Behavioral/Systems/Cognitive

Reinforcement Learning Signals in the Human Striatum Distinguish Learners from Nonlearners during Reward-Based Decision Making

Tom Schönberg,[1] Nathaniel D. Daw,[2] Daphna Joel,[1] and John P. O'Doherty[3]

[1]Department of Psychology, Tel Aviv University, Tel Aviv 69978, Israel, [2]Center for Neural Science and Department of Psychology, New York University, New York, New York 10003, and [3]Division of Humanities and Social Sciences and Computation and Neural Systems Program, Caltech, Pasadena, California 91125

The computational framework of reinforcement learning has been used to forward our understanding of the neural mechanisms underlying reward learning and decision-making behavior. It is known that humans vary widely in their performance in decision-making tasks. Here, we used a simple four-armed bandit task in which subjects are almost evenly split into two groups on the basis of their performance: those who do learn to favor choice of the optimal action and those who do not. Using models of reinforcement learning we sought to determine the neural basis of these intrinsic differences in performance by scanning both groups with functional magnetic resonance imaging. We scanned 29 subjects while they performed the reward-based decision-making task. Our results suggest that these two groups differ markedly in the degree to which reinforcement learning signals in the striatum are engaged during task performance. While the learners showed robust prediction error signals in both the ventral and dorsal striatum during learning, the nonlearner group showed a marked absence of such signals. Moreover, the magnitude of prediction error signals in a region of dorsal striatum correlated significantly with a measure of behavioral performance across all subjects. These findings support a crucial role of prediction error signals, likely originating from dopaminergic midbrain neurons, in enabling learning of action selection preferences on the basis of obtained rewards. Thus, spontaneously observed individual differences in decision making performance demonstrate the suggested dependence of this type of learning on the functional integrity of the dopaminergic striatal system in humans.

Key words: **prediction errors; computational models; fMRI; basal ganglia; instrumental conditioning; associative learning**

Introduction

An accumulating body of research implicates a network of brain regions in reward learning and decision making (Schultz, 2002; O'Doherty, 2004). Single-unit neurophysiology studies implicate dopamine neurons in encoding reward prediction errors (PEs) during classical and instrumental conditioning, consistent with reinforcement learning (RL) models of decision making (Montague et al., 1996; Schultz et al., 1997; Schultz, 2002; Morris et al., 2006). In RL, prediction errors are used to update expectations of future reward associated with a set of stimuli or actions, which are then subsequently used to guide action selection (Sutton and Barto, 1998; Dayan and Abbot, 2001; Dayan and Balleine, 2002; Daw and Doya, 2006). Human neuroimaging studies have revealed activity in target areas of dopamine neurons most prominently in the ventral and dorsal striatum, consistent with RL PE signals (Delgado et al., 2000; Knutson et al., 2000; Pagnoni et al., 2002; O'Doherty et al., 2003, 2004; McClure et al., 2003, 2004; Rodriguez et al., 2006). Error signals in ventral striatum have been found during both classical and instrumental conditioning, indicative of a role for this structure in mediating learning of expected reward in general, whereas dorsal striatum has been found to be more strongly engaged during instrumental conditioning tasks (O'Doherty et al., 2004; Tricomi et al., 2004), implicating this area in learning related to action selection for reward (O'Doherty et al., 2004; Seger and Cincotta, 2005).

A functional interpretation of RL signals in the striatum implies that these signals should relate to behavioral performance on such tasks, a relationship about which much less is known. Such predictions help to distinguish the RL hypothesis from interpretations of striatal activity that are less committed to a specific behavioral function, such as those stressing a relationship between neural signals and stimulus salience (Zink et al., 2006), echoing similar ideas about dopaminergic spiking in animals (Redgrave et al., 1999). Pessiglione et al. (2006) showed previously that PE activity in striatum was modulated by administration of dopamine agonists and antagonists, and that such modulation exerted a corresponding influence on behavioral performance. In this study, we aim to exploit spontaneously oc-

Received June 1, 2007; revised Oct. 5, 2007; accepted Oct. 6, 2007.

This work was supported by Adams Super Center Grant 6194-1 for brain studies in Tel Aviv University (D.J.), a grant from the Chief Scientist Office of the Ministry of Health, Israel (D.J.), and by a grant from the United States–Israel Binational Science Foundation, Jerusalem, Israel (D.J., J.P.O.). T.S. is supported by The Functional Human Brain Mapping Unit, a Joint Project of Levi–Edersheim–Gitter Institute, Tel Aviv University, and Tel Aviv Sourasky Medical Center. J.P.O. is also supported by a Searle scholarship and by a grant from the Gordon and Betty Moore foundation.

Correspondence should be addressed to Tom Schönberg, Department of Psychology, Tel Aviv University, Tel Aviv 69978, Israel. E-mail: tom@post.tau.ac.il.

DOI:10.1523/JNEUROSCI.2496-07.2007

0270-6474/07/2712860-08$15.00/0

curring differences in decision making abilities (Stanovich and West, 2000; Stanovich, 2003; Newell, 2005) of untreated healthy subjects to further study the relations between RL and decision making in humans. For this purpose, we used a simple four-armed bandit decision making task in which subjects choose between one of four options, each with a different, fixed reward probability ranging from 0.75 down to 0.25 (Friedland, 1998). To perform optimally, subjects need to learn to choose the actions associated with the highest probability of reward. Yet, remarkably even on such a simple task, ~50% of subjects fail to choose optimally even after >100 trials (Joel et al., 2005). The aim of the present study is to use functional neuroimaging to study the relationship between these spontaneous differences in behavioral performance and neural signals. We hypothesized that better behavioral performance would accompany more robust PE activity in either the dorsal or ventral striatum.

