◆ Human Brain Mapping 36:793–803 (2015) ◆

Changes in Corticostriatal Connectivity During Reinforcement Learning in Humans

Guillermo Horga,[1] Tiago V. Maia,[1,2] Rachel Marsh,[1] Xuejun Hao,[1] Dongrong Xu,[1] Yunsuo Duan,[1] Gregory Z. Tau,[1] Barbara Graniello,[1] Zhishun Wang,[1] Alayar Kangarlu,[1] Diana Martinez,[1] Mark G. Packard,[3] and Bradley S. Peterson[4]*

[1]*Department of Psychiatry, New York State Psychiatric Institute and College of Physicians and Surgeons, Columbia University, New York, New York*
[2]*Instituto de Medicina Molecular, Faculdade de Medicina da Universidade de Lisboa, Lisboa, Portugal*
[3]*Department of Psychology, Texas A&M University, College Station, Texas*
[4]*Institute for the Developing Mind at Children's Hospital Los Angeles and the Keck School of Medicine, University of Southern California*

Abstract: Many computational models assume that reinforcement learning relies on changes in synaptic efficacy between cortical regions representing stimuli and striatal regions involved in response selection, but this assumption has thus far lacked empirical support in humans. We recorded hemodynamic signals with fMRI while participants navigated a virtual maze to find hidden rewards. We fitted a reinforcement-learning algorithm to participants' choice behavior and evaluated the neural activity and the changes in functional connectivity related to trial-by-trial learning variables. Activity in the posterior putamen during choice periods increased progressively during learning. Furthermore, the functional connections between the sensorimotor cortex and the posterior putamen strengthened progressively as participants learned the task. These changes in corticostriatal connectivity differentiated participants who learned the task from those who did not. These findings provide a direct link between changes in corticostriatal connectivity and learning, thereby supporting a central assumption common to several computational models of reinforcement learning. *Hum Brain Mapp 36:793–803, 2015.* © 2014 Wiley Periodicals, Inc.

Key words: reinforcement learning; computational model-based fMRI; functional connectivity; putamen

Additional Supporting Information may be found in the online version of this article.

Guillermo Horga and Tiago V. Maia contributed equally to this work.

Contract grant sponsor: NIMH; Contract grant numbers: K02-74677, K01-MH077652 and K23MH101637; Contract grant sponsor: NIH/NIBIB; Contract grant numbers: 1R03EB008235 and KL2RR024157; Contract grant sponsors: National Alliance for Research on Schizophrenia and Depression, Sackler Institute for Developmental Psychobiology, Columbia University, and Opening Project of Shanghai Key Laboratory of Functional Magnetic Resonance Imaging (East China Normal University).

*Correspondence to: Bradley S. Peterson, M.D., 4650 Sunset Blvd, Los Angeles, CA 90027. E-mail: bpeterson@chla.usc.edu

Received for publication 22 January 2014; Revised 4 September 2014; Accepted 9 October 2014.

DOI: 10.1002/hbm.22665
Published online 12 November 2014 in Wiley Online Library (wileyonlinelibrary.com).

INTRODUCTION

Computational models form the backbone of our current theoretical understanding of reinforcement learning (RL) in humans and other animals [Glimcher, 2011; Maia and Frank, 2011]. Computational RL models come in two basic flavors that parallel two types of learning that have long been acknowledged in psychology: model-free approaches based on stimulus-response (S-R) learning and model-based approaches akin to planning. Humans and other animals likely implement both types of learning [Daw et al., 2005].

The precise neural instantiation of model-based RL remains unknown despite important recent advances in this area [Balleine and O'Doherty, 2010; Daw et al., 2011]. The neural instantiation of model-free, S-R learning, in contrast, is better understood. Several RL models suggest that S-R learning depends on plasticity in corticostriatal synapses [Barto, 1995; Frank, 2005; Maia, 2009]—more specifically, on changes in the synaptic connections between sensory cortical regions that represent stimuli or situations and the putamen in humans and other primates or its homologue in rodents, the dorsolateral striatum [Balleine and O'Doherty, 2010; Yin and Knowlton, 2006]. The models predict that these changes in corticostriatal synapses in the motor cortico-basal ganglia-thalamo-cortical (CBGTC) loop depend on phasic dopaminergic firing [Maia, 2009], consistent with converging evidence from empirical work in non-human animals [Centonze et al., 2001; Charpier and Deniau, 1997; Pawlak and Kerr, 2008]. Because rapid changes in synaptic efficacy accompany learning-related plasticity [Xu et al., 2009], learning likely modifies the functional coupling between presynaptic and postsynaptic neurons—a phenomenon that can be studied using functional-connectivity tools in fMRI. Some fMRI studies have indeed used RL models to assess changes in neural connectivity during decision making following a learning phase [Wunderlich et al., 2012] or during specific phases of learning, although not as a function of changes in learning signals [van den Bos et al., 2012]. Other studies have instead focused on interindividual differences in structural connections in relation to habitual tendencies [de Wit et al., 2012]. Finally, few fMRI studies have used RL models to specifically assess changes in striatal connectivity as participants learn the relational association between pairs of stimuli (S-S learning) [den Ouden et al., 2010; den Ouden et al., 2009; Wimmer et al., 2012], but none have assessed such changes as a function of S-R learning or other forms of RL.

