Behavioral/Systems/Cognitive

Signals in Human Striatum Are Appropriate for Policy Update Rather than Value Prediction

Jian Li[1,2] **and Nathaniel D. Daw**[1,2]

[1]Department of Psychology and [2]Center for Neural Science, New York University, New York, New York 10003

Influential reinforcement learning theories propose that prediction error signals in the brain's nigrostriatal system guide learning for trial-and-error decision-making. However, since different decision variables can be learned from quantitatively similar error signals, a critical question is: what is the content of decision representations trained by the error signals? We used fMRI to monitor neural activity in a two-armed bandit counterfactual decision task that provided human subjects with information about forgone and obtained monetary outcomes so as to dissociate teaching signals that update expected values for each action, versus signals that train relative preferences between actions (a policy). The reward probabilities of both choices varied independently from each other. This specific design allowed us to test whether subjects' choice behavior was guided by policy-based methods, which directly map states to advantageous actions, or value-based methods such as Q-learning, where choice policies are instead generated by learning an intermediate representation (reward expectancy). Behaviorally, we found human participants' choices were significantly influenced by obtained as well as forgone rewards from the previous trial. We also found subjects' blood oxygen level-dependent responses in striatum were modulated in opposite directions by the experienced and forgone rewards but not by reward expectancy. This neural pattern, as well as subjects' choice behavior, is consistent with a teaching signal for developing habits or relative action preferences, rather than prediction errors for updating separate action values.

Introduction

According to influential theories, the dopamine system broadcasts a prediction error signal for reinforcement learning (RL) (Barto, 1995; Schultz et al., 1997; Dayan and Abbott, 2001; Rangel et al., 2008). However, relatively little is known about the precise action of this signal in guiding subsequent decisions, and indeed, error-driven learning can support qualitatively different decision-making strategies (Sutton and Barto, 1998; Dayan and Abbott, 2001). Two approaches differ in the content of the information learned. Policy-based (direct actor) methods such as the actor/critic learn a policy or direct mapping from situations to advantageous actions, adjusting this in light of received rewards (Barto, 1995; Sutton and Barto, 1998; Dayan and Abbott, 2001). In contrast, value-based (indirect actor) methods, such as Q-learning, produce choice policies indirectly by learning an intermediate representation: the expected reward (value) for each candidate action (Watkins and Dayan, 1992). These intermediate representations can then be compared to derive a policy.

These algorithms formalize important neuropsychological concepts. Policy learning parallels the notion that reinforcement "stamps in" stimulus-response habits, which is central to contemporary accounts of drug abuse (Thorndike, 1898; Dickinson and Balleine, 2002; Everitt and Robbins, 2005). However, either value or policy learning can be accomplished using teaching signals that, in typical tasks, appear nearly identical. In a typical task, where subjects repeatedly select from different options for rewards, action values can be learned from a prediction error (PE) measuring the difference between the received reward and the option's previously predicted value. A policy can also be updated by comparing the received reward to an expected (e.g., reference or state) value prediction (Dayan and Abbott, 2001). Indeed, the actor/critic algorithm learns both values and policies using the same error signal (Barto, 1995). Attempts using such tasks to distinguish versions of these signals have produced inconsistent results (Morris et al., 2006; Roesch et al., 2007).

We studied human choices and fMRI signals in an RL task modified to distinguish signals appropriate for updating policies versus value predictions. Subjects repeatedly chose between two slot machines, associated with independent probabilities of delivering monetary reward. For each choice, the screen displayed the amount of reward subjects won, but also what they would have won, had they chosen the other option. This information should affect teaching signals for values or policies differently, allowing us to distinguish these computational strategies. If values for the two options are learned separately, then two prediction errors are needed: comparing each value prediction to its associated outcome. In contrast, the policy requires only a single update, depending only on the difference between the obtained and forgone outcomes (such that rewards need not be compared with predictions).

Received Dec. 4, 2010; revised Feb. 1, 2011; accepted Feb. 17, 2011.

This work was supported by Grant R01MH087882 from the National Institute of Mental Health as part of the National Science Foundation/National Institutes of Health Collaborative Research in Computational Neuroscience Program, and by a Scholar Award from the McKnight Foundation. We thank Peter Dayan for helpful discussions.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health.

Correspondence should be addressed to either Nathaniel D. Daw or Jian Li at the above addresses. E-mail: nathaniel.daw@nyu.edu or lijian@nyu.edu.

DOI:10.1523/JNEUROSCI.6316-10.2011

 0270-6474/11/315504-08$15.00/0

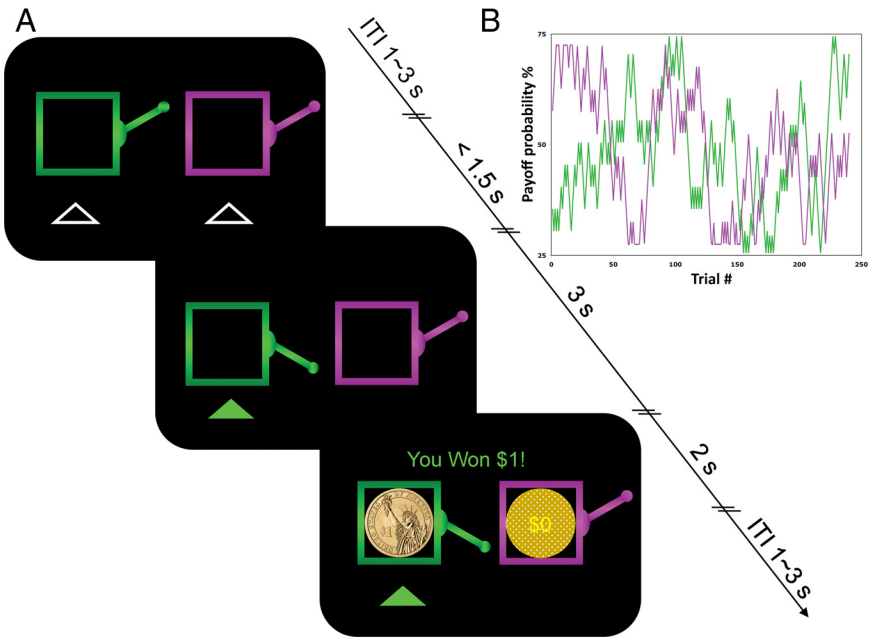

Figure 1. Experimental design. ***A***, Timeline of a single trial. ***B***, Example from one subject of changing reward probabilities for both slots. ITI, Intertrial interval.

