlational structure is embedded within a trial-and-error reward learning task, allowing us to ask whether and how inferred similarity relationships of this kind can affect instrumental choice behavior. Importantly, standard reinforcement models [ranging from Thorndike's (1911) law of effect to more modern TD rules (e.g. Schultz *et al.*, 1997)] should in principle be entirely blind to this kind of structure.

We characterized learning behavior using reinforcement learning models in order to measure the extent to which choices are driven by the correlational structure across option values. We then used functional magnetic resonance imaging (fMRI) to identify regions of the brain where activation covaried with decision variables from the models, to investigate whether the inclusion of this structure implicated the hippocampus instead of (or in addition to) traditional reinforcement learning activations in the striatum. Critically, we could then examine these signals to test whether they reflected value generalization, and specifically whether ventral striatal blood oxygen level-dependent (BOLD) activity reflected relational knowledge. Finally, we used multivariate analyses of the fMRI data to examine whether the use of such structure to guide choices might be reflected in increased functional connectivity between the hippocampus and the striatum.

Materials and methods

Participants

Twenty-four right-handed fluent English speakers with normal or corrected-to-normal vision participated in the study. All participants were free of neurological or psychiatric disorders and fully consented to participate. Informed consent was obtained in a manner approved by the New York University Committee on Activities Involving Human Subjects. Three participants' data were excluded: two due to software problems (one for a partial loss of behavioral responses, one for missing timing information), and a third because the participant elected to leave the experiment before the completion of data acquisition. Behavioral and functional imaging data are presented from the remaining 21 participants (mean age, 19.3 years; range, 18–28; ten females). Participants were paid $20 per hour for the approximately 2-h duration of participation plus one-fifth of the nominal rewards the participant earned in the experimental task.

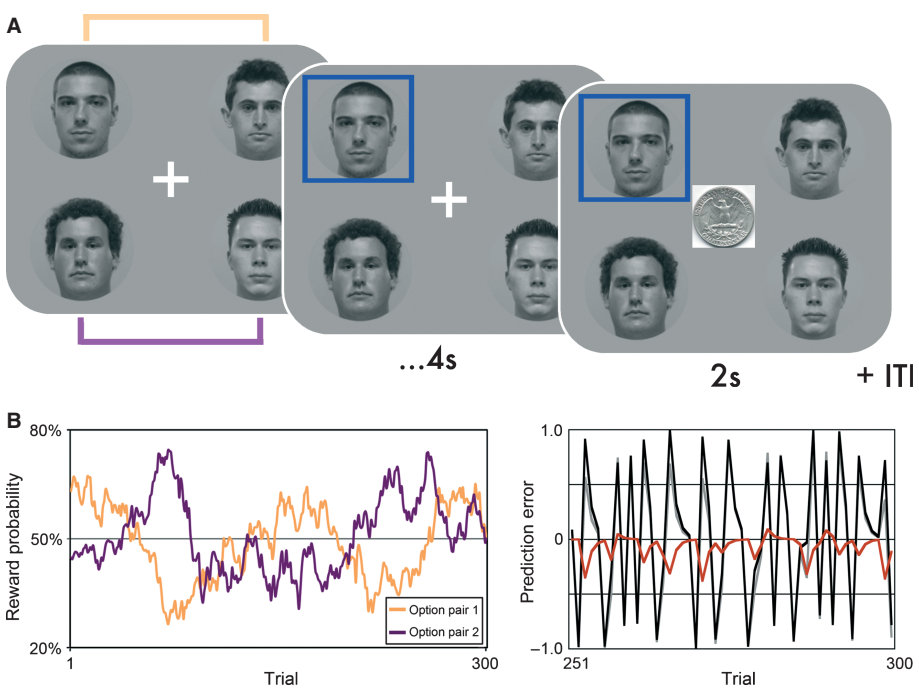

FIG. 1. Design of the reward equivalence paradigm. (A) On each trial, participants chose one of four face options. After a delay, the outcome ($0.25 or $0.00) was revealed. In colored brackets, one example of option pairing is indicated. (B) Drifting reward probability distribution defining the reward equivalence for one example pairing (left). Trial-by-trial reinforcement learning variables for 50 trials from an example participant – fMRI model regressors for prediction error (black) and prediction error difference due to generalization (red), and an illustration of a full generalization model prediction error (grey) (right).

Task

In the experimental task (Fig. 1a; Daw & Shohamy, 2008), on each of 300 trials, participants chose one of four presented face stimuli and then received monetary feedback. This reinforcement learning task is a variant of a 'four-armed bandit' task (Daw *et al.*, 2006; Wittmann *et al.*, 2008). The face stimuli, which were constant across trials and participants, were taken from the Stanford Face Database. The location of the faces was permuted randomly from trial to trial.

On each trial, participants had 2 s to choose between the four options (Fig. 1a), using an MR-compatible button response pad held in the right hand. After the participant made a selection and until the end of the choice period, the selected option was framed in blue and the unchosen options were decreased in brightness. Participants then received binary reward feedback for 2 s, a \$0.25 'win' outcome represented by an image of a quarter-dollar and a \$0.00 'miss' outcome represented by a phase-scrambled image of a quarter-dollar (Fig. 1a). If no choice was recorded during the choice period, no reward outcome was displayed and the face options remained on the screen until the end of the trial. Trials were intermixed with variable duration inter-trial fixation null events (ITI; mean 2 s, range 0–12 s). The total time allotted for null events was equal to one-third of the scan time. The duration and distribution of null events was optimized for estimation of rapid event-related fMRI responses as calculated using OPTSEQ software (http://surfer.nmr.mgh.harvard.edu/optseq/). The task was presented using the Psychophysics Toolbox (Brainard, 1997) and projected onto a mirror screen above the participant's eyes.

Participants were instructed that each face option was associated with a different probability of reward, that these probabilities could change slowly, and that their goal was to attempt to find the most rewarding option at a given time in order to earn the most money. They were also instructed that rewards were tied to the face identity and not the face position. Prior to the scanning session, participants completed a short practice version to familiarize them with the task and to ensure that their button responses reflected their intended choices.

Each of the options (S_1–S_4) was associated with a different probability of monetary reward. Across the 300 trials in the experiment, the reward probabilities diffused gradually according to Gaussian random walks, so as to encourage continual learning. Unbeknownst to the participants, to provide the opportunity of encoding stimulus–stimulus relational structure, the faces were grouped into equivalent pairs (here referred to as faces S_1 & S_3 and S_2 & S_4). The chance of reward on choosing S_1 or S_3 (and similarly S_2 or S_4) was the same on any particular trial; however, trial feedback only displayed the reward outcome for the selected face. The reward probability for each pair of face stimuli changed over time, diffusing between 25 and 75% according to Gaussian random walks with reflecting boundary conditions. Two instantiations of two sets of random walks were generated, and these were then inverted (i.e. subtracting all probabilities from 100%) to give a total of four sequences (Fig. 1b). To ensure that these strong positive correlations did not make the choice problem trivial (i.e. with all four options often having roughly the same value), a more modest negative correlation was included between the two sets of walks within each of these sequences (r^2 between pairs, −0.135 and −0.369; vs. $r^2 = 1$ within paired options). Reward probability sequences were counterbalanced between participants, as was the mapping of particular face stimuli to the underlying reward sequences.

After the completion of scanning, participants answered a series of questions that assessed their strategies during learning and their awareness of the contingencies across options. To further probe any knowledge of the underlying task structure provided by the equivalence relationships, participants were also given a questionnaire that included pictures of the four face stimuli. Participants were instructed to draw lines connecting the pairs of stimuli that for any reason seemed related to one another, and then to describe why they paired those options together (data available for 18 participants). Participants were then informed how much money they had won in the experiment.

Imaging procedure

Whole-brain imaging was conducted on a 3.0-T Siemens Allegra head-only MRI system at NYU's Center for Brain Imaging, using a Nova Medical NM-011 head coil. Head padding was used to minimize head motion; subsequent inspection showed that no participant's motion exceeded 2 mm in any direction from one volume acquisition to the next. Structural images were collected using a high-resolution T1-weighted MPRAGE pulse sequence ($1 \times 1 \times 1$ mm voxel size). Functional images were collected using a gradient echo T2*-weighted echoplanar (EPI) sequence with BOLD contrast (TR = 2000 ms, TE = 15 ms, flip angle = 82, $3 \times 3 \times 3$ mm voxel size; 33 contiguous oblique-axial slices), tilted on a per-participant basis approximately 23° off of the AC–PC axis to optimize sensitivity to signal in the orbitofrontal cortex and the medial temporal lobe (Deichmann *et al.*, 2003). The task was scanned in four blocks each of 310 volumes (10 min 20 s). For each functional scanning block, four discarded volumes were collected prior to the first trial to allow for magnetic field equilibration.

Behavioral analysis

Model-based analyses were used to investigate participants' learning and utilization of the reward equivalence structure to guide choices. Such analyses attempt to explain the timeseries of choices in terms of previous events, allowing precise, quantitative questions to be posed about the dynamics of behavioral adjustment. [See O'Doherty *et al.* (2007) and Daw (2011) for reviews of the methodology.]

First, we sought to test whether participants adjusted their choices dynamically in response to the rewarding outcomes. Because of the fluctuating probability of reward, we could not estimate a learning curve or a percent correct over the course of the task. Instead, as in prior studies, a logistic regression model was fit to explain each participant's sequence of choices in terms of two explanatory variables coding events from the previous trial– the choice made and whether it was rewarded (both coded as binary indicators; Lau & Glimcher, 2005; Gershman *et al.*, 2009; Daw *et al.*, 2011; Li & Daw, 2011). In the present study the dependent variable is multinomial (i.e. choices over four options), so that the appropriate model is a conditional logit (McFadden, 1974), i.e. the link function is the softmax from reinforcement learning (Daw, 2011).

