Behavioral/Cognitive

Hierarchical Learning Induces Two Simultaneous, But Separable, Prediction Errors in Human Basal Ganglia

Carlos Diuk,[1] Karin Tsai,[2] Jonathan Wallis,[3] Matthew Botvinick,[1] and Yael Niv[1]

[1]Psychology Department and Princeton Neuroscience Institute, Princeton University, Princeton, New Jersey 08544, [2]Computer Science Department, Princeton University, Princeton, New Jersey 08540, and [3]Department of Psychology and Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, California 94720

Studies suggest that dopaminergic neurons report a unitary, global reward prediction error signal. However, learning in complex real-life tasks, in particular tasks that show hierarchical structure, requires multiple prediction errors that may coincide in time. We used functional neuroimaging to measure prediction error signals in humans performing such a hierarchical task involving simultaneous, uncorrelated prediction errors. Analysis of signals in a priori anatomical regions of interest in the ventral striatum and the ventral tegmental area indeed evidenced two simultaneous, but separable, prediction error signals corresponding to the two levels of hierarchy in the task. This result suggests that suitably designed tasks may reveal a more intricate pattern of firing in dopaminergic neurons. Moreover, the need for downstream separation of these signals implies possible limitations on the number of different task levels that we can learn about simultaneously.

Introduction

In recent years, computational reinforcement learning (RL; Sutton and Barto, 1998) has provided an indispensable framework for understanding the neural substrates of learning and decision making. Dopaminergic signals projecting into the striatal nuclei, once elusive and misunderstood, are now widely thought to be correlated with a scalar prediction error signal that indicates the difference between reward expectations and actual observations (Barto, 1995; Montague, Dayan, and Sejnowski, 1996). This prediction error signal is key for learning about rewards in the world and is a central element in RL models of learning.

Although the original studies of dopaminergic prediction errors suggested that dopaminergic neurons all report one unitary scalar prediction error signal (Schultz et al., 1997; e.g., Schultz, 2002), computational RL models that attempt to scale beyond simple action-outcome associations into real-world tasks suggest that more than one prediction error may become necessary at each point in time (Sutton et al, 1999). In the present study, we asked the question: can the classic neural correlates of reward prediction errors support more than one prediction error signal type?

Tasks with hierarchical structure constitute one example in which multiple, simultaneous reward prediction errors are needed. This is because in hierarchical settings, outcomes relevant to multiple levels of a task structure might be observed at the same time, and the brain must update its expectations about each level separately. For example, imagine a gambler who arrives at a city with multiple casinos, holding a set of coupons that allow him to enter any one of the casinos and play a number of different games. The gambler enters one casino and plays blackjack, roulette, and a slot machine. Each time he plays a game, he might observe a difference between what he expected to win, and the actual outcome—a "game-level prediction error" that can be used to adjust his future expectations about this game. However, upon playing the last coupon for a casino, he not only learns about the last game itself, but also has enough information to update his knowledge about the casino as a whole: was this a good casino to spend his coupons on? It is at this point that two coincident reward prediction errors would arise: a simple game-related prediction error and a higher-level casino-related prediction error linked to learning the value of the casino as a whole. These prediction errors are not redundant. For example, the slot machine may have been worse than expected but the casino better than expected.

To determine whether concurrent prediction errors occur in the human brain, we designed a task akin to the casino example above—effectively, a hierarchical extension of the classic bandit task used in previous RL research (Daw et al., 2006; Cohen et al., 2007) to a hierarchical setting. We used fMRI to record BOLD signals while participants played this task. We were especially interested in BOLD signals in the ventral striatum (VS), an area

Received Nov. 25, 2012; revised Feb. 5, 2013; accepted Feb. 14, 2013.

Author contributions: C.D., K.T., J.W., M.B., and Y.N. designed research; C.D. and K.T., performed research; C.D., M.B. and Y.N. analyzed data; C.D., M.B., and Y.N. wrote the paper.

This work was supported by the John Templeton Foundation and the National Science Foundation–Collaborative Research in Computational Neuroscience (Grant IIS-1207833). The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation or the National Science Foundation. Y.N. is supported by an Alfred P. Sloan Research Fellowship. C.D. and M.B. were both supported by a Collaborative Activity Award from the James S. McDonnell Foundation. We thank Peter Dayan for extremely helpful comments during the preparation of this manuscript. We also thank Timothy Behrens for useful suggestions.

The authors declare no competing financial interests.

Correspondence should be addressed to either Carlos Diuk or Yael Niv, Psychology Department and Princeton Neuroscience Institute, Green Hall, Princeton, NJ 08540, E-mail: cdiuk@princeton.edu or yael@princeton.edu.

K. Tsai's present address: Computer Science Department, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

DOI:10.1523/JNEUROSCI.5445-12.2013

 0270-6474/13/335797-09$15.00/0

where activity has been shown repeatedly to be correlated with prediction error signals (Hare et al., 2008; Glimcher, 2011; Niv et al., 2012), as well as the ventral tegmental area (VTA) from which dopamine neurons arise. To model learning in this setting, we used the computational framework of hierarchical RL (HRL; Sutton et al, 1999; Dietterich, 2000; Barto and Mahadevan, 2003), an extension of the RL framework for hierarchical settings that was shown recently to be relevant to human learning (Botvinick et al., 2009; Ribas-Fernandes et al., 2011).

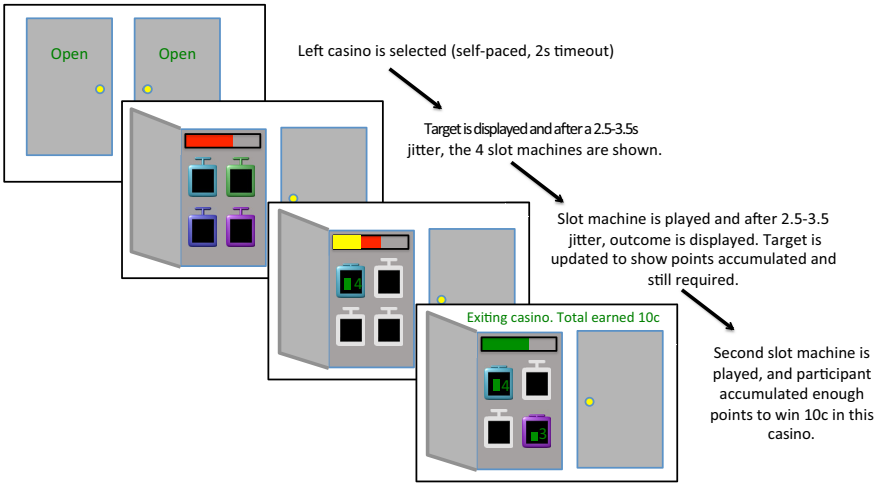

Figure 1. Sample trial: the participant chooses to play in the left casino, the door opens and displays a target number of points (indicated by red bar). After 2.5–3.5 s, the four slot machines appear. The participant plays upper-left slot and, after another 2.5–3.5 s, the points obtained in that machine are shown inside the machine (as a green bar plus a roman numeral). The corresponding part of the target points turns yellow, indicating the points accumulated with the first slot machine play. The rest is still red, indicating the points still necessary to win the casino. The participant plays the bottom-right slot machine and obtains sufficient points to win the casino. The target bar turns green and a message appears indicating the casino win (10¢).