Since both rewards were conditionally independent, it was unambiguous to determine their separate effects on neural activity. We focused on activity in the ventral striatum, which has been shown to correlate with PE signals similar to those seen in dopaminergic neuron recordings (Schultz et al., 1997; Berns et al., 2001; O'Doherty et al., 2003; Lohrenz et al., 2007).

Materials and Methods

Subjects. Twenty subjects participated in the experiment (13 female; mean age, 21.2 ± 5.0 years; age range, 18–39 years). The study was approved by the University Committee on Activities Involving Human Subjects and all subjects provided informed consent before the experiment.

Counterfactual learning task. The task consisted of four sessions of 60 trials each, separated by short breaks. At the start of each trial, subjects were presented with pictures of two differently colored (green and purple) slot machines. Subjects could select their choice of the slot machines using two button boxes (Fig. 1*A*), one for each hand; the left/right ordering of the slot machines was counterbalanced across subjects but kept constant across the task for each individual. Subjects had a maximum of 1.5 s to enter a choice; if no choice was entered during the 1.5 s window, the message "Please respond faster!" was displayed for 1 s, which was then followed by a 1–5 s delay before another trial started. In general, subjects responded well before the timeout, with a mean reaction time of 523 ± 20 ms (mean ± SEM). On valid trials, the chosen slot machine was highlighted and, 3 s later, the outcomes for both chosen and unchosen slot machines were displayed as a picture of a $1 coin or a circle marked $0 overlaid on each machine (Fig. 1*A*). The outcomes stayed on the screen for 2 s, after which the screen was cleared. The trial sequence ended 6.5 s after trial onset, and a randomly jittered intertrial interval with a mean of 2 s was introduced before the beginning of the next trial.

The payoff for each slot machine ($i = \{L, R\}$, where L is left and R is right) on each trial (t) was either $1 or $0 (Fig. 1*A*), with each payoff drawn independently from a binomial distribution according to a machine-specific probability, $p_{i,t}$, that gradually changed over trials (Fig. 1*B*). At the beginning of the task ($t = 1$), both probabilities were independently drawn from a uniform distribution with boundaries of [0.25,0.75]. Following each trial, the probabilities were each diffused either up or down by adding or subtracting 0.05 (equiprobably and independently). The updated probabilities, $p_{i,t+1}$, were then reflected off the boundaries [0.25,0.75] to maintain them within that range.

Subjects were instructed that they would be paid only according to the accumulated outcomes of the slot machines that they actually chose, not the forgone choices. Subjects were also told that the final points they earned would be converted proportionally to dollars but not told the actual scaling factor (which was 0.25).

Behavioral analysis. We first used a logistic regression analysis to estimate how a subject's choice on trial t (dependent variable, for all $t \geq 2$) was influenced by the chosen and unchosen rewards (R_c and R_u, respectively) on the previous trial. We specified three independent variables based on the events on the preceding trial: reward on the chosen slot machine (coded as 0 for no reward, or +1 or −1 for reward following a left or right choice, respectively), reward on the unchosen slot machine (coded similarly), and the choice on that trial [a dummy variable coded as +1 or −1 for left or right choices, respectively, so as to capture any first-order autocorrelation in the choices (Lau and Glimcher, 2005)]. We estimated regression weights for each subject individually using maximum likelihood, and report summary statistics for these quantities across subjects (Table 1).

We also fit the parameters of two learning models (detailed below) to each subject's choices by maximizing the (log) likelihood of the choice sequence:

$$\sum_t \log P(c_{s,t} \mid \Theta), \tag{1}$$

separately for each subject, s. Here, $c_{s,t}$ denotes the choice made by subject s on trial t and Θ is the parameter set. We sought optimal parameters using a nonlinear optimization algorithm (fmincon, Matlab optimization toolbox), and 30 different starting search locations for each subject so as to avoid local maxima. We report negative log likelihoods (smaller values indicate better fit), both pure and penalized (Kass and Raftery, 1995) for model complexity using the Bayesian information criterion. We also report a pseudo-r^2 statistic (Camerer and Ho, 1999; Daw et al., 2006), defined as $(r - l)/r$, where l and r are, respectively, the negative log likelihoods of the data under either fit model and under purely random choices ($P_{c,t} = 0.5$ for all trials and subjects).

Reinforcement learning models: Q-learning model. A Q-learning model (Watkins and Dayan, 1992; Sutton and Barto, 1998; Dayan and Abbott, 2001; Daw et al., 2006; Li et al., 2006) learns an expected value (Q value) for each option based on experienced outcomes, and then chooses accordingly. We adapted a standard model to allow learning from unchosen as well as chosen rewards; this simply updates each machine's value on each trial according to its own prediction error. We allowed the updating of the unchosen option to be controlled by a distinct learning rate to capture any differences in attention to the two outcomes. Specifically, at each trial, both values were updated according to the feedback received:

$$\begin{aligned} Q_{c,t+1} &= Q_{c,t} + \alpha \delta_{Q_{c,t}} \\ Q_{u,t+1} &= Q_{u,t} + \kappa\alpha \delta_{Q_{u,t}}, \end{aligned} \tag{2}$$

where α is a free learning rate parameter, κ modulates the learning rate for the unchosen option, and

$$\begin{aligned} \delta_{Q_{c,t}} &= R_{c,t} - Q_{c,t} \\ \delta_{Q_{u,t}} &= R_{u,t} - Q_{u,t}, \end{aligned} \tag{3}$$

are prediction errors for the rewards, R, received on each machine.

Table 1. Mean parameter fits ± SEM from 20 subjects for three models

	Models				
	Q-learning		Policy-gradient		Logistic regression
$Q_{L,0}$	0.54 ± 0.09		—	R_c	3.44 ± 1.51
$Q_{R,0}$	0.61 ± 0.08	η	0.48 ± 0.06	R_u	2.09 ± 0.053
α	0.35 ± 0.07	α	62.23 ± 20.22	Repetition	1.89 ± 0.58
$\kappa * \alpha$	0.24 ± 0.07	κ	0.48 ± 0.07	Intercept	−0.03 ± 0.06
β	9.08 ± 2.62	W_0	−14.28 ± 35.87		—

We further assumed the probability of choosing either machine ($i \in$ {L, R}) was softmax in its Q value:

$$P_{i,t} = \frac{\exp(\beta Q_{i,t})}{\sum_j \exp(\beta Q_{j,t})}, \quad (4)$$

with free exploration parameter β and initial Q values ($Q_{L,0}$ and $Q_{R,0}$) (Table 1).