Materials and Methods

Subjects. Thirty right-handed healthy normal subjects (one subject discarded because of a clinical finding) participated in the experiment (mean age, 27.4; range, 22–39) of which 15 were female. The subjects were preassessed to exclude those with a prior history of neurological or psychiatric illness. All subjects gave informed consent, and the study was approved by the Ethics committee of the Tel Aviv Sourasky Medical Center and by the Ethics committee in the Department of Psychology in Tel Aviv University.

Imaging procedure. A GE 3.0T Excite scanner (General Electric, Milwaukee, WI) was used to acquire gradient echo T2*-weighted echo-planar images (EPI) with blood oxygenation level-dependent (BOLD) contrast. Each volume comprised 40 axial slices of 3.0 mm thickness and 3.125 in-plane resolution. All images were acquired using a standard quadrature head coil. The following parameters were used: repetition time (TR), 3000 ms; echo time (TE), 30 ms; flip angle, 90°; and a 64×64 matrix with a field of view of 20×20 cm^2. A total of 968 volumes were scanned in four separate scanning runs composed of two task sessions and two control sessions as described below. For each subject, a T1-weighted image was also acquired.

Task. The task is an adaptation of Friedland's card betting task for event-related functional magnetic resonance imaging (fMRI) (Friedland, 1998). In this task, subjects are invited to choose on each trial a card from one of four stationary decks. Each deck contains two types of cards: winning cards and losing cards. When subjects select a winning card they obtain 100 points (which are converted after the experiment into the monetary equivalent of 2 United States cents). Selection of a losing card results in no points. At the beginning of each trial, subjects "bet" a fixed amount of 50 points, thus selection of a winning card results in an overall gain of 50 points, whereas selection of the losing card results in an overall loss of 50 points. The subjects are invited to choose from the decks in order to select as many winning cards as possible and hence win as many points as possible. However, the subjects are not aware that the decks differ in the probability of obtaining a winning card. One deck yields winning cards on 75% of occasions, another deck on 60%, and the two other decks yield winning cards on 40 and 25% occasions, respectively. Thus, the optimal strategy for the subjects is to choose the 75% deck once they have worked out the contingencies. As in the original Friedland's card-betting task used by Joel et al. (2005), all the decks are built in a pseudorandomized manner to ensure that the actual experienced probability of winning on a particular deck does not deviate within every 10 sequential choices, by >10% from the designated probability of winning for that deck. This ensures that no more than four cards of the same type (win/lose) appear one after the other in the 75 and 25% decks, and no more than three cards of the same type in the 60 and 40% decks. Each deck is randomly allocated to one of four positions on the screen at the beginning of the task and remains stationary for the entire task. The colors of the cards are pseudorandomized and counter-balanced across subjects.

Each trial begins with the cards in the four decks facing down and the subjects are prompted to choose a deck, thereby committing to a 50 point bet. The maximum time to choose a deck is 2 s. If a deck is chosen in this time, a green rectangle appears around the chosen deck to indicate that a choice has been made. Three seconds later the card on the top of the chosen deck is turned over to reveal the color of the card, and whether it was a winning card worth 100 points or a losing card with 0 points. The outcome is then presented for 1.5 s. The last part of the trial is a fixation cross, which remains on the screen until a total trial duration of 7 s is reached. On trials where subjects fail to respond in 2 s, a # sign is superimposed over all the decks for 2 s to signify a missed trial. This is then followed by a fixation cross for 3 s until the next trial is triggered. Subjects are not provided with a running total and are only informed of the total points accumulated at the end of the experiment. The trial structure is summarized in Figure 1*A*. In addition to the gambling trials, low-level null event trials are also included whereby a fixation cross is presented for the full 7 s.

Experimental procedure. The task was presented to subjects on a computer monitor using presentation (Neurobehavioral Systems, Albany, CA) projected onto a screen, visible via an angled mirror on top of the fMRI head coil.

Before the experiment began, subjects were informed that the probability of wining on the decks remains stationary throughout the experiment. They were instructed to try and win as many points (later converted to money) as they can with no specific instructions regarding the structure of the task. They were also told that in addition to being reimbursed according to the amount of points won directly on the task, the subject obtaining the highest points over all would obtain a grand prize equivalent to $20 US dollars (USD). Subjects' gained on average $2.5 USD (±$0.25 SD) by the end of the experiment.

In addition to the card betting task, a control task was also performed by the subjects. The control task is identical in length and structure to the card-betting task, except that there are no monetary rewards involved in the control task. The decks in the control task also contain similar percentages of two card colors (signified by different colors than the ones used in the card-betting task) as used in the betting task (75, 60, 40, and 25%). Subjects were explicitly instructed to perform the control task as a condition of their participation in the experiment, although they did not receive rewards during this part of the experiment.

The total time taken for the functional imaging was 48 min and 24 s (for the task and control sessions together). Each of the four parts (two of task and two of control) consisted of 75 trials of either the card-betting task or control task and 25 null event trials (242 volumes in each part including the first six volumes which were discarded in analysis, to allow magnetization stabilization). The order of presentation of the task and control sessions was completely counterbalanced across subjects. Before scanning, the subjects underwent a short practice session which included both task and control trials.

Postexperiment ratings

After subjects were removed from the scanner, they were asked to report pleasantness ratings for each of the decks, using a scale ranging from 1 to 7, where 1 equaled the least pleasant and 7 the most pleasant. The subjects were also asked to rank the decks in order of their preference where 4 equaled the best deck, and 1 the worst, as well as to provide an assessment of the assumed probability of winning on each deck, using a number from 0 to 100.