Having determined that participants' choices were influenced by prior rewards, we next aimed to investigate more detailed aspects of learning using two variations of a reinforcement learning model fit to the choice sequences (Sutton & Barto, 1998), as detailed below.

The model learns to assign an action value to each option, $Q_1 \ldots Q_4$, according to previously experienced rewards. These are assumed to be learned by a delta rule: if option c was chosen and reward r (1 or 0) received, then Q_c is updated according to:

$$Q_{c,t} = Q_{c,t-1} + \alpha * \delta_{c,t} \quad (1)$$

$$\delta_{c,t} = r_t - Q_{c,t-1} \quad (2)$$

where the free parameter α controls the learning rate. To embody possible generalization of value across paired options with yoked

drifting reward probabilities, the model includes a capacity to update the partner option yoked to the current choice. In particular, if option c was chosen, with partner p, then in addition to updating the value of c, Q_c, as in Eqns (1) and (2), Q_p was also updated according to:

$$Q_{p,t} = Q_{p,t-1} + \alpha_2 * \delta_{p,t} \quad (3)$$

$$\delta_{p,t} = r_t - Q_{p,t-1} \quad (4)$$

with the free parameter α_2 controlling generalization learning rate. When α_2 is set to zero, the model is blind to correlational structure, and corresponds to models studied previously (Daw *et al.*, 2006; Schönberg *et al.*, 2007; Gershman *et al.*, 2009). In this sense, this no-generalization limit provides a null hypothesis or baseline model against which to test for generalization effects. With a non-zero generalization learning rate the model allows the reward feedback associated with a selected option (e.g. S_1 or S_2) to update the value of its partner (S_3 or S_4, respectively). Because the models are otherwise identical, this parameter isolates generalization, i.e. we reasoned that if the model with a free generalization learning rate fit significantly better than the baseline one, then such a difference would be attributable to generalization across partners. Moreover, the estimated value of the learning rate measures the strength of the generalization effect (Daw & Shohamy, 2008). Note that a version of the model in which instead of moving partners' values toward the obtained rewards, non-partners' values are moved away from them (reflecting anti-generalization according to negative correlations; Hampton *et al.*, 2006) makes predictions quantitatively similar to the version used here. This is because choice probabilities in the softmax (below) are driven only by the differences between Q values. Thus, for concreteness, and because the positive correlations were stronger in the reward schedules as programmed, we used the positively generalizing form of the rule.

Given value estimates on a particular trial, participants are assumed to choose between the options stochastically with probabilities $P_1 \ldots P_4$ according to a softmax distribution (Daw *et al.*, 2006):

$$P_{c,t} \propto \exp(\beta(Q_{c,t} + \varphi I(c,c_{t-1}))). \quad (5)$$

The free parameter β represents the inverse temperature, which controls the exclusivity with which choices are focused on the highest-valued option. The model also included a free parameter φ, which, when multiplied by the indicator function $I(c,c_{t-1})$, defined as 1 if c is the same choice as that made on the previous trial, and zero otherwise, captures a tendency to choose (for positive φ) or avoid (for negative φ) the same option chosen on the preceding trial (Lau & Glimcher, 2005; Schönberg *et al.*, 2007). Note that because the softmax is also the link function for the conditional logit model discussed above, this analysis also has the form of a regression from Q values onto choices (Lau & Glimcher, 2005; Daw, 2011) except here, rather than as linear effects, the past rewards enter via the recursive learning of Q, controlled, in nonlinear fashion, by the learning rate parameters.

To search for indications of generalization during the task (i.e. exploiting the relational structure underlying the gamble options), we compared the fit of two variants of the Q-learning model described by Eqns (1–5): (1) the 'base' model, where the generalization learning rate, α_2, was set to zero, and (2) the 'generalization' model, where α_2 was a free parameter.

Although equivalence effects would be expected to evolve over time as participants gradually learned the equivalence, for simplicity and lacking a well-supported formal model of the dynamics of such learning, we consider a simplified model in which α_2 is taken as fixed across the experiment. Because the partner learning rate is thus fit to explain choices even over early parts of the task during which it is unlikely that participants will yet have detected any generalization structure, this is a conservative analysis in the sense that it will tend to underestimate the asymptotic equivalence effects (Daw & Shohamy, 2008).

For each participant, maximum-likelihood values for the parameters α, β and φ, as well as α_2 for the generalization model, were estimated using a gradient search (repeated with 20 different starting points, decreasing the chance of local optima) over the likelihood of the participant's observed choice sequence, for each trial conditional on the previous rewards and choices (Lau & Glimcher, 2005; Daw *et al.*, 2006; Daw, 2011). In particular, log likelihood is computed as the sum over trials of $\log(P_c)$ for the actually chosen option using values learned by the model from the previously delivered rewards. A separate set of parameters was optimized for each participant.

To test whether the models provided a reliable account of participants' behavior, we performed several analyses. First, we tested whether the base and full generalization models fit significantly better than chance (i.e. a model with no parameters, with $P_{c,t} = 0.25$ for all t), using likelihood ratio tests. The relative degree of improvement over the chance model provides a standardized descriptive index of how well a model fits, called pseudo-R^2 (Camerer & Ho, 1999; Daw *et al.*, 2006), which we report for comparison with other studies. This is defined as $(R - L)/R$ where L and R are, respectively, the log likelihood of the choices under the model (base or generalization) and under purely random choices ($P_{c,t} = 0.25$ for all t).

For the critical comparison between models, the performance of the base model and the full generalization model were compared using likelihood ratio tests on the individual participant's and summed group log likelihood values. Such a test examines the null hypothesis that any improvement in model fit is due to chance, correcting for the inclusion of additional free parameters (e.g. Stephan *et al.*, 2009).

To reason about the prevalence of the two models across the population as a random effect that might vary across participants, we conducted an additional analysis using the Bayesian Model Selection (BMS) method of Stephan *et al.* (2009). In particular, we estimated Bayes factors (the posterior evidence for one model over the other; Kass & Raftery, 1995) using the AIC criterion (Akaike, 1974), and submitted these to the spm_BMS routine from SPM8 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK).

Imaging analysis

Preprocessing and data analysis was performed using Statistical Parametric Mapping software (SPM5; Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK). Functional images were realigned to correct for participant motion and then spatially normalized by estimating a warping to template space from each participant's anatomical image (SPM5, 'segment and normalize') and applying the resulting transformation to the EPIs. Images were resampled to 2-mm cubic voxels, smoothed with an 8-mm FWHM Gaussian kernel, and filtered with a 128-s high-pass filter.

For reinforcement learning model-based analysis of the fMRI data, we investigated correlations with trial-by-trial parametric signals derived from simulations of the model described above (Eqns 1–5). Data were analysed using SPM5, under the assumptions of the general linear model. The events on each trial were modeled by half-second boxcar regressors at the time of stimulus onset and of outcome feedback. These two events were modulated by parametric regressors: the trial-by-trial probability of the chosen option (Eqn 5) on the stimulus onset, and the trial-by-trial prediction error (Eqn 4) on the outcome. Each event was also modulated by a second parametric

regressor capturing the difference between probabilities or prediction errors in a model with and without generalization (formally, the partial derivative of the modeled quantity with respect to the partner learning rate; see below). Nuisance boxcar regressors were also included during the choice period (4 s) and outcome display periods (2 s) to account for general effects of visual stimulation.

To generate the parametric regressors for the imaging analysis, the free parameters for the learning model were chosen as follows. First, the learning model was re-estimated with the generalization learning rate, α_2, set to zero. This was chosen so as to best characterize values and prediction errors under the null hypothesis of no generalization, allowing us to test (and perhaps reject) it at the neural level. Second, as has been noted previously (Daw *et al.*, 2006), individual parametric fits in tasks and models of this sort tend to be noisy, and regularization of the behaviorally fit parameters across participants tends to improve a model's subsequent fit to fMRI data. Accordingly (following previous work: Daw *et al.*, 2006; Schönberg *et al.*, 2007; Gershman *et al.*, 2009), we generated regressors for each participant using a single setting of the reinforcement learning model's free parameters, here taken as the mean over all participants of the best-fitting individual estimates. The group means estimate the population-level parameters in a random-effects model of inter-subject variability (Holmes & Friston, 1998), and are thus a principled choice for the entire group. Note that although we thus do not characterize individual variability in most of the behavioral model parameters for the purpose of generating fMRI regressors, our approach does capture individual variability in the most important parameter for our questions of interest, the generalization learning rate, α_2, as the prediction error partial derivative 'difference' regressor capturing its effects in the fMRI model (see below) is taken as a random effect across subjects.

To investigate whether value-related neural signals reflect generalization of feedback across partner options, two additional regressors were included to accompany the base parametric reinforcement learning regressors of reward prediction error and choice probability. These two 'difference regressors' each characterize how one of these trial-by-trial parametric timeseries would change if the model included learning from partner option feedback (i.e. if the parameter α_2 were nonzero). Intuitively, these regressors represent the *difference* between the probabilities (or prediction errors) generated according to two competing assumptions about the generalization learning rate α_2 – that it takes on some nonzero value Δ, vs. the null assumption that $\alpha_2 = 0$ (Wittmann *et al.*, 2008; Daw, 2011; Daw *et al.*, 2011). Thus, if the BOLD signal in an area is better correlated with the regressor timeseries for nonzero generalization ($\alpha_2 = \Delta$), then, given the additive nature of the general linear model (GLM), the net BOLD signal will be best explained by a sum of contributions from the main regressor ($\alpha_2 = 0$) plus the difference regressor. If, instead, the BOLD signal correlates are best explained by $\alpha_2 = 0$, there should be no effect of the difference regressor. In other words, this analysis separates a test for generic prediction error (without generalization) and an additional, orthogonal, test of whether such activity would actually be better explained by the prediction error including generalization (the test of the difference regressor in the same voxels). This approach (Wittmann *et al.*, 2008; Daw *et al.*, 2011; Bornstein & Daw, 2012) more cleanly separates these two inferences than simply including regressors generated according to both models and contrasting them (e.g. Hampton *et al.*, 2008) particularly when the signals predicted by the models are correlated.