Materials and Methods

Participants

Thirty participants were recruited from the Princeton University community and gave informed consent. Two participants were excluded due to technical problems during scanning and all data analysis was performed on the remaining 28 participants (ages 18–38, mean 22.04 years, 13 males, all right-handed). Participants received compensation of $20 per hour plus a small bonus based on task performance (participants began the task with a budget of $1 and kept any money earned by playing casinos, resulting in average earnings of $2.34, std = 1.39, min = −0.45, max = 4.55). All experimental procedures were approved by the institutional review board of Princeton University.

Imaging

Functional brain images were acquired using a 3 T Siemens Allegra head-only MRI scanner with a circularly polarized head volume coil. High-resolution (1 mm^3 voxels) T1-weighted structural images were acquired with an MP-RAGE pulse sequence at the beginning of the scanning session. Functional data were acquired using a high-resolution echo-planar imaging pulse sequence (3 mm^3 voxels, 41 contiguous 3 mm thick slices aligned with the anterior commissure-posterior commissure plane, interleaved acquisition, TR 2400 ms, TE 30 ms, flip angle 90°, field of view 192 mm).

Task and procedure

The computerized task was coded using MATLAB (Mathworks) and the Psychophysics Toolbox version 3 (Brainard, 1997). Participants played 120 trials split into four blocks of 30 trials each. Between blocks, they were offered an option to take a break. On each trial, two doors representing the two different casinos appeared (Fig. 1). Each door was marked "Open" or "Closed." In 70 of the 120 trials, both casinos were open, and in 50 trials one of the doors was closed, forcing participants to choose the only open casino. These forced trials, which included 25 trials in which each of the two casinos was open, were interspersed randomly among the 120 trials, and not determined based on participants' actions or earnings. Throughout the task, participants indicated their choices by pressing buttons on a right-hand response trigger box. If no choice was made within 2 s, a message with the text "Timed Out!" appeared and the participant was allowed to keep playing.

Once a participant chose a casino, the casino's door opened and revealed a rectangular bar that graphically indicated, in red, how many points needed to be accumulated to win 10¢ in the casino. This number of "target points" was drawn from a normal distribution with a SD of 2.5 and a mean of 5 for the left casino and 6 for the right casino, rounded to the closest integer and bounded between 2 and 10.

After a jittered time interval that lasted between 2.5 and 3.5 s (uniform distribution), four slot machines were displayed inside the casino, each a different color. Overall, there were eight slot machines in the task, each a unique color and each assigned permanently to one of the casinos. Thus, throughout the task, each casino always contained the same four slot machines in the same four locations, and a slot machine never appeared in both casinos. Participants played the casino by serially selecting two slot machines using one of the four buttons in the trigger box. For every 3 s in which no choice was made, a message with the text "Timed Out!" appeared, a penalty of 5¢ was assessed, and the participant was allowed to keep playing. When the first slot machine was selected, the other three slot machines were temporarily deactivated (graphically depicted by turning gray). The selected slot machine was animated to simulate spinning for 200 ms and then displayed the number of points obtained. The number of points was shown as a green bar inside the slot machine with a Roman numeral to its side. The top bar indicating the casino target points was also updated in the following way: if the points accrued were sufficient to win the casino, the bar turned green; if not, the portion of the total target points just accrued in the slot machine became yellow, with the remaining bar still red. Each slot machine was associated with a normal distribution from which points were drawn (rounded to the closest integer and bounded between 0 and 5). The means of the normal distributions associated with each of the 8 slot machines (4 per casino) drifted independently after each trial by +0.5 or −0.5 (drawn randomly with equal probability) and were bounded between 0 and 5. The SD of the distribution was always 1.

After the first slot machine play, a jittered wait time of 2.5–3.5 s (uniform distribution) was imposed until the remaining slot machines became active again and a second slot machine could be chosen. After the second slot machine was selected, it spun for another 200 ms and displayed the number of accrued points. Once again, the target points bar added the points just accrued and turned green in case of a win or stayed partially yellow/partially red if not. Simultaneously, a message was also displayed indicating the total amount of money earned in the casino and the end of the trial: "Exiting casino. Total Earned 10¢" (or "−10¢"). Trials in which the target points had been achieved resulted in a "win" (and accrual of 10¢), whereas trials in which the target was not reached resulted in a "loss" (and deduction of 10¢ from the total earnings). After a jittered wait time of 2.5–3.5 s (uniform distribution), the casino door closed and a new trial began.

Note that to orthogonalize prediction errors at the two levels of the hierarchy (the casino level and the slot machine level), winnings in the casinos were not based directly on the number of points obtained in slot machines, but instead were thresholded according to the target points required by the casino. In this way, two simultaneous events, learning the outcome of the last slot machine and determining whether sufficient

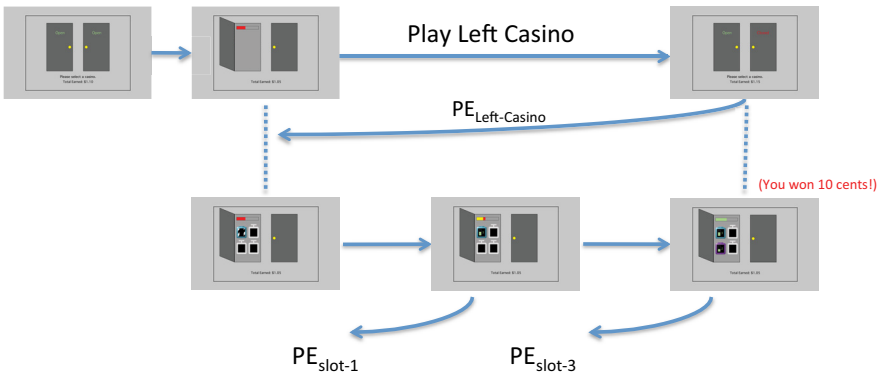

Figure 2. HRL representation of the casino task. The top level shows the task of playing a casino, and the bottom level decomposes this task into the subtasks of playing slot machines. Prediction errors under the Outcome Model and Slot-Points Model are shown (in this example, "slot-3" indicates the name of the slot machine just played). Note that the prediction error for playing the left casino and the second slot machine occur simultaneously.

points were obtained to win in the casino, induced two uncorrelated reward prediction errors. This is because a slot machine could be worse than expected but still lead to an overall win in the casino and vice versa.