Psychological questionnaires and demographic data

After the postexperiment ratings, subjects filled in several questionnaires, all in Hebrew: Friedland's chance-luck questionnaire for identifying tendency toward chance or luck attribution to events (for a full description, see Joel et al., 2005); the obsessive-compulsive personality scale from the Wisconsin Personality Disorders Inventory (Klein et al., 1993) and the obsessive-compulsive inventory (Foa et al., 2002) to assess any possible relation between obsessive-compulsive personality traits or symptoms and task performance; the NEO-R questionnaire (Costa and McCrae, 1992a,b) to try and probe for correlations between the "big 5" personalities and behavior in the task; and the Beck depression inventory (BDI) (Beck et al., 1961; Beck, 1988) to test for correlations between performance in the task and depression.

Subjects were also asked to report their age, years of education and the Israeli equivalent of their SAT scores.

Group level analysis

Subjects were split into groups according to whether they reached the learning criterion or not. The criterion was showing a statistically significant preference for the high-probability (HP) decks (75 and 60%) in the last 40 trials of the task (i.e., choosing these decks on >25 trials according to the binomial distribution around $p = 0.5$). The contrast images computed for each subject were taken to the group random effects level and comparisons were conducted between the learner group and the nonlearner group to determine areas showing enhanced correlations with PE signals in learners compared with nonlearners. The structural T1 images of all subjects were normalized to a standard template. The normalized images were then used to create a normalized structural mean image on which the *t* maps were over laid to obtain anatomical localization.

Reinforcement model-based analysis

Subjects' decisions were modeled as a function of previous choices and rewards using a temporal-difference algorithm. Specifically, the predicted value V_i for each option (deck) i was updated in the direction of the obtained reward using a delta rule with learning rate a whenever deck i was chosen. To capture low-order autocorrelation in the choices (Lau and Glimcher, 2005), we also maintained for each deck i, an index c_i tracking how recently it had been chosen, and allowed this to bias choices. Specifically, when deck i was chosen, c_i was set to one, whereas c_j for all other decks $j \neq i$ were decayed exponentially (by multiplication with a decay factor d). The probability of choosing option i was taken to be softmax in a weighted sum of the value and its choice fraction [i.e., proportional to $\exp(\beta \times (V_i + b \times c_i))$]; note that the coefficient b can be negative to capture a tendency to alternate or positive for perseveration. Free parameters (learning rate a, softmax inverse temperature β, weighting b for the choice recency index, and the time constant d for decay of the choice recency index) were selected to optimize the likelihood of the behavioral data. For fMRI analyses, the behavior of all members of the learner group was fit with a single set of parameters; a second single set of parameters was fit to the behavior of all members of the nonlearner group. A more detailed description of the model fitting procedure as well as additional fits of the model to individual subjects' behavior are provided in the supplemental material. The fit model was then run on each subject's choices and rewards to estimate their trial-by-trial reward predictions V_i and the PE in these values, which were used as parametric regressors for the fMRI analysis as described below. The parametric regressors were modeled by convolving outputs from the RL model with a vector containing impulse events. For the PE regressor we modulated the impulse events at two time points within the trial: the first was the mid time point between stimulus onset and response time and the second time point was at reward delivery. At the first time point, the prediction error was defined as the temporal difference between the value expected given the choice, V_c, defined as above, and the initial value that would be expected before knowing the choice. (We defined the latter as the policy averaged value, i.e., using a separate variable V_{avg} updated according to a delta rule with the same learning rate as the choice values, but on each trial regardless of which option was chosen) (Daw et al., 2006b) (see also O'Doherty et al., 2004). The prediction error at the time of the outcome was then the difference between the observed reward and the reward expected given the choice.

Image analysis

Analysis of fMRI data were performed in SPM2 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK). To correct for subject motion, images were realigned to the first volume, then spatially normalized to a standard T2* template with a resampled voxel size of 3 × 3 mm. Images were then spatially smoothed by applying a Gaussian kernel with a full width at half maximum of 8 mm. High-pass filtering with a cutoff period of 128 s was also applied to the data.

Prediction error signals were generated for each subject as described above and then convolved with a canonical hemodynamic response function and regressed against each subjects' fMRI data. Subjects' gains were entered in the design matrix. The six scan-to-scan motion parameters produced during realignment were included to account for residual effects of scan-to-scan motion. Contrast images were computed at the single subject level by subtracting PE signals during the control task from the PE signals during the card betting task.

Time-course analysis

Time courses were collected at the single subject level. This was done by searching for the peak activation in the contrast of PE signals in the task subtracted by PE signals in the control session of each subject. A sphere of 8 mm was defined around the location of the group peak voxel in the random effects analysis which showed the difference between learners and nonlearners ([18, 15, 15], Montreal Neurological Institute (MNI) coordinates). For each individual subject the locus of the peak activation was checked to ensure that the peak chosen within this sphere fell within the anatomical boundaries of the striatum. This criterion was met for all subjects and hence all were included in this analysis. Then a volume of 8 mm was defined around the specific single subject activation found in the process defined above to collect the raw signal from the peak in that sphere. Effects of no interest were removed except for PE in the task and control sessions of each subject. Then trials which had high positive PE according to the model (>0.4) were binned from each subject's time course. The same procedure was applied to bin trials with negative PE (smaller than −0.4). The data were then smoothed from 3 to 1 s using linear interpolation in Matlab (MathWorks, Natick, MA). All the trials of each type were averaged, first across all trials within each subject and then further averaged across subjects in each of the groups of learners and nonlearners separately to produce group level time courses.

Results

Behavioral results

Learners versus nonlearners

Fifty-nine percent of subjects (17 subjects) were classified as learners and 41% (12 subjects) were classified as nonlearners based on our criterion (Fig. 1*B*), a distribution similar to that reported in previous behavioral studies using this task (Joel et al., 2005).