More formally, this additive approach approximates the (nonlinear) effect of an arbitrary α_2 on the modeled probability or prediction error timeseries in the context of the standard linear analysis of the BOLD response by using a Taylor expansion of this nonlinear function around $\alpha_2 = 0$ and retaining the first-order (linear) term (Friston *et al.*, 1998; Daw, 2011). This corresponds, in the above scheme, to taking the learning rate increment Δ infinitesimally small, or equivalently, to defining the difference regressors as the partial derivatives of the modeled timeseries with respect to α_2, evaluated at $\alpha_2 = 0$. Thus, if the BOLD response is better explained by a timeseries including nonzero generalization, the additive general linear model will explain the BOLD signal via the weighted sum of both regressors; in particular, a significantly positive effect will be estimated for the partial derivative. Voxels that show significant correlations with both prediction error and the prediction error difference regressor, or base chosen value and the chosen value difference regressor, exhibit activity that is better fit by a generalization learning model.

To test whether neural effects related to reinforcement learning variables were better explained by including effects of generalization, we identified activity correlated with basic reinforcement learning variables, then tested for effects of the difference regressors (orthogonalized against the original variables to test only for residual activity), in the vicinity. To test whether these effects were significant in the same voxels (and thus whether activity in a voxel is best described by the weighted sum of both effects) we examined the conjunction of two tests, using SPM's conjunction null (Nichols *et al.*, 2005). Note that although the difference regressors were orthogonalized to the underlying prediction error variables, the validity of conjunction inference using the minimum t statistic does not depend on the conjoined tests being independent (Nichols *et al.*, 2005).

Finally, we examined functional interactions between the striatum and the hippocampus during learning. We focused on a ventral striatum cluster identified in the above GLM as having a significant correlation with the prediction error difference regressor [6-mm spherical region of interest (ROI); coordinates – 14, 8, −8]. A psychophysiological interaction (PPI) analysis was estimated to test for increases in functional correlation between the ventral striatum (the physiological variable) and other brain regions during choice trials (the psychological variable). The time course of activation from the ROI was extracted and deconvolved. This timecourse was interacted with the choice trial boxcar indicator and then convolved with the hemodynamic response function (HRF). The model included the striatal timecourse by trial regressor, the trial regressor and the unmodulated striatal timecourse regressor (Friston *et al.*, 1997). We then correlated the resulting beta values with individual difference measures of the relative fit of the generalization model to behavior [calculated as the difference between choice likelihoods for the base model vs. the generalization model (e.g. Hampton *et al.*, 2008; Simon & Daw, 2011)].

fMRI model regressors were convolved with the canonical HRF and entered into a GLM of each subject's fMRI data. The six scan-to-scan motion parameters produced during realignment were included as additional regressors in the GLM to account for residual effects of subject movement. Linear contrasts of the resulting SPMs were taken to a group-level (random-effects) analysis. We report results small-volume corrected (SVC) for familywise error (FWE) due to multiple comparisons using cluster size (Friston *et al.*, 1993); this approach assesses the spatial extent of clusters defined by an initial and arbitrary uncorrected threshold, which we take as $P < 0.005$ for all analyses. Accordingly, for display purposes, we render all activations at this threshold. We conduct this correction either as whole brain, or within small volumes for which we had an a priori hypothesis. In particular, in the striatum we used a hand-drawn mask of the right nucleus accumbens, based on prior studies showing robust prediction error and model-based influences in this region (Wittmann *et al.*, 2008; Daw *et al.*, 2011; in both cases most robustly on the right). In the medial

temporal lobe we use an anatomically defined mask which included both the hippocampus and parahippocampus, derived from the AAL atlas (Tzourio-Mazoyer *et al.*, 2002). All voxel locations are reported in Montreal Neurological Institute coordinates and results are displayed overlaid on the average of all participants' normalized high-resolution structural images.

Results

Behavioral results

Over the course of the experiment, participants won \$7.56 ± 0.10 (mean ± SEM across participants). Participants were able on most trials to enter a choice within the time constraints (9.6 ± 1.4 missed trials out of 300). On completed trials, response times were 1.16 ± 0.02 s (grand means ± SEMs across participants).

As the task provides only binary feedback about the selected option on each trial, information about similarities between options can only accumulate over multiple trials and switches between options. Because of these properties of the design, knowledge of the task structure may not often reach the level of explicit awareness. Participants shifted their choice selection an average of 115.10 ± 9.47 times (range 32–211), which provides an opportunity for participants to compare values across options. To investigate whether participants displayed explicit awareness of the relational structure of the task, after the experiment, we presented them with a display of the four options and asked them to indicate, by drawing connecting lines, which pairs of options seemed related in any way. We also asked them in a written follow-up question to describe any reasons underlying their answer.

Across the group, pairing performance did not differ from chance (33%; mean correct 22 ± 10%; data available for 18 participants), indicating that participants, collectively, were not explicitly aware of the manipulation. Individually, our criterion for explicit knowledge was both correctly pairing the options and exhibiting some explicit knowledge of the reward equivalence structure on the written question, a combination achieved by only one participant. (On the written question, that participant stated that '… the pairs seemed to alternate when those two faces were *lucky*'.) These post-task measures suggest that the influence of the reward equivalence structure on choice behavior, as discussed below, is probably not due to participants' explicit detection of the relational structure of the task.

Reinforcement learning model of choices

Next, we used the fit of computational models to examine the trial-by-trial dynamics of behavioral adjustment. In particular, such models allow us to quantify how the choices depend on recent feedback, allowing questions to be asked about the specific nature of the updating – in particular, here, whether it reflected generalization between partners.

First, to examine whether participants adjusted their behavior dynamically to previous rewards, we fit a simple regression model to measure the extent to which each participant's choice sequence was predicted by the reward on the previous trial, also controlling for the previous choice as an additional explanatory variable, as done previously (Gershman *et al.*, 2009; Li & Daw, 2011). Consistent with prior reports, we found that across participants, the effect of the previous reward was significant ($\beta = 4.12 \pm 1.12$, $t = 3.67$, $P < 0.005$), indicating participants learned choice preferences from previous rewards, while the effect of the previous choice was not significant ($\beta = 0.06 \pm 0.26$, $t = 0.22$, $P > 0.5$).

Next, to examine whether this adjustment reflected the underlying hidden reward equivalence structure in the gambling task, we tested the fit of more detailed reinforcement learning models characterizing trial-by-trial adjustments in values for each option. In particular, we compared models which differed only in whether they generalized between partners, allowing us to test whether choice behavior revealed any generalization between equivalent options (S_1 & S_3 and S_2 & S_4; Daw & Shohamy, 2008). Standard reinforcement learning models would assume that participants' tendency to choose an option is based on a learned value for that option which is updated only from experience with outcomes from choices of that option. In contrast, a generalization model embodies the idea that outcomes received for one option can influence learning about the value of another option.

To address this question, we considered the fit of two different reinforcement learning models. The 'base' model consisted of a standard reinforcement learning model blind to the relational structure of the task, while the 'generalization' model extended the base model to allow feedback about the present choice to update the value of the unchosen partner option by way of an additional learning rate parameter. The two models coincide when this additional parameter takes on the value zero. A similar generalization model has been shown to better fit participant choice behavior in a prior study that reported the results of a task analogous to the current one (Daw & Shohamy, 2008).

First, we confirmed that both the base and generalization models each explained choices better than chance. This was the case both in the aggregate over participants (likelihood ratio tests; $\chi^2_{63} = 6694.20$; $\chi^2_{84} = 6833.10$; all P values < 1e−16) and also individually for all participants for both the base and the generalization model at $P < 0.0001$. Pseudo-r^2 statistics (a descriptive measure of model fit appropriate for comparing between studies) were 0.38 ± 0.17 for the base model and 0.39 ± 0.17 for the generalization model.

Next, we compared the two models' fits to one another to determine whether there was evidence for generalization. For choice likelihoods aggregated over all participants (equivalent to assuming all participants complied with one model or the other, and testing which one), the difference in log choice likelihoods (Table 1) was 69.4 in favor of the generalization model, i.e. the choices were exp(69.4) more likely given the generalization model than the base model (Kass & Raftery, 1995). We formally tested whether such an improvement was expected due to chance given the extra free parameters with a likelihood ratio test; the restriction to the base model was indeed

TABLE 1. Reinforcement learning model fits

Model	−LL	Aggregate LR test	α	β	φ	α_2
Base	5386.5	–	0.68 ± 0.06	4.50 ± 1.09	0.19 ± 0.04	–
Generalization	5317.1	$\chi^2_{21} = 138.9$ $P <$ 1e−16	0.69 ± 0.06	4.86 ± 1.11	0.19 ± 0.03	0.09 ± 0.04

To compare the base model and the generalization model, which incorporates value generalization across partnered options, we show negative log-likelihood (−LL), aggregate likelihood ratio test statistic (χ^2), and random-effects maximum-likelihood parameter estimates (mean ± SEM across participants) for both the base and generalization reinforcement learning models.

rejected in favor of the generalization model (likelihood ratio test, $\chi^2_{21} = 138.86$;$P < 1e{-}16$; Table 1).