Data analysis

Behavioral data

Logistic regression. We first sought to confirm that participants learned at both levels of the task hierarchy and made decisions accordingly. Specifically, the question was whether participants learned in a hierarchical fashion appropriate for this task, adjusting their casino choices based on the outcomes of casino plays and not on the contingencies of slot machine outcomes and, conversely, basing their slot machine choices on slot outcomes and not casino ones. For this purpose, we used a modified logistic regression to estimate the contribution of past outcomes in both levels (slot machines and casino) to behavioral choices. Specifically, we regressed casino choice on a linear combination of the outcomes of the past 4 choices of this casino (+1 for a win and −1 for a loss) and the total number of slot machine points accrued in each of those past 4 casino choices. One can view this as estimating a "value" for each casino as follows:

$$V_{\text{casino}(i)} = \sum_{j=1}^{4} \beta_j^{\text{casino}} R_{\text{casino}(i)}^{t-j} + \sum_{j=1}^{4} \beta_j^{\text{slot}} R_{\text{slots}(i)}^{t-j}$$

where $R_{\text{casino}(i)}^{t-j}$ is +1 (−1) if the participant won (lost) in casino(i) the last jth time she played in it and $R_{\text{slots}(i)}^{t-j}$ is the total number of points obtained in the slot machines during that trial. Casino choices were then logistically regressed on these values using a soft-max action selection function:

$$p(A) = \frac{e^{V(A)}}{\sum_{j \in \text{Actions}} e^{V(j)}}$$

where p(A) is the probability of choosing casino A, and j enumerates all possible actions.

In a similar fashion, values for each slot machine were computed as a linear combination of the points outcome on the last 4 times this machine was played and the casino outcome (+1 or −1) on each of these trials. These were then logistically regressed on slot machine choices using soft-max as above.

For each logistic regression, we optimized the eight parameters $\beta_{1..4}$ for casinos and $\beta_{1..4}$ for slot machines by minimizing the negative log likelihood of the data given different parameter settings using MATLAB's fminunc function. This function performs unconstrained linear optimization over the space of possible parameter values. The resulting β values were then subjected to between-participants Student's t tests to determine the significance of prior outcomes' influences on action selection. We used Bonferroni correction for multiple comparisons, determining significance at a level of $p < 0.05/8 = 0.00625$.

Temporal difference learning. After confirming learning at each level of the task (see Results), we fit a set of alternative temporal-difference RL models to the data to obtain model-based, trial-by-trial prediction-error regressors to be used in our neuroimaging analysis. Temporal difference (TD) RL (Sutton and Barto, 1998) provides a general framework for understanding trial-by-trial learning and decision making in simple tasks. A number of extensions have been proposed to the case where tasks involve HRL structure (Sutton et al., 1999; Barto and Mahadevan, 2003). Central to these HRL algorithms is the notion of temporally extended action sequences. The implication for human behavior is that some decisions no longer involve a single action, but commit the behaving agent to a sequence of actions over an extended period of time. Playing a casino is one such temporally extended action, which involves a series of subsequent slot machine plays (Fig. 2).

Following the "Options" model from HRL (Sutton et al., 1999), we assumed that participants maintain a value V_{casino} for each of the casinos that directs their choice of casinos. We contrasted two possible models for how this value is learned:

The "Outcome Model" posits that casino values are (correctly) based on the probability of winning 10¢ in each casino, as should be the case for participants learning in a hierarchical fashion. In this way, casino values are (implicitly) based on both the target points and the quality of the slot machines in the casino (as well as the policy used to play the slot machines). According to this model, once the participant observes a casino-level outcome r_{casino} (either +10¢ or −10¢) after playing the second slot machine, a prediction error $\delta_{\text{casino}} = r_{\text{casino}} - V_{\text{casino}(i)}$ is computed and the value of the casino is updated based on $V_{\text{casino}(i)}^{\text{new}} = V_{\text{casino}(i)}^{\text{old}} + \eta_{\text{casino}}\delta$, where η_{casino} is a casino-level learning rate or step-size parameter.

The alternative, straightforward but suboptimal "Target Model" posits that casino values are based on the target number of points required to win each casino. This simple model ignores the fact that the quality of a casino also depends on the expected quality of its slot machines and, in a sense, ignores the hierarchical nature of the task. Using this model, once the casino door opens and the target points p_{casino} are revealed, a casino prediction error $\delta_{\text{casino}} = p_{\text{casino}} - V_{\text{casino}(i)}$ is computed and the value of the casino is updated according to $V_{\text{casino}(i)}^{\text{new}} = V_{\text{casino}(i)}^{\text{old}} + \eta_{\text{casino}}\delta$ as above.

For both models, to fit the free model parameters to behavioral choice at the casino level, we assumed a soft-max action selection function:

$$p(A) = \frac{e^{\tau_{\text{casino}} V(A)}}{\sum_{j \in \text{Actions}} e^{\tau_{\text{casino}} V(j)}}$$

where $p(A)$ is the probability of choosing casino A, τ_{casino} is an inverse temperature parameter, and j enumerates all currently possible actions at this level of the hierarchy.

At the slot machine level of the hierarchy, we also compared two possible learning models:

The "Slot-Points Model" posits that separate values $V_{\text{slot}(i)}$ are maintained for each of the eight slot machines in the game (four in each casino). This enables fast and adaptive learning, especially in light of the fact that slot-machine-expected outcomes drift over time. According to this model, after a slot machine is played and its reward (the number of points obtained, r_{slot}) is observed, a prediction error $\delta_{\text{slot}} = r_{\text{slot}} - V_{\text{slot}(i)}$ is computed and the value of the slot machine is updated according to $V_{\text{slot}(i)}^{\text{new}} = V_{\text{slot}(i)}^{\text{old}} + \eta_{\text{slots}}\delta$, where η_{slots} is a learning rate or step-size parameter specific to learning at the slot machine level (potentially different from the learning rate for casinos). Because the order in which the

two slot machines are chosen within a casino is not relevant (i.e., it is equally optimal to choose the best and second-best slot machines in this order or in the opposite order), we modeled the value-based choice of slot machines as if participants chose a pair of slot machines i and j based on the sum of their values, so that $V_{pair(i,j)} = V_{slot(i)} + V_{slot(j)}$. Therefore, each trial involves six possible options for choosing two slot machines.

The alternative "Six-Armed Bandit Model" (which is suboptimal in our setting) posits that participants do not learn values for individual machines, but rather reinforce their slot machine choices directly based on the overall casino outcome. That is, according to this model participants simply tend to repeat actions that led to monetary gains in the past, as in Thorndike's "Law of Effect" (Thorndike, 1991). Formally, we assumed a value $V_{pair(i,j)}$ for each of six possible pairs of slot machines. After a choice of two slot machines in a casino, we computed a prediction error $\delta_{pair} = r_{pair} - V_{pair(i,j)}$ according to the outcome of the casino ($r_{pair} = 1$ or $r_{pair} = -1$) and updated the value of the pair of slot machines according to $V^{new}_{pair(i,j)} = V^{old}_{pair(i,j)} + \eta_{slots}\delta$.