Demographic data of the two groups classified according to the learning criterion are shown in Table 1. As can be seen, the only significant difference between the two groups was in task performance. No differences were seen in age, years of education or in Israeli SAT scores.

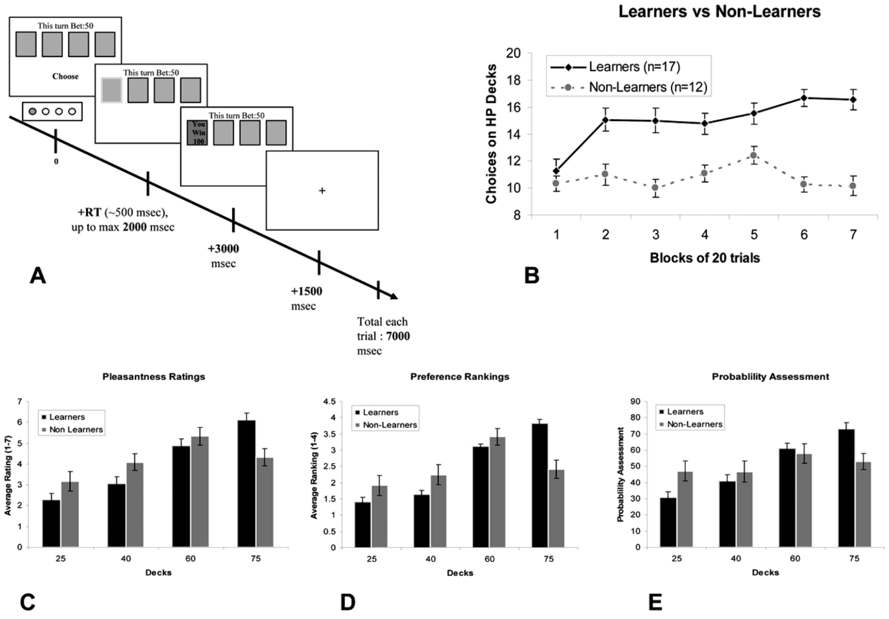

Figure 1. *A*, General outline of a trial in the card-betting task. The task contained four decks of cards. Each deck had a predefined probability of winning of either 75, 60, 40, or 25%. On each trial, subjects had to choose one of the four decks. Participants were unaware of the probability assigned to each deck. ***B***, Subjects' performance during fMRI scanning. Separation into two groups was based on subjects' choices on the two HP decks (75 and 60%) in the last 40 trials of the task. ***C–E***, Postexperiment ratings show significant interaction between groups (learners and nonlearners) in pleasantness ratings (***C***) and preference rankings (***D***). ***E***, In the probability assessment question, a significant linear trend is seen in the learners group but not in the nonlearners group.

Table 1. Demographic and behavioral data of subjects in both groups

	Average (SD) nonlearners	Average (SD) learners
Gender (male/total)	5/12	9/17
Age (years)	27.8 (5.2)	26.6 (3.4)
Education (years)	14.8 (1.7)	15.2 (1.9)
Number of choices of HP decks on last 40 trials of task* (learning criterion)	20.4 (3.4)	33.2 (5.5)
Israeli SAT score	678.1 (66.4)	696.4 (56.1)
Completed trials task (of 150)	148.9 (1.2)	148.52 (2.1)
Completed trials control (of 150)	147.8 (3.3)	148 (3.1)
RT task session (ms)	525 (80)	535 (135)
RT control session (ms)	521 (85)	573 (118)

No significant differences were found between the two groups except for task performance (number of choices on the 60 and 75% decks in the last 40 trials of the task (*$p < 0.000001$). All other independent *t* test comparisons were not significant ($p > 0.2$ to $p > 0.8$). In all of the above shown comparisons, $n = 17$ for learners and $n = 12$ for nonlearners, except for SAT scores where $n = 15$ for learners and 10 for nonlearners.

Pleasantness ratings

We found a significant difference between the learner and nonlearner groups in postexperiment ratings of pleasantness for the different decks, as shown by a significant interaction between deck (25, 40, 60, or 75%) and group (learner vs nonlearner) in a repeated measures ANOVA ($F_{(3,81)} = 5.884$; $p < 0.002$). A Tukey's *post hoc* test revealed that this interaction effect is driven by a significantly decreased pleasantness rating for the 75% deck in the nonlearners compared with the learners (at $p < 0.02$) (Fig. 1*C*).

Preference rankings

A similar effect was also found for preference rankings with a significant deck by group interaction ($F_{(2,54)} = 12.43$; $p < 0.0001$) also being driven by a significantly decreased preference ranking for the 75% deck in the nonlearners compared with the learners (at $p < 0.0001$) (Fig. 1*D*). Because of the dependency between rankings (forcing allocation of 1-2-3-4 to the decks) we performed this ANOVA and subsequent Scheffe *post hoc* analysis only on the 40, 60, and 75% decks.

Probability estimation

We also found a marked difference between the groups in subjects' estimation of the probability of winning on the decks (Fig. 1*E*). The learner group showed average probability estimates that are remarkably close to the actual underlying probabilities of winning on each of the decks (with estimated probabilities of 30, 41, 61, and 73 for the 25, 40, 60, and 75% decks, respectively). It should be noted that at no point were subjects given explicit information about the actual probabilities of winning on each of the decks. These ratings showed a significant linearly increasing trend in probabilities assigned to the decks ($F_{(1,27)} = 43.43$; $p < 0.00001$). However, in the nonlearner group the average probability estimates deviated markedly from the actual probabilities, with estimated probabilities of 47, 47, 58, and 53 assigned to the 25, 40, 60, and 75% decks, respectively. Indeed, no significant linearly increasing trend in probability estimation across the decks was observed in the nonlearners ($F_{(1,27)} = 1.17$; $p > 0.2$).