The foregoing analyses aggregated evidence across participants. We next sought to address whether there were individual differences and whether the effects might be driven by outliers. Examining individuals, likelihood ratio tests also rejected the base model for 11 of 21 participants considered individually (at $P < 0.05$; 12/21 at $P < 0.06$). To more formally examine evidence for either model at the group level, allowing for the possibility that the existence of generalization might vary across participants (i.e. taking the identity of the best-fitting model as a random effect), we conducted an additional Bayesian analysis of the choice fits, fitting a hierarchical model in which participants are assumed to be drawn from either sort and estimating the proportions (Stephan *et al.*, 2009). The estimated fraction of generalizers in the population was 0.853 (compared with 0.147 for non-generalizers); the 'exceedance probability', or posterior probability that the generalization model was the more prevalent of the two, was 99.9%.

The hypothesis of generalization may also be assessed at the group level by treating the learning rate controlling generalization as a random effect analogous to population-level effects in fMRI (Holmes & Friston, 1998). Across participants, the best-fitting estimates were indeed significantly different from zero ($t_{20} = 2.40$, $P < 0.05$; range -0.01 to 0.69, Table 1; note that to render this test meaningful it is important that we did not constrain the estimated parameter to be positive).

Although significant, the generalization effect was modest in size – on average over participants, generalization learning rates were approximately 13% of the primary learning rate. We might expect generalization to be fractional, due to participants' potentially incomplete detection of the relationship. In particular, our model probably underestimates the asymptotic degree of generalization, as for simplicity it treats the parameter as constant throughout the experiment (see Materials and methods), in effect averaging over early parts of the experiment in which the relationship could not yet have been learned. Lacking a well-supported quantitative model of the timecourse of such learning, we separately estimated generalization learning rates for the first and second half of the experiment. The estimated generalization learning rates were significantly greater in the second half (first half, 0.059 ± 0.025; second half, 0.154 ± 0.046, $P < 0.05$ one-tailed, reflecting the directional hypothesis), which suggests that our model is detecting the expected increase in generalization knowledge over the course of the experiment.

Intriguingly, the single participant that displayed clear evidence of being explicitly aware of the generalization task structure showed the greatest model likelihood benefit for the generalization model and the second-highest fit generalization learning rate. Importantly, however, excluding this participant from the group likelihood ratio test, Bayesian model selection analysis and parametric tests did not affect the significance of the results. This suggests that while most participants were not explicitly aware of the generalization structure, our generalization model clearly detected the single participant that did exhibit awareness as an outlier, supporting the validity and sensitivity of our approach.

Together, these results provide evidence that participants utilized the underlying relational reward equivalence structure to generalize reward feedback across equivalent options and guide their choice behavior.

Imaging results

Our analyses of the behavioral data established that the generalization reinforcement learning model, which embodied the generalization of value across pairs of equivalent options, provided a better fit to participants' choice behavior than a reinforcement learning model blind to this relational structure. Thus, we turned to the BOLD fMRI data to investigate neural correlates of this generalization knowledge. In particular, we sought activity correlated with reward predictions and prediction errors as produced by simulations of the reinforcement learning model under the null assumption of no generalization, and then tested whether this activity showed additional evidence of generalization knowledge. We particularly sought to test whether BOLD correlates of reward prediction error in the ventral striatum were naïve to generalization, as would be predicted under the standard model of these responses, and if these signals originate in a procedural learning system entirely separate from a putative cortico-hippocampal system capable of detecting relations and generalizing from them (Daw, 2011; Daw *et al.*, 2011).

We focused first on activity correlated with the reward prediction error when the outcome is revealed. As reward prediction errors report the difference between received and expected rewards, they may reflect the effects of generalization (if any) on the expectations. In particular, if outcomes received for some option also affect the value predicted for its partner, they will affect the prediction error reported on subsequent choices of the partner option. In contrast, such generalization-driven updating of values for an unselected option is not possible in standard stimulus–reward association learning models. To distinguish these possibilities, the fMRI model included a parametric regressor for the base prediction error, assuming no generalization, and a second 'difference regressor' (technically, the partial derivative of the error signal with respect to the generalization learning rate, or equivalently the difference between the signals predicted by models with and without small amounts of generalization), characterizing how it would be expected to change if generalization were included (see Materials and methods). In particular, the sum of the base prediction error and difference regressors, in any weighted combination, corresponds approximately to the prediction error from a model including generalization (Fig. 1b). Thus, because the general linear model used for fMRI analysis is additive, if BOLD responses in a region significantly reflect effects of both regressors, then the net activity there is better explained by a prediction error including generalization, and the region may support the value generalization effect we observed in participants' behavior.

The difference regressor for prediction error across participants included a mean number of 105.3 ± 7.0 positive deflections and 173.4 ± 7.4 negative deflections. Difference regressor values were often most extreme when participants switched choices, as this is when the generalization model makes the most divergent predictions from the base model. To illustrate this effect, consider the case where an option (e.g. S1) has been rewarded on the last several trials, but the participant switches to choosing the partnered option on the next trial (e.g. S3). In the generalization model but not the base model, the value for S3 has increased, and this expectation will modulate the prediction error signal. Here, this will lead the difference regressor to include a negative deviation – if the choice is rewarded, this is less of a positive 'surprise' to the generalization model, while if it is not rewarded, this omission is more of a negative surprise.

Accordingly, we first localized regions where BOLD activity correlated with prediction errors derived from the base reinforcement learning model. Reward prediction error correlates have been found most prominently in the ventral striatum (Knutson *et al.*, 2001; Pagnoni *et al.*, 2002; McClure *et al.*, 2003; O'Doherty *et al.*, 2003; Delgado *et al.*, 2005; Daw *et al.*, 2006; Lohrenz *et al.*, 2007; Schönberg *et al.*, 2007; Hare *et al.*, 2008), a region densely innervated by midbrain dopamine neurons (Falck & Hillarp, 1959; Knutson &

Gibbs, 2007). Replicating these findings, in the current experiment, prediction error at reward outcome correlated with BOLD responses throughout the bilateral ventral striatum [Fig. 2, left; right peak $z = 5.57$ (14, 4, −14), left peak $z = 5.27$ (−22, −4, −16); both clusters were significant whole-brain FWE-corrected for cluster size].

We next examined whether residual activity in this region reflected effects that could be explained by generalization of value between options. Indeed, activation in a region of the right ventral striatum significantly correlated with the difference regressor designed to capture the effects of generalization on prediction error [Fig. 2, center; $z = 3.16$ (14, 8, −8), $P < 0.001$ uncorrected; $P < 0.01$ SVC for FWE in an a priori right nucleus accumbens anatomical ROI]. A conjunction analysis (Fig. 2, right; $P < 0.001$ uncorrected; $P < 0.01$ SVC) verified that this effect was spatially overlapping with the prediction error itself and therefore (see Materials and methods) that the net activity in this region was better explained by a prediction from the generalization model. (A similar sub-threshold cluster was observed in the left ventral striatum.)

Thus, in contrast to predictions based on simple reinforcement learning models, the net BOLD signal in the right ventral striatum, a region often characterized by a reward prediction error response, is best explained by a reinforcement learning model that incorporates generalization knowledge. This result is consistent with other recent indications that the striatal error signal is more sophisticated than previously suspected (Daw *et al.*, 2011; Simon & Daw, 2011).

Next, we asked whether generalization knowledge was also reflected in anticipatory value-related signals during the choice period. Again, we first localized regions where activity correlated with the value of the selected option (the 'chosen value') during the choice period. Here, following evidence from unit recordings that action values in the brain are normalized between options (Platt & Glimcher, 1999; Dorris & Glimcher, 2004; Sugrue *et al.*, 2004), and previous fMRI work (Daw *et al.*, 2006), we define an action's value by the probability that the model predicts it will be chosen, which (Eqn 5) is a normalized transform of the raw value. Prior reinforcement learning studies of learning and decision-making have often found correlates of chosen value in the ventromedial prefrontal cortex (Daw *et al.*, 2006; Kim *et al.*, 2006; Plassmann *et al.*, 2007; Hare *et al.*, 2008; Boorman *et al.*, 2009; Gershman *et al.*, 2009; Palminteri *et al.*, 2009; Smith *et al.*, 2010). Accordingly, at a reduced whole-brain threshold, we also observed a cluster of activation in the ventromedial prefrontal cortex correlated with value [$z = 2.97$ (−6, 58, −18), $P < 0.001$ uncorrected]. The most extensive region of correlation, however, was observed bilaterally in the hippocampus [Fig. 3; right peak $z = 3.88$ (34, −6, −26), left peak $z = 3.49$ (−16, −22, −20); $P < 0.05$ cluster-corrected in a medial temporal lobe mask]. Chosen value correlates in the hippocampus have not often been reported in previous studies of reward learning and decision-making, but this finding is consistent with several more recent reports of value encoding in the hippocampus in categorization and passive viewing tasks (Kumaran *et al.*, 2009; Lebreton *et al.*, 2009; Dickerson *et al.*, 2011).

We then asked whether these responses also reflected generalization knowledge. An analysis of the value difference regressor did not show any significant correlation in the ventromedial prefrontal cortex. In the left hippocampus, we observed a cluster correlated with difference regressor at an uncorrected whole-brain threshold ($P < 0.005$), but this activation did not survive cluster-correction based on a medial temporal lobe mask. The lack of significant evidence for generalization effects in these signals is unexpected in light of our hypothesis that the hippocampal system might support value generalization.