Policy-gradient model. A second approach to reinforcement learning maintains policy parameters specifying a preference over options, and updates this preference with feedback to achieve stochastic gradient ascent on the expected reward (Dayan and Abbott, 2001; Dayan and Daw, 2008). We again represent the selection policy as softmax; here the chance of choosing machine L is

$$P_{L,t} = \frac{1}{1 + \exp(-w_t)}, \quad (5)$$

with the policy parameter w. Here,

$$P_{R,t} = 1 - P_{L,t} = \frac{1}{1 + \exp(w_t)}. \quad (6)$$

If $\langle r_R \rangle$ and $\langle r_L \rangle$ are the average expected reward on either slot machine, then the expected reward given any particular policy is as follows:

$$\langle r \rangle_w = P_L \langle r_L \rangle + P_R \langle r_R \rangle, \quad (7)$$

and its gradient with respect to the policy parameter,

$$\frac{\partial \langle r \rangle_w}{\partial w} = P_L P_R [\langle r_L \rangle - \langle r_R \rangle], \quad (8)$$

is proportional to the difference between the two reward rates. On each trial, then, the gradient can thus be sampled stochastically as the difference between obtained and forgone rewards; note that this does not require separately estimating the average values themselves. This algorithm instead works directly with the policy w.

To write the gradient rule in a way that relates more directly to the standard case when only the chosen reward is received, we formulate it in terms of the chosen and unchosen rewards (rather than left and right), so that

$$w_{t+1} = \begin{cases} \eta w_t + \alpha P_R P_L \delta_{w,t}; & c_{s,t} = L \\ \eta w_t - \alpha P_R P_L \delta_{w,t}; & c_{s,t} = R \end{cases} \quad (9)$$

with error term

$$\delta_{w,t} = R_c - \kappa R_u, \quad (10)$$

stepsize parameter α, decay parameter (to allow learning in the case of nonstationarity, as in the task here) η, and initial w (w_0). The final free parameter, κ, again allows for the gradient to skew (e.g., due to differential attention) toward the chosen or unchosen reward (R_c and R_u, respectively).

Imaging acquisition. Functional images [T2*-weighted echo-planar images with blood oxygenation level-dependent (BOLD) contrast] were collected using a 3T Siemens Allegra head-only scanner and a Nova Medical NM-011 head coil. To optimize functional sensitivity in the orbitofrontal cortex and temporal lobes, we used a tilted acquisition oriented at 30° above the anterior–posterior commissure line (Deichmann et al., 2003). This yielded 33 oblique-axial slices with 3 mm interslice thickness, 3 × 3 mm in-plane resolution, with coverage from the base of the orbitofrontal cortex and medial temporal lobes to the superior border of the dorsal anterior cingulate cortex. Repetition time was 2 s. Subjects' heads were restrained with plastic pads to minimize head movement during the experiment. A T1-weighted structural image (MPRAGE sequence, 1 × 1 × 1 mm) was acquired after the functional run for each subject to allow localization of functional activity. High-pass filtering with a cutoff period of 128 s was also applied to the data.

Functional imaging analysis. Imaging data were preprocessed and analyzed using SPM5 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, U.K.) and xjView (http://www.alivelearn.net/xjview/), except for multiple-comparisons correction on final results, which was done using SPM8. Motion effects were corrected by aligning images in each run to the first volume using a six-parameter rigid body transformation. (To account for additional residual effects of movement, the six scan-to-scan motion parameters produced during realignment were also included as nuisance regressors in the functional analysis.) Mean functional images were then coregistered to the structural image and normalized into MNI template space using a 12-parameter affine transformation (SPM5, segment and normalize, estimated from the structural). Normalized functional images were resampled into 2 × 2 × 2 voxel resolution. A Gaussian kernel with a full width at half maximum of 6 mm was applied for spatial smoothing.

For statistical analysis, we constructed three impulse events for each trial at the times of slot machine presentation, choice entry, and outcome presentation. The first two events were included to control the overall variance; we focused on the outcome event here due to the fact that the key prediction error signal is associated with the outcome. In three separate general linear models (GLMs), we modulated the outcome events with different parametric regressors. First, in an initial attempt to seek activity correlated with teaching signals for either policy gradient or prediction errors associated with chosen and unchosen choices, we constructed a policy gradient regressor as $R_c - \kappa R_u$, the difference between the obtained and forgone reward, and also prediction error regressors for both chosen and unchosen options, $R_c - Q_c$ and $R_u - Q_u$. These were entered into two different GLMs (one for the policy signal and one for both reward prediction errors) for whole-brain regression analyses. Next, to more carefully differentiate effects related to either type of signal, we noted that Equations 3 and 10 both consist of different linear combinations of the four variables R_c, R_u, Q_c, and Q_u; we thus constructed a third GLM containing each of these as parametric regressors modulating the outcome impulse event.

For all these analyses, the chosen and unchosen rewards were taken directly from the outcomes experienced by the subject, and the Q values were those implied by the Q learning model on each trial, using the outcomes experienced by the subject and the free parameters for the model estimated to best fit the subject's choice data. We then convolved these regressors with SPM5's canonical hemodynamic response function, computed parameter estimates for each subject, and took these estimates to the group random-effects level for statistical testing (Friston et al., 1995).

We report significance of activations correcting for whole-brain multiple comparisons using cluster-level false discovery rate (FDR) algorithm implemented in SPM8 on maps generated at an underlying uncorrected threshold of $p < 0.001$ [note that we did not employ voxel-level FDR, which has recently been argued to be invalid (Chumbley and Friston, 2009)]. Accordingly, except where noted, activations are rendered for display using the $p < 0.001$ uncorrected

threshold, but retaining only those clusters that pass the $p < 0.05$ cluster-size correction.

Finally, for the regions of interest (ROI) regression analysis, we first identified two ROIs in striatum using the conjunction (Nichols et al., 2005) of R_c and $-R_u$ (thresholded at $p < 0.001$, uncorrected). We used the average activity from each of these regions for the subsequent regression between the neural effect of R_u (the per-subject regression weight for that variable) and the behavioral effect (the per-subject estimate of κ from the best fitting policy gradient model). Note that this approach largely skirts the problem of multiple comparisons in intersubject regression analyses, since the initial analysis to identify the ROI (the existence of a conjunction effect in the mean across subjects) does not bias the subsequent test for the between-subject pattern of variation. Significance levels need thus be corrected only for two comparisons (two ROIs, left and right) rather than for the whole-brain multiple comparisons used to select the regions.