Reaction times and completed trials

To address the possibility that subjects in the nonlearner group failed to show learning purely because of a lack of engagement or attention to the task, we examined the reaction time taken to respond for each group. Should the groups differ in the time to take a decision this could support the possibility that subjects in the two groups were differentially engaged in the task. However, a direct comparison between reaction times taken to respond to each of the high-probability decks (60 and 75%) separately and together (referred to as HP decks) along the seven blocks of 20 trials showed that the learners and nonlearners did not differ significantly in reaction times (for choosing the 75% deck, $F_{(6,156)} = 1.08$, $p = 0.375$; for choosing the 60% deck, $F_{(6,132)} = 1.06$, $p = 0.387$; for choosing both HP decks, $F_{(6,132)} = 0.926$, $p = 0.478$). The comparison was performed only on the 60 and 75% decks as many of the subjects in the learners group ceased choosing the 25 and 40% decks along the task and therefore only the analysis on the 60 and 75% decks included an adequate number of subjects (at least $n = 12$ in each group). A set of two sample independent Student's *t* tests (Table 1) comparing the total RTs of each group, in the task and control sessions separately, revealed no significant difference between the two groups in either the task or control conditions (task, $p > 0.8$; control, $p > 0.2$). An additional analysis comparing the number of successfully completed trials, where subjects responded within the 2 s limit, in the task and control sessions separately, also revealed no significant differences between groups for both the task and control conditions (task, $p > 0.5$; control, $p > 0.8$).

Psychological questionnaires
We did not find any significant difference (Student's t test, all p values > 0.2) between learners and nonlearners in any of the questionnaires filled by subjects. The only questionnaire showing a tendency was the BDI with $p = 0.09$ (two-tailed, $t_{(27)} = 1.77$) with higher depression scores for the learner group. This trend is consistent with the finding of the study by Joel et al. (2005) in which major depressive disorder patients were most likely to belong to the group of learners compared with the other groups in that study. It should be noted that none of the subjects in our current study were clinically depressed according to BDI criteria (which requires a score higher than 18; the highest score in our sample was 15).

Model fitting
The best fitting model parameters for each group are shown in supplemental Tables 1 and 2, and analyzed in the supplemental Results (available at www.jneurosci.org as supplemental material). The relationship between model parameters and the number of choices of the HP decks (75 and 60%) in the last 40 trials of the task is illustrated in supplemental Figure 1 (available at www.jneurosci.org as supplemental material).

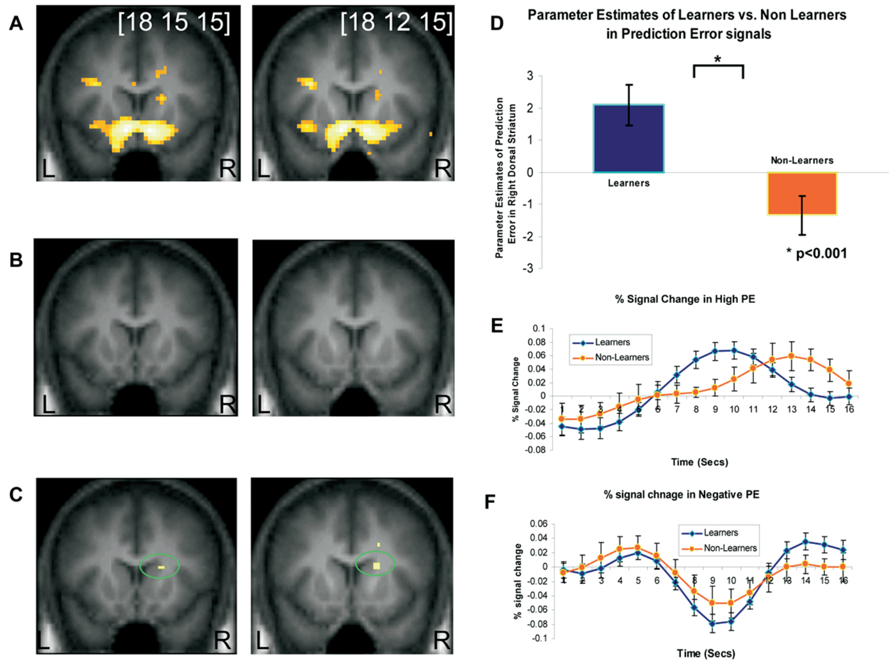

Figure 2. Random effects analysis showing PE correlations in ventral and dorsal striatum. ***A***, The learners group showed significant correlations in bilateral ventral striatum and right dorsal striatum ($n = 17$; $p < 0.001$). ***B***, The nonlearners group did not show significant correlations in a similar threshold ($n = 12$; $p < 0.001$). ***C***, A direct comparison between PE correlated activity in the learners group and the nonlearners group, showed enhanced activity in learners compared with nonlearners in right dorsal striatum ($p < 0.001$). ***D***, Parameter estimates of the direct comparison between learners and nonlearners. ***E***, ***F***, Time courses of the two groups in the right dorsal striatum during trials with high PE (***E***) and negative PE (***F***) learners show stronger activity in both of these trial types than nonlearners.

Neuroimaging results

Learners
In the learner group, we found significant correlations between our model-derived learning signals and neural activity in both the ventral and dorsal striatum, significant at $p < 0.001$ (Fig. 2*A*). These results are consistent with previous reports of prediction error signals in these areas during instrumental conditioning.

Nonlearners
In the nonlearner group, we did not find significant correlations with PE signals in either ventral or dorsal striatum at $p < 0.001$. Instead, we found only weak striatal correlations at a much lower significance level ($p < 0.05$) (Fig. 2*B* presents the nonlearners at $p < 0.001$). The regression coefficients for prediction errors in ventral striatum are plotted separately for learner and nonlearner groups in supplemental Figure 2 (available at www.jneurosci.org as supplemental material).