To examine this hypothesis further, we turned to individual differences in generalization as expressed behaviorally and tested

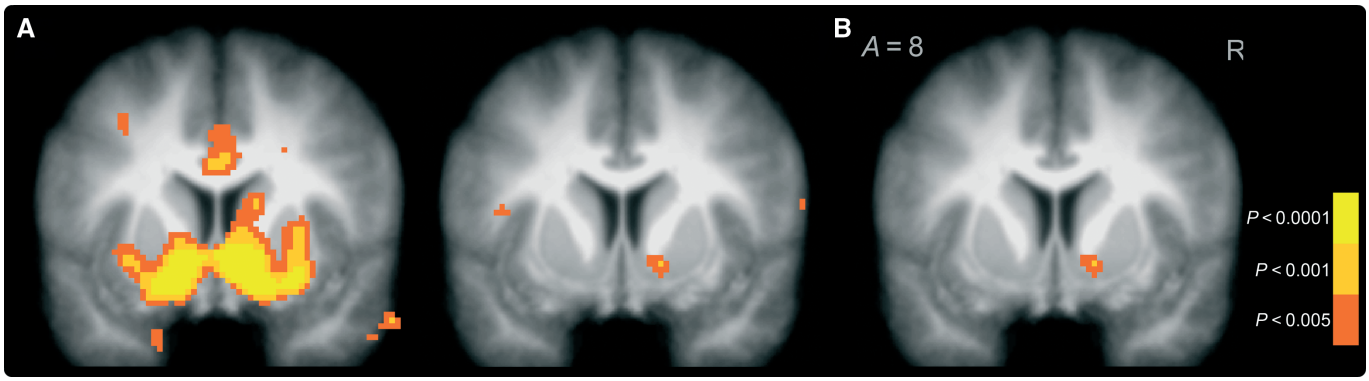

FIG. 2. Ventral striatum BOLD signals are best described by a model that incorporates generalization knowledge. (A) Prediction error, left. Prediction error difference due to value generalization, middle. (B) Conjunction of prediction error and prediction error difference due to generalization ($P < 0.05$, SVC; all MRI images $P < 0.005$ uncorrected, for visualization).

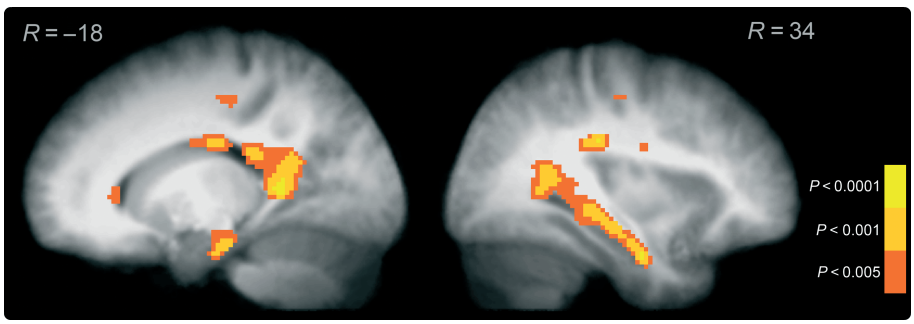

FIG. 3. Hippocampal activation correlated with chosen value during the choice period of the reward equivalence task ($P < 0.05$, SVC).

their relationship to functional connectivity between the striatum and hippocampus. To assess connectivity, we conducted a PPI analysis using as a seed the region of right ventral striatum showing a significant fit to the prediction error difference due to generalization. Overall, trial-related activity in the ventral striatum was significantly correlated with activity in widespread brain regions, including multiple clusters in the hippocampus. We correlated the degree of connectivity from ventral striatum, across participants, with the behavioral model fit benefit provided by the generalization model [the difference in choice likelihoods between the generalization and null models; $n = 20$, excluding a single outlier whose benefit was > 2 SD from the mean (the participant who exhibited high awareness of task structure)]. We found that the degree of striatal-hippocampal connectivity was significantly predicted by the generalization model improvement in fit to choices [$z = 4.0$ (26, −18, −16); Fig. 4]. Further, the cluster showing a connection between connectivity and generalization overlapped with the regions of the hippocampus where the BOLD signal exhibited a significant correlation with choice value. This result is consistent with the hypothesis that the hippocampus (and, more specifically, hippocampal-striatal functional connectivity) contributes to choices that benefit from generalization across correlated options.

Discussion

Our data show that choice behavior and feedback-related BOLD signals in the striatum are both influenced by the generalization of reward across equivalent options, as revealed by a novel reinforcement learning task in which payoff probabilities between pairs of options were correlated, providing an opportunity for participants to encode stimulus–stimulus relations. This structure, wherein individual cues have common outcomes, leading to generalization between them, is conceptually similar to the structure of acquired equivalence tasks used in research on memory to probe hippocampal representations (Myers *et al.*, 2003; Daw & Shohamy, 2008; Shohamy & Wagner, 2008). However, here the common outcomes are likelihood of reward, rather than (as in paired associate learning used in human studies) the identity of the outcome stimulus. Nevertheless, we found that the influence of this shared reward probability on both choice behavior and striatal BOLD signaling was captured by a reinforcement learning model that, whenever feedback was received about an option's value, also fractionally updated the value of its equivalent partner. Notably, such generalization on the basis of correlational structure is not predicted by standard reinforcement learning models commonly used to describe reward-driven learning and associated neural responses in the midbrain dopamine system.

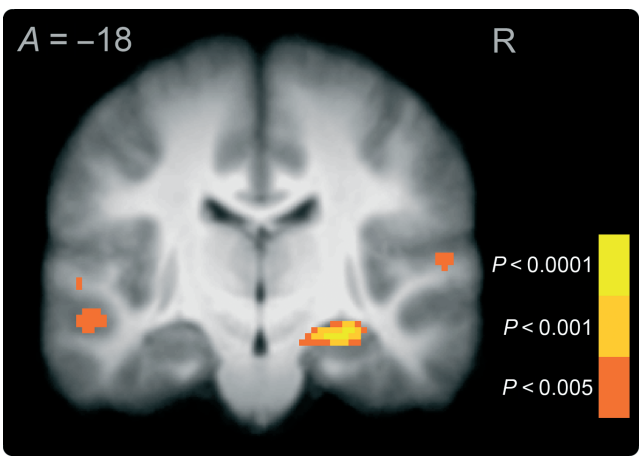

FIG. 4. Psychophysiological interaction (PPI) between task and ventral striatal activity is predicted by the degree that a participant's choice behavior is better fit by the generalization reinforcement learning model ($n = 20$; $P < 0.05$ SVC).

The present results suggest that human participants do indeed encode structure and generalize across correlated choice options during reinforcement learning. One ambiguity that remains is to what extent our effects are driven by positive correlations between 'equivalent' options or, instead or additionally, by weaker negative correlations that were also included between the two equivalent pairs in our reward schedules. For concreteness (and because the positive correlations were objectively much stronger), our analysis assumed positive generalization. However, in our model and overall framework (see Materials and methods) generalization driven by correlations could, in principle, also arise due to negative generalization between anti-correlated non-partners. Both conceptually and mathematically (due to symmetries in the softmax choice equations), positive and negative generalization might be expected to have quantitatively quite similar effects on choices and BOLD signals. Therefore, disentangling the relative contributions of similarity and distinctiveness to generalization awaits further experiments manipulating the positive and negative correlations independently. In reinforcement learning tasks, unambiguous negative generalization between options has been observed when their values are strongly anti-correlated due to a serial reversal contingency (Hampton *et al.*, 2006; Bromberg-Martin *et al.*, 2010). Importantly, the central cognitive and computational issues and our basic conclusions about generalization according to structure crosscut this distinction between positive and negative generalization.

Hippocampus and value

On the basis of the analogy between the correlational structure embedded in our task and that in acquired equivalence studies (Coutureau *et al.*, 2002; Myers *et al.*, 2003; Shohamy & Wagner, 2008), we hypothesized that learning of correlational structure would implicate the hippocampus. Our data provide somewhat mixed support for this hypothesis. The most direct evidence in favor was our finding that connectivity between the striatum and hippocampus predicted the degree to which participants' choice behavior was better described by the generalization model.

Also consistent with hippocampal involvement in this task, we found strong and widespread covariation of the BOLD signal in bilateral hippocampus with chosen option value, derived from the reinforcement learning model. This activation stands out in the context of the literature on reinforcement learning tasks similar to ours, particularly as similar activity is much more widely reported in ventromedial prefrontal cortex (Daw *et al.*, 2006; Kim *et al.*, 2006; Plassmann *et al.*, 2007; Hare *et al.*, 2008; Boorman *et al.*, 2009; Gershman *et al.*, 2009; Palminteri *et al.*, 2009; Smith *et al.*, 2010), where value-related activity was relatively modest in the present study.

An intriguing possibility is that the inclusion of structure in the present task recruited systems for valuation at least partly distinct from those exercised by other tasks. This hypothesis is consistent with other recent reports of hippocampal activation related in some way to stimulus value, which used task designs (active learning, passive observation, or model-based reinforcement learning) that might enhance the relevance of relational information relative to standard reinforcement learning tasks (Kumaran *et al.*, 2009; Lebreton *et al.*, 2009; Dickerson *et al.*, 2011; Simon & Daw, 2011). However, future studies will be necessary to test this hypothesis directly by specifically comparing learning with a relational component vs. without within a single study.

We were unable to demonstrate the effects of value generalization quantitatively in hippocampal correlates of value, even though effects of generalization were visible in the striatum. Based on our hypothesis that the hippocampus supports generalization between option values, this result is puzzling and it may indicate that the hypothesis was incorrect. At the same time, this null result should not be over-interpreted; this may be due, for instance, to our less refined quantitative characterization of the neural correlates of chosen value in the hippocampus, relative to prediction errors as studied in the striatum. In particular, it has been persistently unclear whether neural activity in many different parts of the brain correlates with chosen value linearly, or better via some nonlinear transform or normalization such as the softmax employed here (Platt & Glimcher, 1999; Corrado *et al.*, 2005; Daw & Doya, 2006; Daw *et al.*, 2006). However, the form of this relationship is not well specified, and because our analysis seeking neural correlates of generalization is based on a linear approximation (a first-order Taylor expansion of the modeled signal's dependence on the learning rate for generalization), it is probably particularly sensitive to any misspecification of this sort. [See also Daw *et al.*'s (2011) discussion of value-related BOLD activity in ventromedial prefrontal cortex vs. striatum for a similarly equivocal result from a similar analysis.]