All results reported herein were qualitatively the same when Q values were computed using a common set of parameters across subjects, taken as the average over all subjects of those from the individual fits. Results were also invariant to changes in the ordering of the entry of regressors in the design matrix (which, due to serial orthogonalization of parametric regressors in SPM, might hypothetically have impacted their relative significance).

Table 2. Qualities of behavioral fits of both models

	Direct actor	Q-learning
−LL	107.6	114.3
Pseudo-R^2	0.3534	0.3131
Number of parameters	4	5
BIC	118.6	128.0

Q-learning and policy-gradient models were fit to 20 subjects individually. Average quantities are reported. −LL, Negative log likelihood; BIC, Bayesian information criterion.

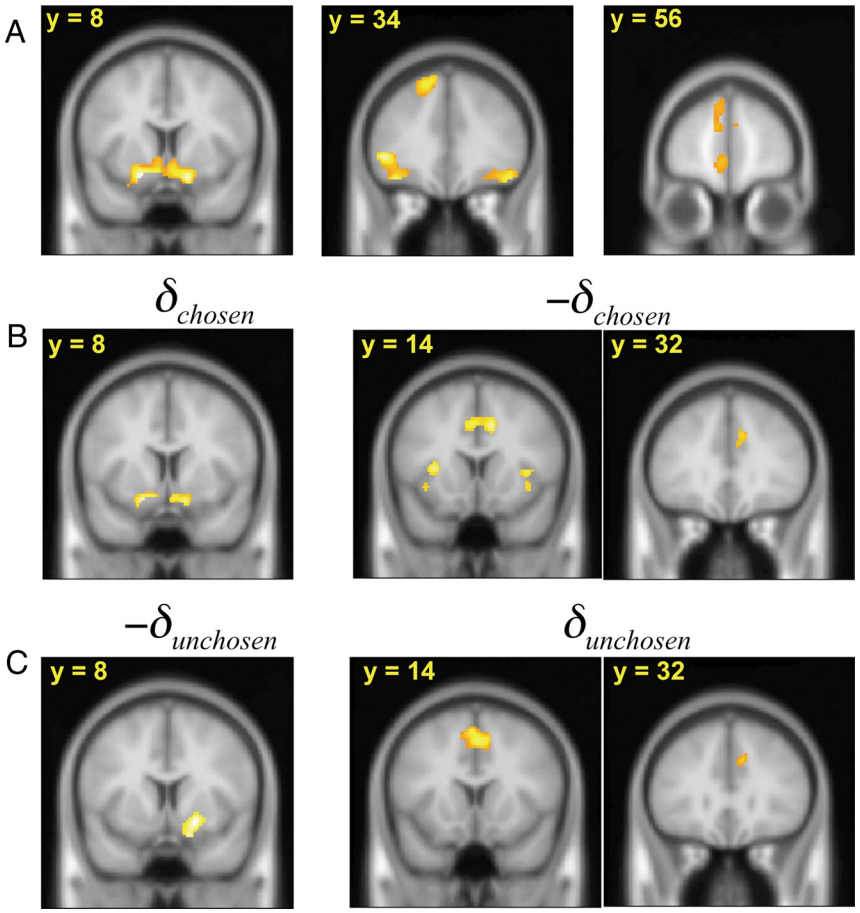

Figure 2. ***A***, Neural correlates of $R_c - \kappa R_u$. ***B***, ***C***, Overlapping view of the neural correlates of prediction errors of chosen choices (***B***, δ_{chosen} and $-\delta_{chosen}$) and forgone choices (***C***, $\delta_{unchosen}$ and $-\delta_{unchosen}$). $p < 0.05$.

Results

Forgone reward and action selection

First, we assessed subjects' behavioral sensitivity to the experienced and forgone rewards by using logistic regression to predict each subject's choices as a function of the feedback she received on the previous trial (Table 1). Across subjects, both experienced ($t_{(19)} = 2.27$, $p < 0.02$) and forgone ($t_{(19)} = 3.94$, $p < 0.001$) rewards significantly predicted the subject's next choice, with no significant difference in the strength between the two influences (paired samples, $t_{(19)} = 0.84$, $p > 0.4$). This result is consistent with the previous literature on counterfactual effects, indicating that humans' choices are affected not only by "what was" but also by "what might have been" (Camille et al., 2004; Coricelli et al., 2005; Lohrenz et al., 2007).

We used two RL models to examine more detailed hypotheses about how experience drove choices via learning (Sutton and Barto, 1998; Dayan and Abbott, 2001; Bhatnagar et al., 2008). A standard value-based model, Q-learning (Watkins and Dayan, 1992), separately tracks the expected value for each option and compares these predicted values at choice times to derive a policy for action selection. However, a policy-based model, the direct actor, learns a policy representing the probability of choosing either action, with an update that is determined by stochastic gradient ascent on the overall expected reward (Dayan and Abbott, 2001; Bhatnagar et al., 2008). Both models were straightforwardly adapted to incorporate forgone rewards, and both included an additional free parameter, κ, to allow for a possible difference in weighting (e.g., attention to) experienced rewards and forgone rewards (see Materials and Methods) (Table 1). We estimated free parameters and compared the models' fits by maximizing the likelihood of each subject's choices (Table 1). The policy-based model generally performed better than the value-based model. Individually, the policy-based model outperformed the Q-learning model for 19 of 20 subjects (Table 2) according to the Bayesian information criterion.

Neural correlates of Q-learning and policy gradient

Choice behavior was thus most consistent with a policy-based strategy. However, because the behavioral predictions of the two learning strategies are qualitatively similar to one another, we next tested for neural signatures of teaching signals, about which the two hypotheses make quite qualitatively distinct predictions. Specifically, learning the value of each option separately requires two independent prediction error signals, measuring the difference between the rewards received (R_c and R_u) and expected (Q_c and Q_u) for both chosen and unchosen options ($R_c - Q_c$ and $R_u - Q_u$) (see Materials and Methods, above) (Camerer and Ho, 1999). In contrast, due to the symmetry of the task, updating