Learners and nonlearners
Moreover, in a direct statistical comparison between prediction error-related activity in the learner and nonlearner groups, we found that learners showed significantly greater prediction error-related activity in the dorsal striatum than nonlearners (Fig. 2*C*). This analysis is significant at $p < 0.001$ uncorrected, and also at $p < 0.05$ false discovery rate after small volume correction of the anatomically defined caudate nucleus. The group peak MNI coordinates are [18 15 15]. Parameter estimates of the direct comparison between learners and nonlearners in these coordinates can be seen in Figure 2*D*. At the whole brain level, no regions were found to survive whole brain correction at $p < 0.05$ (familywise error corrected). We report areas showing effects outside of our striatal regions of interest at $p < 0.001$ uncorrected in supplemental Table 3 (available at www.jneurosci.org as supplemental material).

Time course plots from dorsal striatum
For illustration purposes and to examine the trial averaged time course of activity in this area, we separated trials according to the magnitude of prediction errors predicted to occur on those trials. We extracted time courses from all trials with a high positive prediction error at the time of outcome (PE $>$0.4) from the peak voxels showing PE related activity in each individual subject separately for the learner and nonlearner groups. The rationale for searching individual peaks is to overcome anatomical misalignment. The group averaged time course data for high positive prediction errors are reported in Figure 2*E*. This plot reveals that the learner group shows a strong positive BOLD signal increase at 9 s into the trial, consistent with a prediction error being solicited at approx. 3.5 s into the trial at the time when the outcome was revealed (taking into account a 6 s hemodynamic lag). However, the nonlearners did not show a strong PE signal at the expected time in the trial. Instead, this group showed evidence of a later peak in signal, a response profile not accounted for by the RL model. A similar time course analysis was conducted for trials associated with a strong negative prediction error (where PE is less than -0.4) (Fig. 2*F*). A difference was also observed between the groups in the magnitude of negative prediction errors elicited in nonlearners compared with learners, although the difference between groups appeared more subtle than in the positive prediction error case.

Differences between effects of positive and negative prediction errors

To test for a significant difference between the degree of impairment of nonlearners in generating positive compared with negative prediction error signals, we modeled positive and negative prediction errors responses as separate regressors in an additional analysis and computed an interaction effect between group (learner vs nonlearner) and error type (positive vs negative) at the group random effects level. We found no significant interaction effect even at $p < 0.05$ uncorrected anywhere in the striatum, but confirmed our finding of a significant difference for the main effect of group in dorsal striatum (pooling over positive and negative error signals) at $p < 0.001$. These results indicate that nonlearners were not differentially impaired at generating positive compared with negative prediction errors, but rather showed an impairment at prediction error signaling in general.

Differences in responses to outcome receipt

We also tested for differences between the groups in responses to receipt of outcomes. No significant differences were found between the groups for this contrast even at $p < 0.005$ uncorrected. We then tested for differences in activity on gain compared with loss trials between the groups. Again, no significant differences were found even at $p < 0.005$ uncorrected.

Correlation analysis with learning efficacy

To address whether the results described above are an artifact of the specific categorization procedure used to assign subjects as learners and nonlearners, we also conducted a correlation analysis between the degree of learning in each subject as measured by the number of choices on the two high-probability decks in the last two blocks and the degree of prediction error activity across subjects. This analysis revealed an area of dorsal striatum whereby neural responses to prediction errors were correlated across subjects with the degree of learning exhibited by those subjects (Fig. 3*A*). This result bolsters the above categorical group analysis by demonstrating that the above results are not dependent on the specific group categorization procedure used. Figure 3*B* shows the scatter plot of single-subject parameter estimates versus learning criterion (number of choices on the 60 and 75% decks in the last 40 trials of the task).

Use of unitary versus separate model-parameter fits for the two groups

In all of the fMRI results reported above the model parameters (learning rate, exploration constant) were derived from behavioral fits to the learner group. These parameters were then used to generate regressors for both learner and nonlearner groups. This leaves open the possibility that nonlearners show poor correlations with prediction error signals because this group uses different model parameters than learners. To address this possibility we also derived model parameters separately from the nonlearner group and used these parameters to generate regressors for the nonlearners. Even in this case, we found the same pattern of results showing activity in ventral and dorsal striatum in learners, no observed activity in nonlearners and significant differences between learners and nonlearners in the same area of dorsal striatum (all at $p < 0.001$), suggesting that these effects are robust to the use of different model parameters in the nonlearner group.

Discussion

Human subjects vary widely in performance on choice and decision tasks (Stanovich and West, 2000; Stanovich, 2003; Newell, 2005). Here, we used a simple four-armed bandit task in which subjects are almost evenly split into two groups on the basis of their performance: those who do learn to favor choice of the optimal action, and those who do not. To determine the neural basis of this group difference we scanned both groups with fMRI while they performed the task and analyzed their neural activity using a reinforcement learning model. We found that these groups differ markedly in the degree to which RL signals in the striatum are engaged. Whereas learners showed robust prediction error signals in both the ventral and dorsal striatum during learning, nonlearners showed a significantly decreased prediction error signal. Moreover, neural activity in a region of dorsal striatum (and only there) was significantly more correlated with PE signals in learners than in nonlearners. Note that these results are not inconsistent, because they arise from tests asking different statistical questions. Although the visible difference between the threshold maps for each group separately might suggest that they differ throughout both dorsal and ventral striatum, when we formally compare the groups, we can only reject the null hypothesis of no difference in a more constrained area of dorsal striatum.