Because our results give mixed support to the hypothesized role of the hippocampus in value generalization, it is worth considering whether our design changes some key aspects of acquired equivalence that engage the hippocampus. The chief difference from most acquired equivalence studies is that in our task equivalence is driven by value (e.g. stimulus–reward equivalences) rather than arbitrary stimulus–stimulus associations of the sort used in prior human and animal acquired equivalence studies (Coutureau *et al.*, 2002; Myers *et al.*, 2003; Shohamy & Wagner, 2008). However, acquired equivalence has also been demonstrated via value in rodents (albeit without neural manipulations to test hippocampal involvement; Honey & Hall, 1989). Moreover, entorhinal lesions have been shown in rodents to affect acquired equivalence using a task that counterbalances both stimulus–stimulus and stimulus–reward outcomes between partners and nonpartners (Coutureau *et al.*, 2002). This suggests that the medial temporal lobe memory system may be implicated more generally in encoding and inferring equivalence, rather than specifically for stimulus–stimulus encoding.

Finally, although conscious awareness is sometimes viewed as a characteristic of hippocampal episodic representations, the finding that most participants in our task did not report awareness of the relational structure does not preclude hippocampal involvement. Work in both memory and decision-making isolates different types of representations operationally by the nature of the information coded rather than by self-report; in this context, there is much evidence of hippocampal involvement in relational coding absent conscious awareness (Greene *et al.*, 2006; Shohamy & Wagner, 2008; Hannula & Ranganath, 2009). Nonetheless, it is interesting to note that in the current study the single participant who showed clear evidence of awareness of the task structure also showed the strongest evidence of generalization in trial-by-trial choices. Future studies are necessary to more directly probe the role of awareness of structure on generalization and on the role of the hippocampus and the striatum in generalization-guided choices.

Ventral striatum and value generalization

Although our task elicited relational coding not typically implicated in reinforcement learning tasks and may have recruited additional neural circuitry subserving this function, we nevertheless also observed the now-standard correlates of reward prediction error in the ventral striatum (Knutson *et al.*, 2001; McClure *et al.*, 2003; O'Doherty *et al.*, 2003; Delgado *et al.*, 2005; Lohrenz *et al.*, 2007; Hare *et al.*, 2008). However, here the net striatal activation was better explained by error signals from the augmented model that learned its predictions about an option's value not just from feedback about that option, but also by generalizing from its partner. By design, such a finding goes beyond what can be explained by standard reinforcement learning models without such augmentation, and demonstrates that the ventral striatum has access to information about correlational structure of a sort that goes beyond the simple, stimulus–reward learning normally associated with this area. The question of whether striatal value signals reflect such generalization was left open by a related study by Hampton *et al.* (2006), who investigated generalization in a serial reversal task (which causes two option values to be negatively correlated, rather than positively, as here). There, value correlates in ventromedial prefrontal cortex were shown to reflect generalization, but the same question was not asked about prediction errors in the striatum. (Also, unlike the present study, participants in the Hampton task were instructed as to the reversal contingency.)

The finding of generalization in the striatal error signal also cuts against two-system accounts of both reinforcement learning and of memory systems, which envision that a standard temporal-difference learning system is responsible for limited, 'habitual' behaviors, whereas more sophisticated decision-making phenomena drawing on cognitive maps or action–outcome associations (in memory terms, relational representations) are segregated in a parallel, competing network for 'model-based' reinforcement learning (Doya, 1999; Daw *et al.*, 2005; Balleine *et al.*, 2008; Rangel *et al.*, 2008; Redish *et al.*, 2008). Contrary to our results, such an architecture predicts that signals originating within the putative temporal-difference system (notably, the ventral striatal prediction error) will be naïve to relational information even when behavior, under the control of the more sophisticated system, reflects it.

Two other recent results attempting to interrogate the model-free vs. model-based distinction more explicitly also found evidence for model-based effects on striatal prediction error signals (Daw *et al.*, 2011; Simon & Daw, 2011). Altogether, these results suggest that the systems are more interacting than separate, an idea even more directly supported by the present study's results regarding functional connectivity between striatum and hippocampus. That said, another possibility regarding the present dataset is that generalization effects do not arise from a full model-based planning system, but rather, from standard temporal-difference learning operating over an input representation that reflects the relationship between the options (i.e. which maps options to values but with equivalent options coded in an overlapping fashion; Gluck & Myers, 1993; Moustafa *et al.*, 2009). Such an interpretation is also consistent with recent evidence from a two-phase acquired equivalence task that suggested that generalization effects arose already during the initial learning phase rather than via inference about equivalent relationships conducted during the probe phase (Shohamy & Wagner, 2008; as would be expected from a model-based reinforcement learning system).

Although some results suggest that prediction errors in striatal BOLD signal may in part reflect dopaminergic inputs there (Pessiglione *et al.*, 2006; Knutson & Gibbs, 2007; Schott *et al.*, 2008; Schönberg *et al.*, 2010), it is not possible to isolate the underlying neural cause for our effect, or in particular to conclude whether prediction errors carried by dopamine neurons also similarly reflect generalization. A related point is that the net BOLD signal in an area probably superimposes multiple underlying neural causes – including local processing and activity from different inputs. Thus, although our

analysis uses the conjunction of multiple additive effects to assess what sort of prediction error signal best explains the net BOLD response, it is not possible to exclude the possibility that these effects have different neural sources, and in particular that the generalization-related activity originates from a different source than the prediction error. All these questions could best be answered using unit recordings. However, in this respect it is interesting that our results are strongly reminiscent of a recent neurophysiological study in nonhuman primates, which showed that dopamine neurons also reflect values learned by generalization between two (negatively correlated) options in a serial reversal task (Bromberg-Martin *et al.*, 2010).

All these results (but not the idea of strictly segregated learning systems) are broadly consistent with strong anatomical connections between the hippocampus and the mesolimbic dopamine system. Intriguingly, in the present dataset, we find that functional connectivity between these regions, the ventral striatum and hippocampus, is predicted by the degree that participant's choices were fit by the generalization model. Anatomically, the ventral striatum may gain access to relational representations via direct projections there from the hippocampus and medial temporal lobe (Kelley & Domesick, 1982; Cohen *et al.*, 2009). Conversely, value information in the hippocampus may arrive via significant projections from midbrain dopaminergic neurons of the ventral tegmental area (Dahlström & Fuxe, 1964; Swanson, 1982; Frey *et al.*, 1990; Gasbarri *et al.*, 1994; Huang & Kandel, 1995; Otmakhova & Lisman, 1996). These latter connections have broader implications for how hippocampal memories are influenced by reward, motivation and predictions (e.g. Adcock *et al.*, 2006; Shohamy & Wagner, 2008; Kuhl *et al.*, 2010; for a review see Shohamy & Adcock, 2010).

Limitations and future directions

One limitation of the present study is that, although our findings demonstrate that participants used the equivalence between the options to guide choices and that this effect increases in the second half of the experiment, our reinforcement learning model does not explicitly characterize the learning of the equivalence. To focus on the question of whether participants' value learning reflected the equivalence structure, we took the degree of such learning and the underlying equivalence structure over which it operated as fixed throughout the task. For the questions of the present study, the main consequence of this approach is likely to underestimate the asymptotic size of the generalization effect, but it leaves open the question of how learning of the equivalence structure occurred. Accounts of such learning are reasonably well understood (at least in the abstract, it can be accomplished by Bayesian model comparison; Griffiths & Tenenbaum, 2005; Courville *et al.*, 2006; Kemp & Tenenbaum, 2008); however, the present experimental design is not well suited to testing them. In particular, because the actual equivalence structure was fixed throughout the task, learning of it occurred alongside many other potentially confounding changes (e.g. representational, strategic or habituation) that may occur simply with time on task; a more targeted design would incorporate dynamic equivalencies so as to test different dynamic accounts of how participants follow them.

In general, our results highlight the promise of integrated investigations of memory and decision-making. While often studied separately, it is clear that memory, if is it to be behaviorally beneficial, exists to guide decisions (Buckner, 2010; Shohamy & Adcock, 2010). A growing number of studies already focus on the cognitive and neural underpinnings of the use of different types of information in decision-making (Johnson *et al.*, 2007; Daw & Shohamy, 2008; van der Meer *et al.*, 2010; Shohamy & Adcock, 2010). Future studies may further probe how and when these different types of memory are reassembled into behavior by studying more complex decision processes and environments in conjunction with computational models (e.g. Daw *et al.*, 2005). In this respect, our data point to the ability of the striatum to utilize information characteristic of relational memory systems, thus suggesting at least one underexplored way in which past experience can drive future choices.

Acknowledgements

This work was supported in part by NIDA (R03 DA026957 to D.S.), NINDS (R01 NS 078784 to D.S. and N.D.), the Brain and Behavior Research Foundation (NARSAD Young Investigator Award to D.S. and N.D.), a Scholar Award from the McKnight Foundation (N.D.), and an Award in Understanding Human Cognition from the McDonnell Foundation (N.D.). We are grateful to Sam Gershman and Dylan Simon for assistance in data analysis.

Abbreviations

BMS, Bayesian Model Selection; BOLD, blood oxygen level-dependent; fMRI, functional magnetic resonance imaging; FWE, familywise error; GLM, general linear model; ITI, intertrial interval; PPI, psycho-physiological interaction; ROI, region of interest; SVC, small-volume corrected.