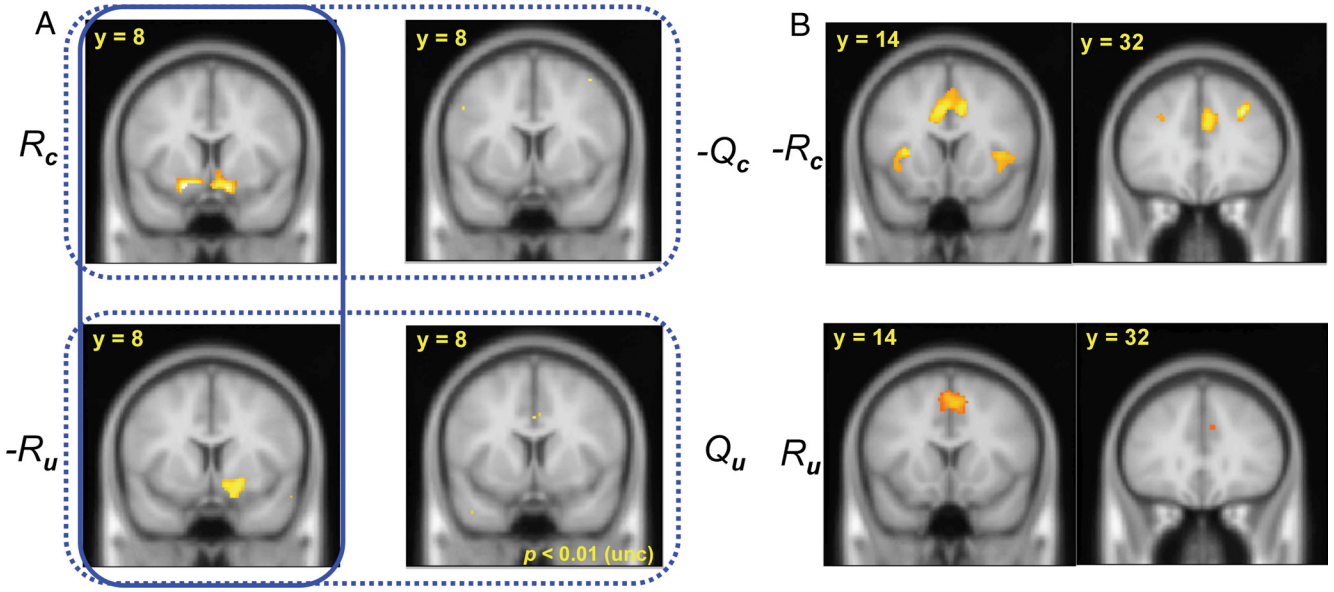

Figure 3. ***A***, Effect of decision variables in striatum. BOLD activity positively correlates with R_c and $-R_u$ ($p < 0.05$) but not with $-Q_c$ or Q_u, even at a loose threshold [$p = 0.01$, uncorrected (unc)]. The dotted boxes surround the pairs of effects expected to be significant for value-based learning [$R_c - Q_c$ and $-(R_u - Q_u)$] and the solid box surrounds those for policy learning ($R_c - \kappa R_u$). ***B***, Similar activities (bilateral insula, anterior cingulate cortex, and dorsolateral prefrontal cortex) were positively correlated with $-R_c$ and R_u ($p < 0.05$).

action preferences requires a unitary signal (formally, the gradient of the expected reward with respect to the policy) proportional to the difference between the obtained and forgone rewards, $R_c - \kappa R_u$.

As a first step, we verified whether these three candidate error signals correlated with BOLD activity in each trial at the time when outcomes were revealed and learning was expected to take place for both *Q*-learning and policy gradient models. These three signals were extracted from each model using the parameters that best fit the behavior. We focused on ventral striatum, where many previous studies using fMRI have shown error-related neural activity (Berns et al., 2001; McClure et al., 2003; O'Doherty et al., 2003, 2004; Delgado et al., 2005; Daw and Doya, 2006). Accordingly, both the prediction error for the chosen action's value in the value-based model, $R_c - Q_c$, and that for the policy in the policy-based model, $R_c - \kappa R_u$, correlated positively with BOLD activities in the ventral striatum ($p < 0.05$; unless otherwise noted, all statistics concerning fMRI activations are corrected at the cluster level for false discovery rate due to whole-brain multiple comparisons) (Fig. 2*A*,*B*); the prediction error for the unchosen option, $R_u - Q_u$, correlated negatively with activity in a similar region ($p < 0.05$) (Fig. 2*C*).

The finding that error signals from both models were correlated with BOLD activity in ventral striatum presumably reflects the fact that the candidate error signals are themselves mutually correlated, necessitating finer investigation to separate them. For instance, the chosen value error ($R_c - Q_c$) and policy error ($R_c - \kappa R_u$) signals might potentially each correlate significantly with striatal BOLD in virtue of sharing the same term R_c. More particularly, all these signals consist of different linear combinations of the same variables: the chosen and unchosen rewards, R_c and R_u; and the expected rewards, Q_c and Q_u. Taking advantage of this decomposition and the additivity of the GLM to examine more carefully which error signal was most consistent with striatal BOLD activity, we estimated a GLM that included these four factors as separate regressors. The policy-based model predicts that R_c and R_u should together modulate any teaching-related brain activity in the same voxels but with opposite directions (see Materials and Methods); the weighted combination of these two effects corresponds to the policy error signal. In contrast, the value-based model (*Q*-learning) makes no particular claim about spatial overlap or relative direction of R_c and R_u effects (which, according to this hypothesis, are part of separate error signals that are not necessarily spatially distinct). However, this value-based model does predict that each effect (R_c or R_u) should be accompanied by spatially overlapping activity correlated with its correspondingly associated predictions (Q_c or Q_u), with opposite signs for the *R* and *Q* components ($R_c - Q_c$ and $R_u - Q_u$).

Activity in bilateral ventral striatum positively correlated with R_c ($p < 0.05$) (Fig. 3*A*), which is a component of teaching signals in both hypotheses. However, as predicted by the policy model, R_u was negatively correlated with BOLD activity in substantially the same region of striatum ($p < 0.05$) (Fig. 3*A*). A conjunction analysis confirmed that these two effects occurred in overlapping voxels (Fig. 4*A*, bilaterally at $p < 0.001$, uncorrected; the cluster on the right survived whole-brain correction at $p < 0.05$). Furthermore, we tested for differences in the spatial expression of these effects by using the contrasts $R_c > -R_u$ and $R_c < -R_u$ to identify voxels where the positive effect of R_c and the negative effect of R_u differed in size. Consistent with the policy model, no such differences were found in striatum (at $p = 0.05$). Finally, we tested whether value prediction signals Q_c and Q_u were correlated with striatal BOLD activity in the directions opposite to their associated rewards (Fig. 3*A*), as predicted by the value-based model. However, no such correlations were found for either signal, even at a much weaker threshold (Fig. 3*A*, $p < 0.01$, uncorrected) and in either positive or negative directions. In all, the results are consistent with net ventral striatal BOLD activity expressing a single teaching signal proportional to $R_c - \kappa R_u$, as predicted by the policy gradient algorithm, but do not demonstrate any evidence for a pair of value prediction errors for each of the chosen and unchosen options, either overlapping or spatially separate (Lohrenz et al., 2007). Indeed, the results also are not consistent with an alternative value-based decision variable, i.e., the relative action value ($Q_c - Q_u$), since the prediction error, $(R_c - R_u) - (Q_c - Q_u)$, for this combined