To fit the parameters of both slot-machine-level models to behavioral choices, we again assumed a soft-max action selection function (see above), albeit with six possible (pair) choices on each trial. The inverse temperature parameters were allowed to differ from that of the casino choices.

For each of the 4 models (2×2 casino level and slot-machine-level learning models) we used each participant's behavioral data to fit the models' free parameters (η_{slots} and τ_{slots} for the slot machine level and η_{casino} and τ_{casino} for the casino-level choice). Model likelihoods were computed by assigning probabilities to each choice for each participant, according to the soft-max function specified above. In the case of the slot machines, there were 120 choice trials (each modeled as a choice of one of six possible pairs of slot machines). In the case of the casinos, likelihood was only estimated based on the 70 trials in which both casinos were "open" and thus choice behavior was available, although we modeled learning of casino values on all 120 trials. Note that each level of our task could be fit independently because, given the actual slot machine choices, there was no interaction between slot machine values and learning at the casino level.

We optimized model parameters by minimizing the negative log likelihood of the data given different parameter settings using MATLAB's fmincon function. This function performs constrained linear optimization over the space of possible parameter values. We constrained η_{slots} and η_{casino} to be between 0 and 1 and τ_{slots} and τ_{casino} to be positive. To facilitate finding the global minimum of the negative log likelihood, for each model and each participant we ran the routine four times from different, randomly chosen initial values for the two parameters in each fit and kept track of the best fit over the four runs. We compared the two alternative casino and slot-machine-level models by comparing the likelihoods of the models directly, because in both the casino- and slot-machine-level cases, the compared models have the same number of free parameters determined by the same number of data points, so no penalties for model complexity and potential overfitting needed to be established. For a general reference on our approach to model fitting, see Daw (2009).

Our models included one learning rate per hierarchy level, thereby assuming that participants learned equally from the outcomes of free-choice trials (when both casinos were open) as from those of forced trials (when one of the casinos was closed). To test the validity of this assumption, we also fit models with two distinct learning rates, one for forced and one for free choice trials. Models with two learning rates did not explain the behavioral data better than models with a single rate after accounting for the extra degree of freedom afforded by the extra parameters. Using both the Bayesian Information Criterion and the more lenient Akaike's Information Criterion, we found that models with a single learning rate were favored for all 28 participants. As a result, models with separate learning rates for forced and free-choice trials will not be discussed further.

Imaging data

Preprocessing. Preprocessing of the images and whole-brain image analysis were performed using SPM8 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK). Preprocessing of EPI images included motion correction (rigid-body realignment of all images to the first volume), and spatial normalization to a standard T2* template in Montreal Neurological Institute space. Anatomical regions of interest (ROIs) were marked for each participant using MRIcron (Center for Advanced Brain Imaging, Georgia State University and Georgia Tech University, Atlanta). Whole-brain images were then further preprocessed by spatially smoothing the images using a Gaussian kernel with a full width at half maximum of 8 mm to allow for statistical parametric mapping analysis.

Region of interest analysis. Based on the extensive existing literature on BOLD correlates of prediction error signals in the human brain, we focused our analysis on two a priori anatomically defined ROIs in the VS (McClure et al, 2003; O'Doherty et al., 2003; O'Doherty et al., 2004; Delgado et al., 2005; Abler et al., 2006; Li et al., 2006; Preuschoff et al., 2006; Hare et al., 2008; Glimcher, 2011; Niv et al., 2012) and VTA (D'Ardenne et al., 2008; Klein-Flügge et al., 2011).

The VS (nucleus accumbens) ROI was delineated separately for each participant using their structural brain image. The nucleus accumbens was anatomically defined as the area bordered superiorly by the internal capsule, caudate, and putamen; inferiorly by white matter or, in its most posterior extent, by the subcallosal gyrus; medially by the septal nuclei and/or the lateral ventricle; and laterally by the putamen. The border with the caudate was taken to be at the inferior margin of the lateral ventricle and with the putamen at the thinnest part of gray matter. We considered the anterior-most border to be at the axial slice in which the caudate and putamen were fully separated and the posterior border where the anterior commissure was fully attached between hemispheres. Only voxels wholly within these boundaries were considered part of the ROI. According to Klein-Flügge et al. (2011), the VTA ROI was defined as the anatomical region within Montreal Neurological Institute coordinates: x: -8 to $+6$; y: -26 to -14; and z: -20 to -12.

To analyze ROI time courses, we used the methods of Niv et al. (2012). We first averaged, for each participant and each ROI (VS and VTA, combining bilateral ROIs to one), the BOLD signal in all the ROI voxels using singular value decomposition. This resulted in a single time course per ROI per participant. We then removed from the time courses effects of no interest due to scanner drift and participant motion by estimating and subtracting from the data, for each session separately, a linear regression model that included the six motion regressors (3D translation and rotation), two trend regressors (linear and quadratic), and a baseline. To determine whether the resulting signal corresponded to prediction error signals, we regressed against each ROI time course a linear model that included three regressors of interest: FirstSlot, LastSlot, and Casino. These regressors were obtained from the model fits described previously. Although FirstSlot always included the prediction error corresponding to the outcome of the first slot machine chosen, LastSlot included the prediction error corresponding to the outcome of either the second slot machine chosen or the first slot machine in those trials in which one slot machine was sufficient to win the whole casino (this occurred in ~10% of the trials). In all cases, the prediction error regressors corresponded to the onsets of the relevant outcome events and the casino prediction errors corresponded to the onset of the LastSlot outcome. To conclude that BOLD activity corresponds to a prediction-error regressor, we required significant correlations at $p < 0.05$ across participants and that these correlations be positive.

To search for potential anatomical separation between activations for casino and slot machine prediction errors, we regressed against each voxel in the VS ROI, a linear model with two regressors from the best fitting models (see Results below): a "CombinedSlot" regressor that combined prediction errors for both slot machine plays from the Slot-Points Model, and a Casino regressor (with casino-level prediction errors from the Outcome Model). We defined the CombinedSlot regressor to increase power in our analysis of slot-machine-related activity. For each regressor, we then identified the peak activation voxel for each participant in each of the two bilateral ROIs. For each side of the brain, we computed a within-participant coordinate difference between the peak voxels for the Casino and CombinedSlot regressors, resulting in three values representing the distances along each coordinate: Δx, Δy, and Δz,

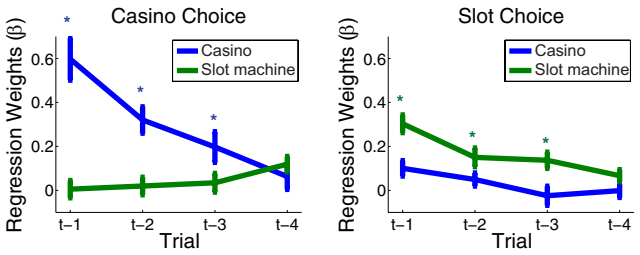

Figure 3. Logistic regression on casino (left) and slot machine (right) choices. We estimated the relationship between casino choices and the outcome of the casino on the last four times it was chosen, as well as the total slot machine points obtained in the corresponding trials. We similarly estimated the relationship between choices of each slot machine and the outcomes of the last four plays of this slot machine, as well as the casino outcomes during those same trials. Plotted are the regression weights for the last four outcomes of each type. Stars indicate significance at a between-participants Bonferroni-corrected level ($p < 0.0063$).