References

Adcock, R.A., Thangavel, A., Whitfield-Gabrieli, S., Knutson, B. & Gabrieli, J.D. (2006) Reward-motivated learning: mesolimbic activation precedes memory formation. *Neuron*, **50**, 507–517.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, **19**, 716–723.

Balleine, B.W., Daw, N.D. & O'Doherty, J.P. (2008) Multiple forms of value learning and the function of dopamine. In Glimcher, P.W., Camerer, C.F., Poldrack, R.A. & Fehr, E. (Eds), *Neuroeconomics: Decision Making and the Brain*. Academic Press, New York, NY.

Boorman, E.D., Behrens, T.E., Woolrich, M.W. & Rushworth, M.F. (2009) How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action *Neuron*, **62**, 733–743.

Bornstein, A.M. & Daw, N.D. (2012) Dissociating hippocampal and striatal contributions to sequential prediction learning. *Eur. J. Neurosci.*, **35**, in press.

Brainard, D.H. (1997) The Psychophysics Toolbox. *Spat. Vis.*, **10**, 433–436.

Bromberg-Martin, E.S., Matsumoto, M., Hong, S. & Hikosaka, O. (2010) A pallidus-habenula-dopamine pathway signals inferred stimulus values. *J. Neurophysiol.*, **104**, 1068–1076.

Buckner, R.L. (2010) The role of the hippocampus in prediction and imagination. *Annu. Rev. Psychol.* **61**, 27–48, C21–28.

Camerer, C. & Ho, T.H. (1999) Experience-weighted attraction in learning normal-form games. *Econometrica*, **67**, 827–874.

Cohen, N.J. & Eichenbaum, H. (1993) *Memory, Amnesia, and the Hippocampal System*. MIT Press, Cambridge, MA.

Cohen, M.X., Schoene-Bake, J.C., Elger, C.E. & Weber, B. (2009) Connectivity-based segregation of the human striatum predicts personality characteristics. *Nat. Neurosci.*, **12**, 32–34.

Corrado, G.S., Sugrue, L.P., Seung, H.S. & Newsome, W.T. (2005) Linear-Nonlinear-Poisson models of primate choice dynamics. *J. Exp. Anal. Behav.*, **84**, 581–617.

Courville, A.C., Daw, N.D. & Touretzky, D.S. (2006) Bayesian theories of conditioning in a changing world. *Trends Cogn. Sci.*, **10**, 294–300.

Coutureau, E., Killcross, A.S., Good, M., Marshall, V.J., Ward-Robinson, J. & Honey, R.C. (2002) Acquired equivalence and distinctiveness of cues: II. Neural manipulations and their implications. *J. Exp. Psychol. Anim. Behav. Process.*, **28**, 388–396.

Dahlström, A. & Fuxe, K. (1964) Localization of monoamines in the lower brain stem. *Experientia*, **20**, 398–399.

Davachi, L. (2006) Item, context and relational episodic encoding in humans. *Curr. Opin. Neurobiol.*, **16**, 693–700.

Daw, N.D. (2011) Trial-by-trial data analysis using computational models. In: Delgado, M.R., Phelps, E.A. & Robbins, T.W. (Eds), *Decision making,*

affect, and learning: attention and performance XXIII. Oxford University Press, Oxford, UK, pp. 3–38.
Daw, N.D. & Doya, K. (2006) The computational neurobiology of learning and reward. *Curr. Opin. Neurobiol.*, **16**, 199–204.
Daw, N.D. & Shohamy, D. (2008) The cognitive neuroscience of motivation and learning. *Soc. Cogn.*, **26**, 593–620.
Daw, N.D., Niv, Y. & Dayan, P. (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.*, **8**, 1704–1711.
Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B. & Dolan, R.J. (2006) Cortical substrates for exploratory decisions in humans. *Nature*, **441**, 876–879.
Daw, N.D., Gershman, S.J., Seymour, B., Dayan, P. & Dolan, R.J. (2011) Model-based influences on humans' choices and striatal prediction errors. *Neuron*, **69**, 1204–1215.
Deichmann, R., Gottfried, J.A., Hutton, C. & Turner, R. (2003) Optimized EPI for fMRI studies of the orbitofrontal cortex. *Neuroimage*, **19**, 430–441.
Delgado, M.R., Miller, M.M., Inati, S. & Phelps, E.A. (2005) An fMRI study of reward-related probability learning. *Neuroimage*, **24**, 862–873.
Dickerson, K.C., Li, J. & Delgado, M.R. (2011) Parallel contributions of distinct human memory systems during probabilistic learning. *Neuroimage*, **55**, 266–276.
Dorris, M.C. & Glimcher, P.W. (2004) Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron*, **44**, 365–378.
Doya, K. (1999) What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw.*, **12**, 961–974.
Dusek, J.A. & Eichenbaum, H. (1997) The hippocampus and memory for orderly stimulus relations. *Proc. Natl. Acad. Sci. USA*, **94**, 7109–7114.
Eichenbaum, H. & Cohen, N.J. (2001) *From Conditioning to Conscious Recollection: Memory Systems of the Brain*. Oxford University Press, New York, NY.
Everitt, B.J. & Robbins, T.W. (2005) Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. *Nat. Neurosci.*, **8**, 1481–1489.
Falck, B. & Hillarp, N.A. (1959) On the cellular localization of catechol amines in the brain. *Acta Anat.*, **38**, 277–279.
Foerde, K., Knowlton, B.J. & Poldrack, R.A. (2006) Modulation of competing memory systems by distraction. *Proc. Natl. Acad. Sci. USA*, **103**, 11778–11783.
Frank, M.J., Seeberger, L.C. & O'Reilly R, C. (2004) By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, **306**, 1940–1943.
Frey, U., Schroeder, H. & Matthies, H. (1990) Dopaminergic antagonists prevent long-term maintenance of posttetanic LTP in the CA1 region of rat hippocampal slices. *Brain Res.*, **522**, 69–75.
Friston, K.J., Worsley, K.J., Frackowiak, S.J., Mazziotta, J.C. & Evans, A.C. (1993) Assessing the significance of focal activations using their spatial extent. *Hum. Brain Mapp.*, **1**, 210–220.
Friston, K.J., Buechel, C., Fink, G.R., Morris, J., Rolls, E. & Dolan, R.J. (1997) Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage*, **6**, 218–229.
Friston, K.J., Josephs, O., Rees, G. & Turner, R. (1998) Nonlinear event-related responses in fMRI. *Magn. Reson. Med.*, **39**, 41–52.
Gabrieli, J.D. (1998) Cognitive neuroscience of human memory. *Annu. Rev. Psychol.*, **49**, 87–115.
Gasbarri, A., Verney, C., Innocenzi, R., Campana, E. & Pacitti, C. (1994) Mesolimbic dopaminergic neurons innervating the hippocampal formation in the rat: a combined retrograde tracing and immunohistochemical study. *Brain Res.*, **668**, 71–79.
Gershman, S.J., Pesaran, B. & Daw, N.D. (2009) Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *J. Neurosci.*, **29**, 13524–13531.
Gluck, M.A. & Myers, C.E. (1993) Hippocampal mediation of stimulus representation: a computational theory. *Hippocampus*, **3**, 491–516.
Greene, A.J., Gross, W.L., Elsinger, C.L. & Rao, S.M. (2006) An FMRI analysis of the human hippocampus: inference, context, and task awareness. *J. Cogn. Neurosci.*, **18**, 1156–1173.
Griffiths, T.L. & Tenenbaum, J.B. (2005) Structure and strength in causal induction. *Cogn. Psychol.*, **51**, 334–384.
Hampton, A.N., Bossaerts, P. & O'Doherty, J.P. (2006) The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *J. Neurosci.*, **26**, 8360–8367.
Hampton, A.N., Bossaerts, P. & O'Doherty, J.P. (2008) Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proc. Natl. Acad. Sci. USA*, **105**, 6741–6746.
Hannula, D.E. & Ranganath, C. (2009) The eyes have it: hippocampal activity predicts expression of memory in eye movements. *Neuron*, **63**, 592–599.
Hare, T.A., O'Doherty, J., Camerer, C.F., Schultz, W. & Rangel, A. (2008) Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *J. Neurosci.*, **28**, 5623–5630.
Hartley, T. & Burgess, N. (2005) Complementary memory systems: competition, cooperation and compensation. *Trends Neurosci.*, **28**, 169–170.
Holmes, A.P. & Friston, K.J. (1998) Generalisability, random effects and population inference. *Neuroimage*, **7**, S754.
Honey, R.C. & Hall, G. (1989) Acquired equivalence and distinctiveness of cues. *J. Exp. Psychol. Anim. Behav. Process.*, **15**, 338–346.
Houk, J.C., Adams, J.L. & Barto, A.G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In Houk, J.C., Davis, J.L. & Beiser, D.G. (Eds), *Models of Information Processing in the Basal Ganglia*. MIT Press, Cambridge, MA, pp. 249–270.
Huang, Y.Y. & Kandel, E.R. (1995) D1/D5 receptor agonists induce a protein synthesis-dependent late potentiation in the CA1 region of the hippocampus. *Proc. Natl. Acad. Sci. USA*, **92**, 2446–2450.
Johnson, A., van der Meer, M.A. & Redish, A.D. (2007) Integrating hippocampus and striatum in decision-making. *Curr. Opin. Neurobiol.*, **17**, 692–697.
Kass, R.E. & Raftery, A.E. (1995) Bayes factors. *J. Am. Stat. Assoc.*, **90**, 7 73–795.
Kelley, A.E. & Domesick, V.B. (1982) The distribution of the projection from the hippocampal formation to the nucleus accumbens in the rat: an anterograde- and retrograde-horseradish peroxidase study. *Neuroscience*, **7**, 2321–2335.
Kemp, C. & Tenenbaum, J.B. (2008) The discovery of structural form. *Proc. Natl. Acad. Sci. USA*, **105**, 10687–10692.
Kim, H., Shimojo, S. & O'Doherty J, P. (2006) Is avoiding an aversive outcome rewarding? neural substrates of avoidance learning in the human brain. *PLoS Biol.*, **4**, e233.
Knowlton, B.J., Mangels, J.A. & Squire, L.R. (1996) A neostriatal habit learning system in humans. *Science*, **273**, 1399–1402.
Knutson, B. & Gibbs, S.E. (2007) Linking nucleus accumbens dopamine and blood oxygenation. *Psychopharmacology (Berl)*, **191**, 813–822.
Knutson, B., Adams, C.M., Fong, G.W. & Hommer, D. (2001) Anticipation of increasing monetary reward selectively recruits nucleus accumbens. *J. Neurosci.*, **21**, RC159.
Kuhl, B.A., Shah, A.T., DuBrow, S. & Wagner, A.D. (2010) Resistance to forgetting associated with hippocampus-mediated reactivation during new learning. *Nat. Neurosci.*, **13**, 501–506.
Kumaran, D., Summerfield, J.J., Hassabis, D. & Maguire, E.A. (2009) Tracking the emergence of conceptual knowledge during human decision making. *Neuron*, **63**, 889–901.
Lau, B. & Glimcher, P.W. (2005) Dynamic response-by-response models of matching behavior in rhesus monkeys. *J. Exp. Anal. Behav.*, **84**, 555–579.
Lebreton, M., Jorge, S., Michel, V., Thirion, B. & Pessiglione, M. (2009) An automatic valuation system in the human brain: evidence from functional neuroimaging. *Neuron*, **64**, 431–439.
Li, J. & Daw, N.D. (2011) Signals in human striatum are appropriate for policy update rather than value prediction. *J. Neurosci.*, **31**, 5504–5511.
Lohrenz, T., McCabe, K., Camerer, C.F. & Montague, P.R. (2007) Neural signature of fictive learning signals in a sequential investment task. *Proc. Natl. Acad. Sci. USA*, **104**, 9493–9498.
Mattfeld, A.T. & Stark, C.E. (2011) Striatal and medial temporal lobe functional interactions during visuomotor associative learning. *Cereb. Cortex*, **21**, 647–658.
McClure, S.M., Berns, G.S. & Montague, P.R. (2003) Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, **38**, 339–346.
McFadden, D. (1974) Conditional logit analysis of qualitative choice behavior. In Zarembka, P. (Ed.), *Frontiers in Econometrics*. Academic Press, New York, NY, pp. 105–142.
van der Meer, M.A., Johnson, A., Schmitzer-Torbert, N.C. & Redish, A.D. (2010) Triple dissociation of information processing in dorsal striatum, ventral striatum, and hippocampus on a learned spatial decision task. *Neuron*, **67**, 25–32.
Moustafa, A.A., Myers, C.E. & Gluck, M.A. (2009) A neurocomputational model of classical conditioning phenomena: a putative role for the hippocampal region in associative learning. *Brain Res.*, **1276**, 180–195.
Myers, C.E., Shohamy, D., Gluck, M.A., Grossman, S., Kluger, A., Ferris, S., Golomb, J., Schnirman, G. & Schwartz, R. (2003) Dissociating hippocampal versus basal ganglia contributions to learning and transfer. *J. Cogn. Neurosci.*, **15**, 185–193.