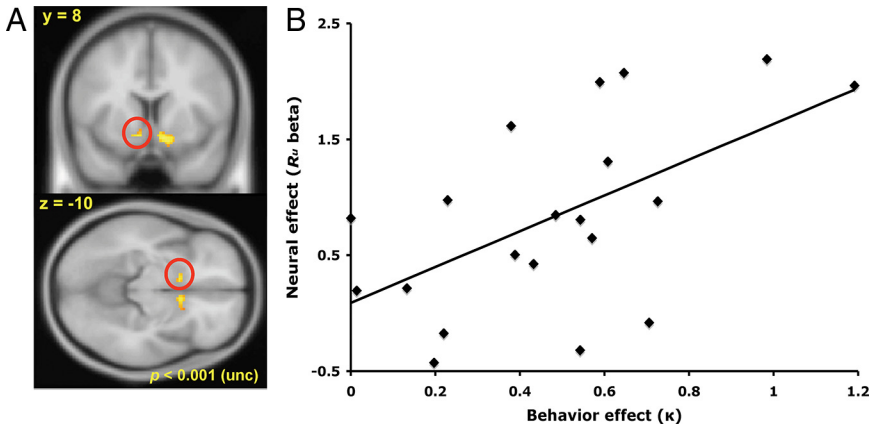

Figure 4. ***A***, Error signaling ROI in left ventral striatum, identified from the conjunction of R_c and $-R_u$ across subjects [$p <$ 0.001, uncorrected (unc)]. ***B***, In left striatum (***A***, circles), the neural effect size for $-R_u$ was positively correlated, across subjects, with the weight for the unchosen reward, κ, estimated from choice behavior ($p = 0.018$, Bonferroni corrected; $r = 0.57$).

quantity also predicts effects of both *Q*s. The lack of significant value prediction-related activity in our study seems to contrast with many other studies in which striatal BOLD activity was demonstrably modulated by value expectation, as for prediction errors (Berns et al., 2001; O'Doherty et al., 2003; Tanaka et al., 2004; Hare et al., 2008). This apparent difference may be due to the design of the current task, which differs from many others in that, because of its symmetric form and the inclusion of counterfactual feedback, the teaching signal for the policy contains no reward expectation term.

Similarly, and strikingly, throughout the rest of the brain, activity correlated with R_c and R_u was observed in almost the same set of neural pathways (Fig. 3*B*), but with opposite directions of effect. Evidence for spatially nonoverlapping effects was found ($R_c > -R_u$, $p < 0.05$) only in posterior portions of the brain (occipital visual cortex and fusiform areas). Over the whole brain, no activity was found to correlate positively or negatively with Q_c or Q_u, even at a lower threshold ($p < 0.01$ uncorrected).

Correlation of behavioral and neural sensitivities to forgone reward

Finally, we compared neural and behavioral variation across subjects to investigate whether forgone outcome signaling in ventral striatum was related to choice behavior. In particular, we tested whether, across subjects, there was covariation between estimates of the weight (e.g., attention) given to the forgone outcome as assessed from choice behavior (the parameter κ from the policy-gradient model) and from neural error-related activity in striatum (the effect size for the forgone reward). We first identified areas of error signaling in left and right ventral striatum (using the conjunction of R_c and $-R_u$ effects) (Fig. 4*A*), then tested within each cluster whether the contribution of R_u to this signal covaried across subjects with the weight to forgone rewards estimated from choice behavior. The predicted correlation between the behavioral and neural effects of the forgone reward was significant in left striatum ($p = 0.018$; Bonferroni corrected for two comparisons, left and right striatum) (Fig. 4*A*, red circle, *B*), and trended in the same direction, though not significantly so, in right striatum ($p = 0.45$ corrected) (data not shown), suggesting that both neural and behavioral analyses consistently tapped a common learning process.

Discussion

A longstanding question in psychology—dating back to early debates surrounding behaviorism (Thorndike, 1898; Tolman, 1949; Dickinson and Balleine, 2002)—is the representational question: what exactly is learned from reinforcement? Error-driven RL theories are surprisingly ambivalent on this issue. Value-based (Q-learning) models propose that the difference between obtained and predicted rewards is used to update expected action values, and choice policies are subsequently derived by comparing these intermediate quantities (Barraclough et al., 2004; Daw et al., 2005; O'Reilly and Frank, 2006; Hare et al., 2008; Boorman et al., 2009). Policy-based approaches instead update a choice policy directly, though often alongside a value representation. What makes these two approaches difficult to differentiate is that in most circumstances, the policy update signal, derived from the gradient of the expected rewards with respect to the policy, also takes the form of a difference between obtained and expected rewards. Indeed, the signals are so similar that the prominent actor/critic algorithm actually uses the same error signal to update both state values and policies.

Here, we studied a task in which this is not the case, allowing us to distinguish a teaching signal for direct policy preferences from signals for learning value predictions. In particular, forgone reward takes the place of the expected reward in a policy teaching signal, but not in an action value teaching signal. We found evidence that net outcome-related BOLD activity in the striatum is appropriate to learning policies, but no similar evidence for signals appropriate for updating separate action values. That said, the latter (negative) conclusion relies on a stronger test for PE signaling than is often used in the literature. In many studies, including ours (Fig. 2), striatal BOLD activity correlated with a PE signal for *Q* values. We decomposed the activity to show that, in our task, this correlation was likely due only to outcome- and not prediction-related activity, supporting the policy model. Although most previous authors did not report this particular analysis, many showed other evidence that the striatal BOLD response was modulated by predictions as well as outcomes [e.g., by including outcome as an effect of no interest or contrasts between expected and unexpected outcomes (Berns et al., 2001; McClure et al., 2003; Li et al., 2006, 2011; Hare et al., 2008)]. Interestingly, Behrens and colleagues (2007, 2008) separately tested both effects in the same manner we did and also found no effect of predictions. This may be because in their task, like ours, forgone rewards were known, so that expectancies would also disappear from the policy update signal.