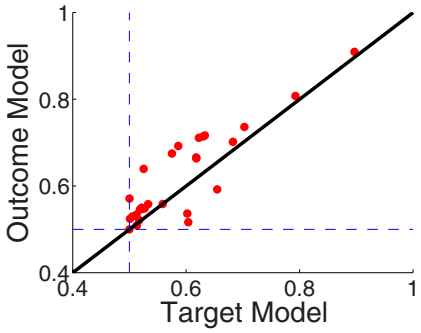

Figure 4. Average posterior probability per choice trial for the Outcome Model and the Target Model per participant. The Outcome Model assigns a higher average probability per trial to the choices of 22 of 28 participants (points lying above the solid equal-likelihood line). The average probability of a choice trial was calculated as the likelihood of the whole sequence of choice data divided by the number of choice trials. Dashed lines indicate chance.

and performed three *t* tests, one for each coordinate, to test for significant anatomical differences in any of the dimensions.

Whole-brain analysis. We used SPM8 to conduct a supplemental whole-brain analysis in which we searched for brain areas in which BOLD activity was correlated with the prediction error signals induced by our models. The design matrix comprised, for each of the four sessions: (1) three parametric regressors for slot machine prediction errors (FirstSlot and LastSlot) and casino prediction errors according to the Slot-Points Model and Outcome Model, respectively, similar to those used for the ROI analysis; (2) two stick-function regressors, one for the onsets of the casino door opening and one for the onsets of all slot machine outcomes; and (3) nuisance covariate regressors for motion, linear and quadratic drift, and baseline. The prediction error regressors were added as covariate regressors by convolving the punctate prediction errors as assigned by the model with the canonical hemodynamic response function. Other stick regressors were convolved with the hemodynamic response function, as is usual in SPM8. The six scan-to-scan motion parameters produced during preprocessing were used as nuisance motion regressors to account for residual effects of movement. This design matrix was entered into a regression analysis of the fMRI data of each participant. A linear contrast of regressor coefficients was then computed at the single-participant level for each regressor of interest. Each contrast was analyzed separately as a random effect at a second, between-participants level by including the contrast images of each participant in a one-way ANOVA with no mean term. Group-level activations were localized using a group-averaged structural scan and visualizations were generated using xjView (http://www.alivelearn.net/xjview8/).

Results

Behavioral results: participants learn at multiple levels of the hierarchy

We first tested in a model-free way, using modified logistic regression, whether participants learned at both levels of the task and whether they used the hierarchical structure of the task correctly in their learning; that is, we tested whether choices of casinos were informed by previous outcomes at the casino level, regardless of the specific slot machine outcomes during those plays (note that slot machine outcomes were, by design, uncorrelated from casino outcomes). Similarly, we tested whether slot machine choices were informed by previous slot machine outcomes and not the casino-level outcome on trials in which the slot machine had been chosen. The resulting regression weights revealed that casino choices were influenced by the outcome of the past three trials in which that casino was chosen (between participants, $p < 0.005$, Bonferroni corrected), but not by slot machine points obtained on those trials ($p > 0.05$; Fig. 3, left). Conversely, slot machine choices were influenced by the prior three slot machine outcomes (between participants, $p < 0.005$, Bonferroni corrected) but not by the casino outcomes on those trials ($p > 0.05$; Fig. 3, right).

Having established this hierarchical decomposition of learning in the task to two levels, we next used a model-based analysis to test for correct temporal-difference learning at both levels of the hierarchy. For this, we fit the choice data at each level to a learning model prescribed by the HRL framework and an alternative, simpler but suboptimal model (see Materials and Methods for details of all models).

At the casino level, we compared an Outcome Model that assumes that the participant updated the expected value of a casino based on the casino's true outcomes after playing the two slot machines and a Target Model that assumes that the participant updated the expected values of the casinos based only on the point target of the casino. Formal model comparison showed strong support for the Outcome Model, which provided a better fit for the choices of 22 of the 28 participants (Fig. 4; $p < 0.002$, one-tailed paired Student's *t* test on the difference in log-likelihoods of the Outcome Model and the Target Model). We thus used the Outcome Model in all our further fMRI analyses: for each participant, we generated regressors for expected casino-level prediction error activations by using that individual's best fit learning rate and inverse temperature parameters (mean across participants, 0.31 and 1.69, respectively; median, 0.19 and 0.3, respectively).

At the slot machine level, we compared a Slot-Points Model, which assumes that participants chose which two slot machines to play based on the sum of their expected outcomes and updated the expected value of each slot machine after observing its outcome according to a slot-level prediction error, and a Six-Armed Bandit Model, which treated each of the six possible pairs of slot machines as a distinct option (an "arm" in a bandit problem), with the value of the option updated at the end of the trial based on the overall casino outcome (win or lose). The first model represents a standard TD learning model, but accounts for the fact that the order of the two chosen slot machines is inconsequential. The second model also uses TD learning, but reinforces compound actions directly without learning about each of the slot machine outcomes. Again, formal model comparison favored the optimal Slot-Points Model: this model provided a better fit for the choices of 25 of the 28 participants, with the remaining three participants equally well fit by the two models (Fig. 5; $p < 10^{-6}$, one-tailed paired Student's *t* test on the difference in log-likelihoods of the Slot-Points Model and the Six-

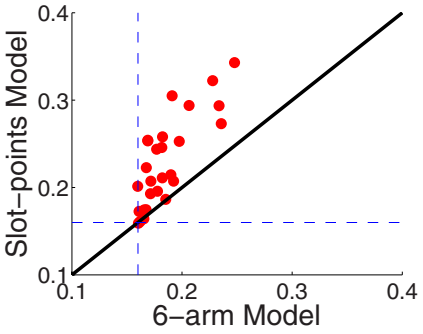

Figure 5. Average posterior probability per choice trial for the Slot-Points Model and Six-Armed-Bandit Model per participant. The Slot-Points Model assigns a higher average probability per trial to the choices of 25 of 28 participants (points lying above the solid equal-likelihood line). Chance = 0.16, indicated by dashed lines.