Nichols, T., Brett, M., Andersson, J., Wager, T. & Poline, J.B. (2005) Valid conjunction inference with the minimum statistic. *Neuroimage*, **25**, 653–660.
O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H. & Dolan, R.J. (2003) Temporal difference models and reward-related learning in the human brain. *Neuron*, **38**, 329–337.
O'Doherty, J.P., Hampton, A. & Kim, H. (2007) Model-based fMRI and its application to reward learning and decision making. *Ann. NY Acad. Sci.*, **1104**, 35–53.
Otmakhova, N.A. & Lisman, J.E. (1996) D1/D5 dopamine receptor activation increases the magnitude of early long-term potentiation at CA1 hippocampal synapses. *J. Neurosci.*, **16**, 7478–7486.
Pagnoni, G., Zink, C.F., Montague, P.R. & Berns, G.S. (2002) Activity in human ventral striatum locked to errors of reward prediction. *Nat. Neurosci.*, **5**, 97–98.
Palminteri, S., Boraud, T., Lafargue, G., Dubois, B. & Pessiglione, M. (2009) Brain hemispheres selectively track the expected value of contralateral options. *J. Neurosci.*, **29**, 13465–13472.
Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J. & Frith, C.D. (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, **442**, 1042–1045.
Plassmann, H., O'Doherty, J. & Rangel, A. (2007) Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. *J. Neurosci.*, **27**, 9984–9988.
Platt, M.L. & Glimcher, P.W. (1999) Neural correlates of decision variables in parietal cortex. *Nature*, **400**, 233–238.
Poldrack, R.A., Clark, J., Pare-Blagoev, E.J., Shohamy, D., Creso Moyano, J., Myers, C. & Gluck, M.A. (2001) Interactive memory systems in the human brain. *Nature*, **414**, 546–550.
Preston, A.R., Shrager, Y., Dudukovic, N.M. & Gabrieli, J.D. (2004) Hippocampal contribution to the novel use of relational information in declarative memory. *Hippocampus*, **14**, 148–152.
Rangel, A., Camerer, C. & Montague, P.R. (2008) A framework for studying the neurobiology of value-based decision making. *Nat. Rev. Neurosci.*, **9**, 545–556.
Redish, A.D., Jensen, S. & Johnson, A. (2008) A unified framework for addiction: vulnerabilities in the decision process. *Behav. Brain Sci.*, **31**, 415–437; discussion 437–487.
Schacter, D.L. (1990) Perceptual representation systems and implicit memory. Toward a resolution of the multiple memory systems debate. *Ann. NY Acad. Sci.*, **608**, 543–567; discussion 567–571.
Schönberg, T., Daw, N.D., Joel, D. & O'Doherty, J.P. (2007) Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *J. Neurosci.*, **27**, 12860–12867.
Schönberg, T., O'Doherty, J.P., Joel, D., Inzelberg, R., Segev, Y. & Daw, N.D. (2010) Selective impairment of prediction error signaling in human dorsolateral but not ventral striatum in Parkinson's disease patients: evidence from a model-based fMRI study. *Neuroimage*, **49**, 772–781.
Schott, B.H., Minuzzi, L., Krebs, R.M., Elmenhorst, D., Lang, M., Winz, O.H., Seidenbecher, C.I., Coenen, H.H., Heinze, H.J., Zilles, K., Duzel, E. & Bauer, A. (2008) Mesolimbic functional magnetic resonance imaging activations during reward anticipation correlate with reward-related ventral striatal dopamine release. *J. Neurosci.*, **28**, 14311–14319.
Schultz, W. (2006) Behavioral theories and the neurophysiology of reward. *Annu. Rev. Psychol.*, **57**, 87–115.
Schultz, W., Dayan, P. & Montague, P.R. (1997) A neural substrate of prediction and reward. *Science*, **275**, 1593–1599.
Shohamy, D. & Adcock, R.A. (2010) Dopamine and adaptive memory. *Trends Cogn. Sci.*, **14**, 464–472.
Shohamy, D. & Wagner, A.D. (2008) Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. *Neuron*, **60**, 378–389.
Shohamy, D., Myers, C.E., Geghman, K.D., Sage, J. & Gluck, M.A. (2006) L-dopa impairs learning, but spares generalization, in Parkinson's disease. *Neuropsychologia*, **44**, 774–784.
Shohamy, D., Myers, C.E., Kalanithi, J. & Gluck, M.A. (2008) Basal ganglia and dopamine contributions to probabilistic category learning. *Neurosci. Biobehav. Rev.*, **32**, 219–236.
Simon, D.A. & Daw, N.D. (2011) Neural correlates of forward planning in a spatial decision task in humans. *J. Neurosci.*, **31**, 5526–5539.
Smith, D.V., Hayden, B.Y., Truong, T.K., Song, A.W., Platt, M.L. & Huettel, S.A. (2010) Distinct value signals in anterior and posterior ventromedial prefrontal cortex. *J. Neurosci.*, **30**, 2490–2495.
Squire, L.R. (1992) Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychol. Rev.*, **99**, 195–231.
Staresina, B.P. & Davachi, L. (2009) Mind the gap: binding experiences across space and time in the human hippocampus. *Neuron*, **63**, 267–276.
Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J. & Friston, K.J. (2009) Bayesian model selection for group studies. *Neuroimage*, **46**, 1004–1017.
Sugrue, L.P., Corrado, G.S. & Newsome, W.T. (2004) Matching behavior and the representation of value in the parietal cortex. *Science*, **304**, 1782–1787.
Sutton, R.S. & Barto, A.G. (1998) *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.
Swanson, L.W. (1982) The projections of the ventral tegmental area and adjacent regions: a combined fluorescent retrograde tracer and immunofluorescence study in the rat. *Brain Res. Bull.*, **9**, 321–353.
Thorndike, E.L. (1911) *Animal Intelligence*. Hafner, Darien, CT.
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B. & Joliot, M. (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, **15**, 273–289.
Wittmann, B.C., Daw, N.D., Seymour, B. & Dolan, R.J. (2008) Striatal activity underlies novelty-based choice in humans. *Neuron*, **58**, 967–973.
Zeithamova, D. & Preston, A.R. (2010) Flexible memories: differential roles for medial temporal lobe and prefrontal cortex in cross-episode binding. *J. Neurosci.*, **30**, 14676–14684.