A number of recent studies suggest that error-related BOLD activity in striatum may reflect, at least in part, the dopaminergic input from midbrain (Pessiglione et al., 2006; Knutson and Gibbs, 2007; Schonberg et al., 2010). Of course, since the fMRI BOLD signal is a generic metabolic signal not specific to a single underlying neural cause, unit recordings will be required to determine whether our results generalize to the prediction error responses of midbrain dopamine neurons.

Our results do, however, provide positive evidence in the human brain for a teaching signal specifically appropriate for updating action policies. Unlike error signals previously reported in other tasks, which did not probe the distinction since value and policy errors coincided (Daw et al., 2006; Hampton et al., 2006; Morris et al., 2006; Schonberg et al., 2007, 2010), this signal can-

not alternatively be interpreted as a prediction error for values. Perhaps the best previous evidence for a policy-specific update signal was an influential report of spatially distinct correlates in striatum of a prediction error during a free-choice compared with an instructed-choice condition (O'Doherty et al., 2004; also see Tricomi et al., 2004). Although in that task also, modeled value and policy teaching signals were substantially the same and activity in the dorsal striatum was specific for the free-choice condition and interpreted as a policy teaching signal on that basis. Interestingly, standard actor/critic models do not obviously predict that policy teaching signals will be specific to free-choice conditions; instead, they assert that both actor (policy learning) and critic (value learning) modules just use a common error signal (Barto, 1995). Here, we skirt this interpretational difficulty by investigating whether the information carried by the signal itself is appropriate for training policies or values.

The comparison with the O'Doherty et al. (2004) study also points to another unresolved puzzle in both datasets (and in studies of striatal error signaling in humans more generally), which is that lesion work in rodents identifies stimulus–response policy learning specifically within the dorsolateral striatum (Yin et al., 2004), which corresponds to the dorsal putamen in humans. Most PE-related activity in humans, including, surprisingly, the policy errors in the present study, is instead focused more ventrally in ventral putamen and caudate (but see Schonberg et al., 2010). Meanwhile, in the O'Doherty et al. (2004) study, the putative policy error in the free-choice condition was localized in the dorsal caudate (dorsomedial striatum), which is not consistent with stimulus–response learning in rats (Yin et al., 2005). Thus, further work remains to detect counterparts in human striatal error signaling of the striatal suborganization suggested by rodent work (Tricomi et al., 2009; Schonberg et al., 2010).

Alongside our positive evidence for a policy update signal, our accompanying failure to detect evidence for value update signals in the present task should not be interpreted as contradicting the overwhelming evidence that neural signals, including those in the nigrostriatal system, reflect action value expectations in many other circumstances (Barraclough et al., 2004; Daw et al., 2005; Samejima et al., 2005; O'Reilly and Frank, 2006; Hare et al., 2008; Rangel et al., 2008; Seo and Lee, 2008; Boorman et al., 2009). First, we cannot rule out that our results are idiosyncratic to the present task. For instance, subjects' brains might have adopted the direct actor strategy here because it's particularly straightforward and efficient due to the symmetric information structure of our task design. Anyway, a true neural value representation would be required even under the wildest possible extrapolation from the current results—that all previously reported error signals, too, can be understood as policy rather than value update signals—since in those tasks even the error signal for policies itself contains value predictions.

The necessity of value predictions for producing an error signal (for policies) does not imply that those value predictions are themselves also learned from that error signal, although the latter is also assumed in standard models like the actor/critic. For instance, even if value predictions are not trained by a nigrostriatal error signal, they may arise from prediction processes involving other brain systems. A longstanding idea in psychology and cognitive neuroscience is that the brain relies on multiple learning and memory systems, which can each learn different representations (Knowlton et al., 1996; Poldrack et al., 2001; Dickinson and Balleine, 2002; Daw et al., 2005). In addition to values directly learned from prediction errors, value predictions are also believed to be constructed from still more fundamental information—e.g., from action-outcome associations, cognitive maps, or model-based RL, often associated with prefrontal or declarative memory systems (Poldrack et al., 2001; Dickinson and Balleine, 2002; Daw et al., 2005). These predictions may, in turn, inform nigrostriatal signals for policy update (Doll et al., 2009; Daw et al., 2011; Simon and Daw, 2011).

Such model-based evaluation processes likely played a more prominent role in a pair of recent studies of counterfactual effects on choice (Camille et al., 2004; Coricelli et al., 2005), because in those studies subjects had to evaluate new options on each trial based on a pictorial description, rather than learn a choice policy by trial-and-error as in the present task. The requirement for such explicit evaluation may help to explain why, in those studies, the level of regret related to forgone options was specifically associated with distinct neural correlates in orbital prefrontal cortex, whereas both actual and forgone rewards engaged largely identical brain networks in the present study.

In another related study, Lohrenz et al. (2007) reported correlations in striatal BOLD with fictive error signals in a decision task in which subjects learned how much to bet on a simulated trading market. This fictive error signal (the difference between the obtained reward and the largest possible reward) can be viewed as another instance of the policy gradient signal reported here, and indeed, the authors noted this potential interpretation. However, because that task involved only a single dimension of action (how much to bet) with a single reward (the movement of the market), rather than two actions with two independent rewards, it would have been difficult to separate the constituents of the signal as we have done in our task, so as to distinguish a counterfactual policy update signal (how much to move the bet policy toward the all-in bet) from other correlated quantities such as a counterfactual value update signal (e.g., updating a stored value prediction for the all-in bet).

Our findings thus provide specific new evidence about the computational form and function of error-driven learning in the human striatum, adding to an accumulating body of evidence about these processes. The suggestion that these signals directly reinforce choice policies, rather than training value predictions that serve as intermediate quantities in evaluating candidate choices, may have particular relevance for dysfunctions involving compulsive action, such as drug abuse, gambling, and impulse control disorders (Dickinson and Balleine, 2002; Everitt and Robbins, 2005).

References

Barraclough DJ, Conroy ML, Lee D (2004) Prefrontal cortex and decision making in a mixed-strategy game. Nat Neurosci 7:404–410.

Barto AG (1995) Adaptive critics and the basal ganglia. In: Models of information processing in the basal ganglia (Houk JC, Davis J, Beiser D, eds), pp. 215–232. Cambridge, MA: MIT.

Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. Nat Neurosci 10:1214–1221.

Behrens TE, Hunt LT, Woolrich MW, Rushworth MF (2008) Associative learning of social value. Nature 456:245–249.

Berns GS, McClure SM, Pagnoni G, Montague PR (2001) Predictability modulates human brain response to reward. J Neurosci 21:2793–2798.