Armed-Bandit Model). Once again, the resulting parameter fits from the Slot-Points Model were used to generate slot-level prediction error regressors that were tailored to each individual's learning parameters for further fMRI analyses. The mean learning rates and inverse temperatures across participants were 0.37 (median 0.4) and 12.25 (median 0.98), respectively.

These results suggest that participants were simultaneously learning about slot machines and casinos based on separate prediction errors at each level of the hierarchy. Therefore, we should expect concurrent, distinct prediction errors, at least at the time of the last slot machine play, at which point information about that slot machine *and* about the overall worth of the casino became available simultaneously.

fMRI results: two different prediction error signals in VS

Supported by the behavioral results described above, we used the framework of HRL to model the participants' learning (see Materials and Methods for details) and to generate prediction error regressors for analysis of the fMRI data. At the lower, slot machine level, we modeled standard TD learning of separate values for each slot machine as in the Slot-Points Model, in which value estimates were updated when the outcome of a slot machine was encountered, according to the difference between the expected value of that slot machine and the actual number of points obtained. At the higher, casino level, a separate TD learning mechanism kept track of the value of each casino, and this value was updated when the casino outcome was revealed as per the Outcome Model.

This hierarchical model thus induced three regressors of interest in each trial for each participant: (1) a prediction error for the first slot machine played ("FirstSlot"), (2) a prediction error for the second slot machine played ("LastSlot"), and (3) a prediction error for the chosen casino ("Casino"). We modeled each regressor at the onset of the outcome that led to that prediction error, with LastSlot and Casino occurring simultaneously. Through the design of the task, these regressors were nearly orthogonal (mean correlation coefficient, −0.029, std 0.07) allowing us to search for neural correlates of each despite their temporal cooccurrence.

All three regressors of interest were significantly correlated with VS BOLD activity (FirstSlot, $p = 0.004$; LastSlot, $p = 0.0269$; Casino, $p = 5.5 \times 10^{-5}$), indicating that indeed two distinct, but temporally coincident prediction error signals, LastSlot and Casino, can coexist in the VS. To verify the result from our behavioral model comparison, we also tested whether the VS signal correlated with a prediction error regressor based on target only (from the Target Model) and found no significant effect ($p = 0.22$). In the VTA, our analysis involved a much smaller and less well defined ROI and a noisy signal due to pulsatility (D'Ardenne et al., 2008). We therefore performed one-tailed t tests looking for positive correlations. We found that VTA BOLD activity was significantly correlated with the FirstSlot regressor ($p = 0.0336$; one-tailed) and a trend in the case of Casino ($p = 0.051$; one-tailed). We did not find significant correlation with LastSlot ($p = 0.19$).

Testing for anatomical separation between slot and casino activations

The presence of two simultaneous prediction error signals in the VS led us to investigate whether there is an anatomical separation within striatum between areas activated by the slot machine prediction errors and areas activated by casino prediction errors. Visual inspection of the relative magnitudes of activation within VS suggested that slot machine activations might be more medial and casino activations more lateral. However, we did not find significant separation along any coordinate (between-participant, two-tailed t tests on each of x-, y-, and z-coordinate differences between peak voxels on each side of the brain were all $p > 0.3$). This null result could imply that there is indeed no anatomical separation, and that a single neural population is involved in producing an additive prediction error signal, making credit assignment a difficult problem. Alternatively, prediction errors at the two levels may be intertwined (or separable only at a subvoxel resolution) or participants may allocate each level of the task to a separate anatomical location idiosyncratically such that the allocation is inconsistent across participants (e.g., the casino level may be more laterally represented for some and more medially for others and vice versa for the slot level).

Whole-brain analysis

Our previous analysis concentrated on a priori, anatomically defined ROIs in the VS and VTA. To supplement this, we conducted a whole-brain analysis searching for areas correlating with two regressors of interest (CombinedSlot and Casino). At a whole-brain corrected threshold of $p < 0.05$, the only significant positive correlation found was between the Casino regressor and bilateral VS (ventral putamen; Fig. 6*A*). We also observed large clusters negatively correlated with the Casino regressor in visual areas (Table 1). Although the origin of this correlation is not clear, we speculate that losing in a casino caused an increase in attentive visual processing (when participants lost, the achieved points were denoted on the points bar by the color yellow and the number of remaining points was indicated in red; in contrast, if they won the casino, the entire points bar turned green), leading to correlation between negative prediction errors and visual cortical activation. A combined slot machine prediction–error regressor did not reveal any activation that survived whole-brain correction. However, bilateral ventral striatal activations did survive the widely used uncorrected $p < 0.001$ threshold (O'Doherty et al., 2004; Li and Daw, 2011), as expected (Fig. 6*B*).

The relatively weaker activations for the CombinedSlot regressor compared with the Casino regressor may have been due to the fact that outcomes at the casino level were binary, whereas slot machine outcomes were distributed normally. As a result, the variance of the Casino regressor was an order of magnitude larger than that of the slot machine regressors. This speculation is consistent with previous work showing that prediction errors due to binary outcomes are easier to detect in ventral stri-

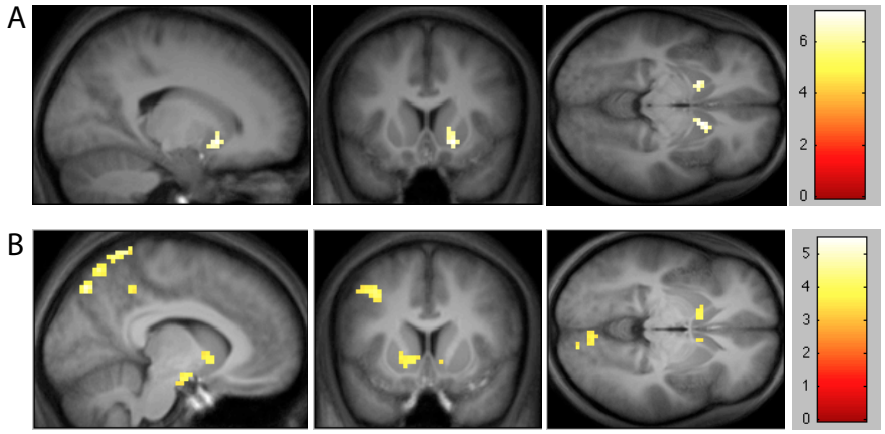

Figure 6. ***A***, Activations that survived a whole-brain FWE-corrected threshold of $p < 0.05$, cluster size >5, in the random effects contrast for the Casino regressor. Images are centered at voxel (18,14,−8) to better depict the extent of the activation. ***B***, Activations that survived an uncorrected threshold of $p < 0.001$, cluster size >5, in the random effects contrast for the combined Slot regressor. Images are centered at voxel (−9, 11, −5).