These findings suggest that one crucial factor which distinguishes those subjects who succeed in learning to choose advantageously on simple decision making tasks from those who do not, is the degree to which RL signals are engaged. These results are consistent with reinforcement learning theories of instrumental conditioning, whereby a reward prediction error signal is functionally involved in behaviorally expressed learning. Such signals are suggested to be carried by the phasic activity of dopamine neurons, which project heavily to both the ventral and dorsal striatum (Schultz, 2002). Previous human imaging studies have reported PE signals throughout the striatum during both classical and instrumental conditioning (O'Doherty, 2004). Modulation of dopamine by systemic administration of dopami-

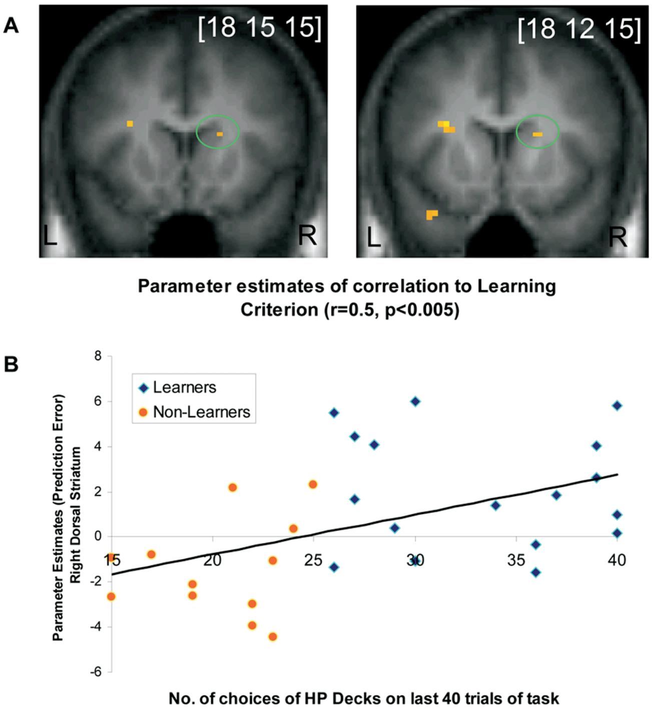

Figure 3. Second-level analysis showing simple regression between the learning criterion and PE in right dorsal striatum. ***A***, Simple regression analysis shows correlation in right dorsal striatum between the learning criterion used (number of choices on the two HP decks in the last 40 trials of the task) and PE contrast maps of each subject. ***B***, Scatterplot of the learning criterion and parameter estimates in the simple regression analysis shown in ***A***.

nergic agonists and antagonists has also been shown to modulate PE activity in striatum as well as to alter behavioral performance (Pessiglione et al., 2006). The present study demonstrates that even under natural conditions, without the external administration of drugs, neural PE signals relate to spontaneously occurring differences in behavioral performance. We also extend these results, by showing a graded relationship (in addition to a groupwise difference) between behavioral and neural measures. This latter finding helps to rule out the possibility that our results are attributable to our group categorization procedure. Rather these findings suggest that our results reflect a more continuous relationship between performance and prediction error signaling. When taken together with previous findings, these results lend support to the possibility that error signals in the striatum are *causally* related to behavioral performance in reward-related instrumental decision-making tasks in humans.

As discussed above, whereas nonlearners showed an absence of significant prediction error signals in both ventral and dorsal striatum, we only observed a significant difference in PE activity between groups in the dorsal striatum. These findings are broadly consistent with a role for dorsal striatum analogous to the "actor" in actor/critic models of instrumental conditioning (Joel et al., 2002; Suri, 2002; O'Doherty, 2004) whereby the dorsal striatum in particular is involved in implementing reward-based action selection (O'Doherty et al., 2004; Morris et al., 2006), whereas the ventral striatum is involved in reward-learning more generally. Our finding of a significant correlation across both groups between individual subject performance and the degree of prediction error activity in dorsal striatum indicate that our results are not merely an artifact of the specific criteria we used to split subjects into groups of learners and nonlearners, but rather, even when subjects are not split into arbitrary groups, performance on the task can be explained by activity in dorsal striatum. A previous study by Lohrenz et al. (2007) adds additional support to the claim that dorsal striatum is involved in implementing reward-based action selection. In that study the authors used a more complex decision making reward task and found that signals in dorsal striatum correlated with a novel "fictive error" signal that strongly predicted subjects' behavior in subsequent decisions.

Although we have shown that prediction error activity in striatum discriminates learners from nonlearners, the present study has not addressed why these two groups differ in the degree to which PE activity is elicited. One mundane possibility is that subjects in the nonlearner group were simply less motivated to perform well on the task, and/or failed to attend to the task. Our findings that the learner and nonlearner groups did not differ significantly in reaction times during task and/or control sessions, and additionally did not differ in the number of successfully completed trials, provides relatively strong evidence against this interpretation. Had subjects in one group been less motivated or less engaged in the task, then this should have been reflected by a consistent difference in the amount of time taken to make a choice on each trial or in the number of missed trials. The learner and nonlearner groups did not differ in age, years of education or in the Israeli equivalent of SAT scores. Both groups had on average 15 years of formal education and were approx. 1.5 SDs above the general Israeli population mean in SAT scores. Thus, the suboptimal performance on the task by the nonlearners is unlikely to be attributable to low intellect or education. Furthermore, the groups did not differ on standardized personality questionnaires suggesting that performance differences appear to be unrelated to commonly measured personality traits.