Bhatnagar S, Sutton RS, Ghavamzadeh M, Lee M (2008) Incremental natural actor-critic algorithms. In: Advances in neural information processing systems 20 (Platt J, Koller D, Singer Y, Roweis S, eds), pp. 105–112. Cambridge, MA: MIT.

Boorman ED, Behrens TE, Woolrich MW, Rushworth MF (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.

Camerer C, Ho TH (1999) Experience-weighted attraction learning in normal form games. Econometrica 67:827–874.
Camille N, Coricelli G, Sallet J, Pradat-Diehl P, Duhamel JR, Sirigu A (2004) The involvement of the orbitofrontal cortex in the experience of regret. Science 304:1167–1170.
Chumbley JR, Friston KJ (2009) False discovery rate revisited: FDR and topological inference using Gaussian random fields. Neuroimage 44:62–70.
Coricelli G, Critchley HD, Joffily M, O'Doherty JP, Sirigu A, Dolan RJ (2005) Regret and its avoidance: a neuroimaging study of choice behavior. Nat Neurosci 8:1255–1262.
Daw ND, Doya K (2006) The computational neurobiology of learning and reward. Curr Opin Neurobiol 16:199–204.
Daw ND, Niv Y, Dayan P (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8:1704–1711.
Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876–879.
Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. Neuron 69:1204–1215.
Dayan P, Abbott LF (2001) Theoretical neuroscience: computational and mathematical modeling of neural systems. Cambridge, MA: MIT.
Dayan P, Daw ND (2008) Decision theory, reinforcement learning, and the brain. Cogn Affect Behav Neurosci 8:429–453.
Deichmann R, Gottfried JA, Hutton C, Turner R (2003) Optimized EPI for fMRI studies of the orbitofrontal cortex. Neuroimage 19:430–441.
Delgado MR, Miller MM, Inati S, Phelps EA (2005) An fMRI study of reward-related probability learning. Neuroimage 24:862–873.
Dickinson A, Balleine B (2002) The role of learning in the operation of motivational systems (Pashler H, Gallistel R, eds), p 533. New York: Wiley.
Doll BB, Jacobs WJ, Sanfey AG, Frank MJ (2009) Instructional control of reinforcement learning: a behavioral and neurocomputational investigation. Brain Res 1299:74–94.
Everitt BJ, Robbins TW (2005) Neural systems of reinforcement for drug addiction: from actions to habits to compulsion. Nat Neurosci 8:1481–1489.
Friston KJ, Frith CD, Frackowiak RS, Turner R (1995) Characterizing dynamic brain responses with fMRI: a multivariate approach. Neuroimage 2:166–172.
Hampton AN, Bossaerts P, O'Doherty JP (2006) The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J Neurosci 26:8360–8367.
Hare TA, O'Doherty J, Camerer CF, 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.
Kass RE, Raftery AE (1995) Bayes factors. J Amer Stat Assoc 90:773–795.
Knowlton BJ, Mangels JA, Squire LR (1996) A neostriatal habit learning system in humans. Science 273:1399–1402.
Knutson B, Gibbs SE (2007) Linking nucleus accumbens dopamine and blood oxygenation. Psychopharmacology 191:813–822.
Lau B, Glimcher PW (2005) Dynamic response-by-response models of matching behavior in rhesus monkeys. J Exp Anal Behav 84:555–579.
Li J, McClure SM, King-Casas B, Montague PR (2006) Policy adjustment in a dynamic economic game. PLoS One 1:e103.
Li J, Delgado MR, Phelps EA (2011) How instructed knowledge modulates the neural systems of reward learning. Proc Natl Acad Sci U S A 108:55–60.
Lohrenz T, McCabe K, Camerer CF, Montague PR (2007) Neural signature of fictive learning signals in a sequential investment task. Proc Natl Acad Sci U S A 104:9493–9498.
McClure SM, Berns GS, Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. Neuron 38:339–346.
Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H (2006) Midbrain dopamine neurons encode decisions for future action. Nat Neurosci 9:1057–1063.
Nichols T, Brett M, Andersson J, Wager T, Poline JB (2005) Valid conjunction inference with the minimum statistic. Neuroimage 25:653–660.
O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ (2003) Temporal difference models and reward-related learning in the human brain. Neuron 38:329–337.
O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston K, Dolan RJ (2004) Dissociable roles of ventral and dorsal striatum in instrumental conditioning. Science 304:452–454.
O'Reilly RC, Frank MJ (2006) Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. Neural Comput 18:283–328.
Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD (2006) Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature 442:1042–1045.
Poldrack RA, Clark J, Paré-Blagoev EJ, Shohamy D, Creso Moyano J, Myers C, Gluck MA (2001) Interactive memory systems in the human brain. Nature 414:546–550.
Rangel A, Camerer C, Montague PR (2008) A framework for studying the neurobiology of value-based decision making. Nat Rev Neurosci 9:545–556.
Roesch MR, Calu DJ, Schoenbaum G (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. Nat Neurosci 10:1615–1624.
Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of action-specific reward values in the striatum. Science 310:1337–1340.
Schonberg T, Daw ND, Joel D, O'Doherty JP (2007) Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. J Neurosci 27:12860–12867.
Schonberg T, O'Doherty JP, Joel D, Inzelberg R, Segev Y, Daw ND (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.
Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599.
Seo H, Lee D (2008) Cortical mechanisms for reinforcement learning in competitive games. Philos Trans R Soc Lond B Biol Sci 363:3845–3857.
Simon DA, Daw ND (2011) Neural correlates of forward planning in a spatial decision task in humans. J Neurosci, in press.
Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Cambridge, MA: MIT.
Tanaka SC, Doya K, Okada G, Ueda K, Okamoto Y, Yamawaki S (2004) Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. Nat Neurosci 7:887–893.
Thorndike EL (1898) Animal intelligence: an experimental study of the associative processes in animals. New York: Macmillan.
Tolman EC (1949) Purposive behavior in animals and men. Los Angeles: University of California.
Tricomi E, Balleine BW, O'Doherty JP (2009) A specific role for posterior dorsolateral striatum in human habit learning. Eur J Neurosci 29:2225–2232.
Tricomi EM, Delgado MR, Fiez JA (2004) Modulation of caudate activity by action contingency. Neuron 41:281–292.
Watkins CJ, Dayan P (1992) Q-learning. Mach Learn 8:279–292.
Yin HH, Knowlton BJ, Balleine BW (2004) Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur J Neurosci 19:181–189.
Yin HH, Ostlund SB, Knowlton BJ, Balleine BW (2005) The role of the dorsomedial striatum in instrumental conditioning. Eur J Neurosci 22:513–523.