Table 1. All activations that survived a whole-brain FWE-corrected threshold of $p < 0.05$ in the random-effects contrast for a casino prediction error signal

Anatomical location	Peak *x,y,z* (mm)	Cluster size	Peak intensity (T)
Right ventral putamen	15, 11, −8	29	7.15
Left ventral putamen	−18, 8, −11	17	7.05
Left lingual gyrus	−3, −64, 1	31	−6.91
Right lingual gyrus	9, −73, −5	7	−6.66
Occipital lobe	−12, −88, 25	485	−12.68

Anatomical locations were determined through inspection with respect to the average anatomical image of all 28 participants.

atal BOLD signals than those derived from normally distributed outcomes (compare Fig. 2*A* in Schönberg et al., 2007 with the activations in supplemental Table 3 in Daw et al., 2006). We verified this hypothesis in a companion experiment (Diuk et al., 2010) in which slot machine outcomes were binary (win/lose) and the casino outcome was multiple valued (specifically, participants could play up to four slot machines each costing 5¢ to play, and had to win at least two slot machines to earn a certain amount of money the casino offered, which varied in each trial according to a casino-specific distribution). In this companion experiment, whole-brain analysis (FWE corrected at the $p < 0.05$ level) revealed significant bilateral VS activation for the slot machine prediction-error regressors, but a trend was only observed for the casino regressor ($p < 0.001$, uncorrected). Because prediction errors at level two of the hierarchy were not orthogonal in that study by design, we have not reported its results here.

Another factor that might explain the difference in activation strengths between casino and slot machine prediction errors is that slot machine outcomes did not lead to direct monetary reward, but rather acted as subgoals toward a reward that was only obtained if the casino was won. Therefore, slot machine outcomes and prediction errors might have been less salient. However, this was also the case in the companion experiment in which slot machine prediction errors were stronger than casino prediction errors despite the fact that the latter were the only ones related to actual monetary outcomes.

Finally, correlations between prediction errors and outcomes may have affected the relative strengths of our activations. These correlations exist because prediction errors reflect the difference between outcome and value and, by definition, the values are correlated with previous outcomes due to learning. As a result, dissociating outcomes from prediction errors is difficult in our task, as was the case in many tasks used previously (e.g., most variations of bandit tasks; but see Hare et al., 2008 for a notable exception in which the task was specifically designed to dissociate these signals). Nevertheless, we do not believe that these correlations could explain the large difference in activations, because relatively strong correlations occurred at both levels of the hierarchy and in both the current task and the companion task described above.

Discussion

We tested for the existence of simultaneous, orthogonal prediction errors supporting learning on different levels of a task hierarchy in the VS and VTA. To generate simultaneous prediction errors, we made use of hierarchy and predictions from the computational framework of HRL (Sutton et al., 1999; Dietterich, 2000; Barto and Mahadevan, 2003). Despite differences between existing HRL models, common to most of them is the existence of multiple prediction errors that occur when a subtask ends and new knowledge about multiple levels of the hierarchy becomes available. Recent work (Botvinick et al., 2009) derived a set of predictions from the HRL framework, evaluating the extent to which current scientific knowledge accorded with each of the theory's elements. Only recently, experimental data have been produced that support some of these predictions directly (Ribas-Fernandes et al., 2011; Botvinick, 2012). Our results add to this body of work, further demonstrating the plausibility of HRL in the human brain.

Specifically, we investigated whether, when confronted with a hierarchical task that requires learning in parallel about two levels of task structure, the human brain is capable of generating two concurrent but distinct reward prediction error signals. Our behavioral results showed that participants learned the task successfully at both levels of the hierarchy. Moreover, participants' choice behavior at each level of the task (casino choice or slot machine choice) showed dependence on the outcomes of that level only and independence of the outcomes of the alternative level, thereby confirming that the hierarchical structure of the task was used and encapsulated correctly.

In our task, the high-level casino outcome was revealed at the exact same time as an outcome of a lower-level task (second slot machine play). Learning about these two events thus required the presence of two separable but simultaneous reward prediction errors. Using fMRI and anatomically defined ROIs, we found that BOLD signals in the VS indeed were correlated with each of these prediction error signals, with a trend in the case of VTA. These results provide evidence for a key prediction of HRL, namely the existence of coincident reward prediction errors corresponding to the different levels of the hierarchy and the existence of high-level prediction errors that span lower-level actions and transition.

Our work adds to a growing body of evidence indicating the existence of more than one prediction error signal in the brain. Neuroimaging data have suggested the existence of functionally distinct error signals. Gläscher et al. (2010) found evidence for a distinct reward and a state prediction error signal in different brain areas. Two recent studies in the social domain found dis-

tinct error signals in prefrontal areas for predictions about the observed actions of others and the outcomes of such actions (Burke et al., 2010; Suzuki et al., 2012). Other work has found evidence of both a standard and a "fictive" reward prediction error in VS (Lohrenz et al., 2007, but see Li and Daw, 2011). Another result shows dual prediction errors for reward and social reputation in striatum and prefrontal cortex (Behrens et al., 2008). Finally, a recent study (Daw et al., 2011) indicated the presence of a striatal prediction error signal based on combined predictions from two learning systems (model-based and model-free). In that work, however, the two signals were not orthogonal and only their combined effect could be observed. Physiological findings have also suggested that the population of dopaminergic neurons may not be homogenous (Brischoux et al., 2009). Our results extend beyond these previous studies by showing that reward prediction error signals projecting into the same brain area (VS and, slightly less compellingly, the VTA) can signal more than one quantity at the same time. Our results are most similar to those by Gershman et al. (2009), who showed the presence of prediction errors from two simultaneously performed tasks in the VS. That study concentrated on effector-specific decomposition of prediction errors rather than their superposition and showed that activity was strongest for prediction errors from the task performed by the contralateral hand. We add to these results by showing that prediction errors are not only represented simultaneously, but can also span different temporal resolutions, allowing hierarchical decomposition of learning and decision making.

We provide functional evidence against the unitary nature of prediction errors and suggest that, in more realistic learning scenarios, several prediction errors may be used to learn in parallel at different levels of the task. This finding contrasts with previous empirical work suggesting that prediction errors are scalar and unitary, resulting from dopamine neurons signaling a single, global difference between obtained and expected reward (Schultz et al., 1997; Glimcher, 2011). However, it does not contradict those previous data: in the simple tasks examined previously, only one prediction error signal was available and required for learning at each point in time. Our finding of two concurrent prediction error signals suggests that these prior results are a special case of the function of prediction errors in the RL machinery of the basal ganglia, in which the VS and VTA are key players, and that a more detailed parcellation may be uncovered by using more complex tasks.