Dopaminergic drugs have been shown to have varying effects across individuals (Cools and Robbins, 2004), presumably reflecting differences in underlying dopaminergic function. In a similar manner our two groups may differ in the degree of endogenous striatal dopamine release or else in the degree of sensitivity of striatal neurons to afferent dopaminergic modulation. An important future step will be to measure dopamine uptake using ligand PET measures in these two groups, to determine whether they do differ in basic dopamine function. Such functional differences could emerge as a result of a genetic polymorphism or else as a result of experience-dependent effects. Moreover, it will be useful to determine whether differences in performance between subjects persist after repeated exposure to the same or related decision making tasks. If these groups demonstrate stable and consistent differences in performance, this would suggest that these groups represent distinct subpopulations with differential capacity to generate reward PE signals and, thus, to learn to choose adaptively on simple choice tasks.

However, a more parsimonious hypothesis is that learners and nonlearners differ not in dopaminergic physiology, but rather in their construal of the relevant features of the decision problem, or in reinforcement-learning terms: their model of the state space (Daw et al., 2006a). According to this explanation, the reward PE signal is engaged in the nonlearners, but responds to different and irrelevant features of the decision problem. This would lead to a failure to learn the task, and a failure on our part to detect the signals because we examined their relationship to task-relevant features. If this hypothesis is true, then subjects in the nonlearner group would be able to learn the task successfully once given appropriate instructions as to the relevant stimuli and states in the task.

To conclude, in the present study we show that prediction error activity in human striatum correlates with differences in behavioral performance on a simple choice-based decision making task. These findings suggest that the engagement of PE signals, likely originating from dopaminergic neurons in the midbrain, may play a critical role in facilitating appetitive instrumental learning in humans.

References

Beck A (1988) Beck depression inventory. London: The Psychological Corporation.

Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J (1961) An inventory for measuring depression. Arch Gen Psychiatry 4:561–571.

Cools R, Robbins TW (2004) Chemistry of the adaptive mind. Philos Transact A Math Phys Eng Sci 362:2871–2888.

Costa Jr PT, McCrae RR (1992a) Normal personality assessment in clinical practice: the NEO personality inventory. Psychol Assess 4:5–13.

Costa Jr PT, McCrae RR (1992b) Revised NEO personality inventory (NEO-PI-R) and NEO five-factor inventory (NEO-FFI) professional manual. Odessa, FL: Psychological Assessment Resources.

Daw ND, Doya K (2006) The computational neurobiology of learning and reward. Curr Opin Neurobiol 16:199–204.

Daw ND, Courville AC, Touretzky DS (2006a) Representation and timing in theories of the dopamine system. Neural Comput 18:1637–1677.

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

Dayan P, Abbott LF (2001) Theoretical neuroscience: computational and mathematical modeling of neural systems. Cambridge, MA: MIT.

Dayan P, Balleine BW (2002) Reward, motivation, and reinforcement learning. Neuron 36:285–298.

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

Foa EB, Huppert JD, Leiberg S, Langner R, Kichic R, Hajcak G, Salkovskis PM (2002) The obsessive-compulsive inventory: development and validation of a short version. Psychol Assess 14:485–496.

Friedland N (1998) Games of luck and games of chance: the effect of luck-versus chance-orientation on gambling decisions. J Behav Decision Making 11:161–179.

Joel D, Niv Y, Ruppin E (2002) Actor-critic models of the basal ganglia: new anatomical and computational perspectives. Neural Netw 15:535–547.

Joel D, Zohar O, Afek M, Hermesh H, Lerner L, Kuperman R, Gross-Isseroff R, Weizman A, Inzelberg R (2005) Impaired procedural learning in obsessive-compulsive disorder and Parkinson's disease, but not in major depressive disorder. Behav Brain Res 157:253–263.

Klein MH, Benjamin LS, Rosenfeld R, Treece C, Husted J, Greist JH (1993) The Wisconsin Personality Disorders Inventory: development, reliability and validity. J Personal Disord 285–303.

Knutson B, Westdorp A, Kaiser E, Hommer D (2000) FMRI visualization of brain activity during a monetary incentive delay task. Neuroimage 12:20–27.

Lau B, Glimcher PW (2005) Dynamic response-by-response models of matching behavior in rhesus monkeys. J Exp Anal Behav 84:555–579.

Lohrenz T, McCabe K, Camerer CF, Montague PR (2007) Neural signature of fictive learning signals in a sequential investment task. Proc Natl Acad Sci USA 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.

McClure SM, York MK, Montague PR (2004) The neural substrates of reward processing in humans: the modern role of FMRI. Neuroscientist 10:260–268.

Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936–1947.

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

Newell BR (2005) Re-visions of rationality? Trends Cogn Sci 9:11–15.

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'Doherty JP (2004) Reward representations and reward-related learning in the human brain: insights from neuroimaging. Curr Opin Neurobiol 14:769–776.

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.

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

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.

Redgrave P, Prescott TJ, Gurney K (1999) Is the short-latency dopamine response too short to signal reward error? Trends Neurosci 22:146–151.

Rodriguez PF, Aron AR, Poldrack RA (2006) Ventral-striatal/nucleus-accumbens sensitivity to prediction errors during classification learning. Hum Brain Mapp 27:306–313.

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

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

Seger CA, Cincotta CM (2005) The roles of the caudate nucleus in human classification learning. J Neurosci 25:2941–2951.

Stanovich KE (2003) Is probability matching smart? Associations between probabilistic choices and cognitive ability. Mem Cognit 31:243–251.

Stanovich KE, West RF (2000) Individual differences in reasoning: implications for the rationality debate? Behav Brain Sci 23:645–665.

Suri RE (2002) TD models of reward predictive responses in dopamine neurons. Neural Netw 15:523–533.

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

Tricomi EM, Delgado MR, Fiez JA (2004) Modulation of caudate activity by action contingency. Neuron 41:281–292.

Zink CF, Pagnoni G, Chappelow J, Martin-Skurski M, Berns GS (2006) Human striatal activation reflects degree of stimulus saliency. Neuroimage 29:977–983.