Indeed, our present results and previous results suggest that the so-called "scalar prediction error signal" may be more of a vector-valued signal, as required by a number of RL extensions such as learning of successor representations (Dayan, 1993; Hayes et al., 2011), factored representations (Koller and Parr, 1999), and HRL (Barto and Mahadevan, 2003; Botvinick et al., 2009). However, this raises the problem of spatial credit assignment: how are the different prediction error signals distinguished in downstream areas to learn separate reward predictions? Does this rely on a "hard-wired" anatomical separation of predictions for different levels of task hierarchy (which implies a limit on the number of nested levels that can be learned about at each given time), or is the decomposition of prediction errors more flexible? That we did not find a consistent anatomical separation of the two prediction errors across participants may support the latter option, but given the null effect nature of our results, more research on this question is clearly warranted.

In conclusion, our results have two key implications: the first is that more than one concurrent prediction error signal may be calculated and used for learning in the brain. This may not be surprising from a theoretical point of view, because learning about two (or more) separate reward predictions within any given scenario requires the calculation of two separate prediction errors. Such a dual-task situation may be common in daily life. However, RL tasks examined previously in laboratory settings did not test this prediction directly. The second implication is that the human brain can calculate prediction errors that temporally span over several states and actions, a fundamental element in existing HRL models (Botvinick et al., 2009). The fact that a casino-level prediction error was apparent at the end of the casino play suggests that the predicted value of the casino was maintained in memory throughout the casino play (which included two lower-level slot machine plays) to be compared with the actual outcome of the casino (see Materials and Methods and Fig. 2). Our task cannot determine whether the value, or rather the state (which casino was chosen), was maintained in memory, because these are equivalent from the point of view of prediction errors. Moreover, our task design purposefully made it easy to remember the current high-level state throughout the trial because we were interested in uncovering these temporally extended prediction errors. However, that a prediction error was computed abstracting over intervening actions and state transitions is of major interest to understanding how high-level decision making is accomplished in the brain.

References

Abler B, Walter H, Erk S, Kammerer H, Spitzer M (2006) Prediction error as a linear function of reward probability is coded in human nucleus accumbens. Neuroimage 31:790–795. CrossRef Medline

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.

Barto AG, Mahadevan S (2003) Recent advances in hierarchical reinforcement learning. Discrete Event Dynamic Systems 13:341–379. CrossRef

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

Botvinick MM (2012) Hierarchical reinforcement learning and decision making. Curr Opin Neurobiol 22:956–962. CrossRef Medline

Botvinick MM, Niv Y, Barto AC (2009) Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition 113:262–280. CrossRef Medline

Brainard DH (1997) The psychophysics toolbox. Spat Vis 10:443–446. CrossRef Medline

Brischoux F, Chakraborty S, Brierley DI, Ungless MA (2009) Phasic excitation of dopamine neurons in ventral VTA by noxious stimuli. Proc Natl Acad Sci U S A 106:4894–4899. CrossRef Medline

Burke CJ, Tobler PN, Baddeley M, Schultz W (2010) Neural mechanisms of observational learning. Proc Natl Acad Sci U S A 107:14431–14436. CrossRef Medline

Cohen JD, McClure SM, Yu AJ (2007) Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. Philos Trans R Soc Lond B Biol Sci 362:933–942. CrossRef Medline

D'Ardenne K, McClure SM, Nystrom LE, Cohen JD (2008) BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. Science 319:1264–1267. CrossRef Medline

Daw ND (2009) Trial-by-trial data analysis using computational models. Attention and Performance (1–26).

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876–879. CrossRef Medline

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. CrossRef Medline

Dayan P (1993) Improving generalization for temporal difference learning: the successor representation. Neural Computation 5:613–624. CrossRef

Delgado MR, Miller MM, Inati S, Phelps EA (2005) An fMRI study of reward-related probability learning. Neuroimage 24: 862–73. CrossRef Medline

Dietterich TG (2000) Hierarchical reinforcement learning with the MAXQ value function decomposition. Journal of Artificial Intelligence Research 13:227–303.

Diuk C, Botvinick MM, Barto AG, Niv Y (2010) Program No. 36:907.13. 2010 Neuroscience Meeting Planner, San Diego, CA: Society for Neuroscience.

Gershman SJ, Pesaran B, Daw ND (2009) Human reinforcement learning subdivides structured action spaces by learning effector-specific values. J Neurosci 29:13524–13531. CrossRef Medline

Gläscher J, Daw N, Dayan P, O'Doherty JP (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron 66:585–595. CrossRef Medline

Glimcher PW (2011) Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. Proc Natl Acad Sci U S A 108.

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. CrossRef Medline

Hayes TR, Petrov AA, Sederberg PB (2011) A novel method for analyzing sequential eye movements reveals strategic influence on Raven's advanced progressive matrices. J Vis 11:10. CrossRef Medline

Klein-Flügge MC, Hunt LT, Bach DR, Dolan RJ, Behrens TE (2011) Dissociable reward and timing signals in human midbrain and ventral striatum. Neuron 72:654–664. CrossRef Medline

Koller D, Parr R (1999) Computing factored value functions for policies in structured MDPs. In: Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (Dean T, ed), pp 1332–1339. San Francisco: Morgan Kaufman.

Li J, McClure SM, King-Casas B, Montague PR (2006) Policy adjustment in a dynamic economic game. PLoS ONE 1:e103. CrossRef Medline

Li J, Daw ND (2011) Signals in human striatum are appropriate for policy update rather than value prediction. J Neurosci 31:5504–5511. CrossRef Medline

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. CrossRef Medline

McClure SM, Berns GS, Montague PR (2003) Temporal prediction errors in a passive learning task activate human striatum. Neuron 38:339–346. CrossRef Medline

Montague PR, Dayan P, Sejnowski TJ (1996) A Framework for Mesencephalic Predictive Hebbian Learning. J Neurosci 16:1936–1947. Medline

Niv Y, Edlund JA, Dayan P, O'Doherty JP (2012) Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. J Neurosci 32:551–562. CrossRef Medline

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. CrossRef Medline

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. CrossRef Medline

Preuschoff K, Bossaerts P, Quartz SR (2006) Neural differentiation of expected reward and risk in human subcortical structures. Neuron.

Ribas-Fernandes JJ, Solway A, Diuk C, McGuire JT, Barto AG, Niv Y, Botvinick MM (2011) A neural signature of hierarchical reinforcement learning. Neuron 71:370–379. CrossRef Medline

Schönberg 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. CrossRef Medline

Schultz W (2002) Getting formal with dopamine and reward. Neuron 36: 241–263. CrossRef Medline

Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599. CrossRef Medline

Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Cambridge, MA: MIT.

Sutton RS, Precup D, Singh S (1999) Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. Artificial Intelligence 112:181–211. CrossRef

Suzuki S, Harasawa N, Ueno K, Gardner JL, Ichinohe N, Haruno M, Cheng K, Nakahara H (2012) Learning to simulate others' decisions. Neuron 74: 1125–1137. CrossRef Medline

Thorndike EL (1991) Animal intelligence: experimental studies. New York, NY: Macmillan.