We used fMRI based on computational modeling (i.e., computational fMRI) [O'Doherty et al., 2007] to study changes in corticostriatal connectivity as human participants performed a task analogous to the "win-stay" version of a radial-arm maze task [Packard et al., 1989; Packard and Knowlton, 2002]. Participants were instructed to find hidden monetary rewards by navigating a virtual-reality eight-arm radial maze (Fig. 1). They had to learn to enter lit maze arms, which contained a reward at the end of the arm. Prior work in rodents has shown that forming an association between the light and the response of entering the lit arm on this version of the task is insensitive to outcome devaluation and, therefore, relies on S-R learning [Sage and Knowlton, 2000], which is mediated by the dorsolateral striatum [McDonald and White, 1993; Packard et al., 1989]. Building on this prior work, we hypothesized that human learning of this putative S-R association would be accompanied—indeed, driven—by two main changes at the neural level: (1) an increase in functional connectivity between sensory cortices and the putamen, reflecting the strengthening of the S-R association, and (2) an increase in activity in the putamen, reflecting the increasing engagement of the habit system (driven by the increased connectivity between the sensory representation of the light and the representation of the response in the putamen).

MATERIALS AND METHODS

Participants

The participants in this study were 55 healthy individuals, aged 14–59 years (mean $\pm$ s.d., 27 ± 10 years; 41 females), who had no history of neurological illness or any lifetime Axis I psychiatric disorder. The study was approved by the Institutional Review Board of the New York State Psychiatric Institute and the Department of Psychiatry of Columbia University.

Behavioral Paradigm

Virtual environments were generated with C++ and OpenGL. The virtual environments consisted of an eight-arm radial maze with a central starting location and a low outer-perimeter wall. The maze was surrounded by a naturalistic landscape. Prior to scanning, participants underwent a training session on a desktop computer to practice using a joystick to navigate freely about a virtual maze that was similar in appearance to the maze used during scanning. During scanning, stimuli were presented through nonmagnetic goggles (Resonance Technology, refresh rate = 60 Hz), and participants used an MRI-compatible joystick (Current Designs, Inc.) to navigate the maze. Before entering the scanner, participants were told that they would find themselves in the center of a virtual maze with eight identical runways extending outwards and that hidden rewards ($) would be available at the end of some runways. Participants were instructed to find the rewards; they were unaware that they would not be given actual monetary rewards for their performance but rather paid a fixed amount for their participation at the end of the study. In a first session, participants executed a spatial-learning version of the task in which they had to learn to use fixed extra-maze cues to navigate and earn

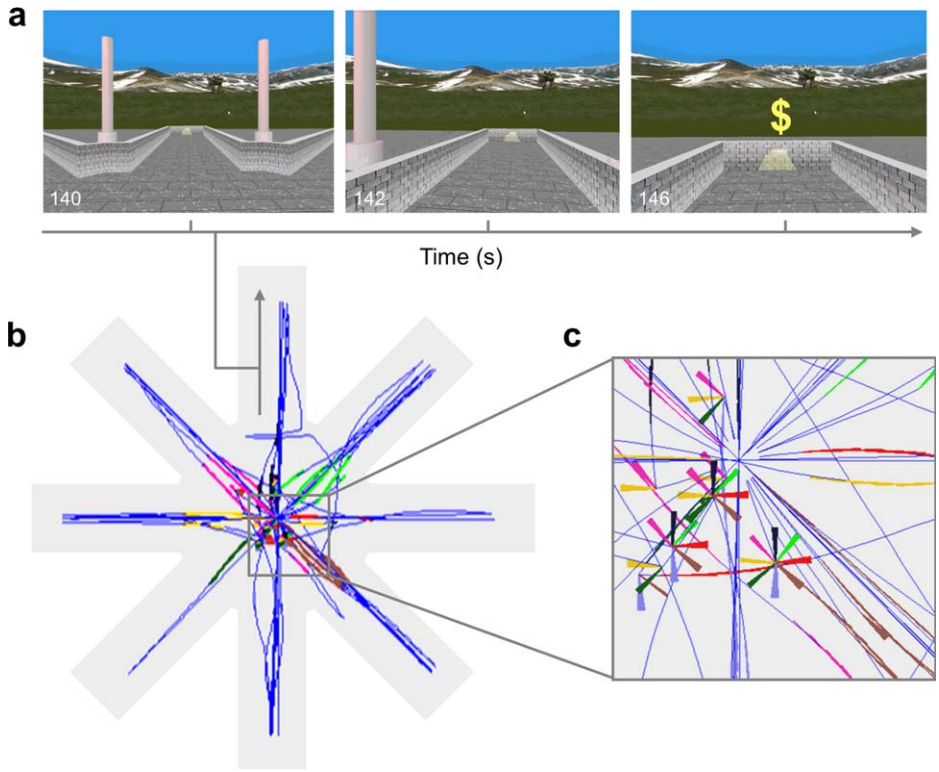

Figure 1.

The virtual-reality maze task. (**a**) Screen captures of the participant's view when traversing a lit arm and when arriving at the end of the arm. (**b**) Aerial view of the virtual maze (not seen by participants) showing a reconstructed trajectory of a simulated participant along the maze (dark blue) and the points along the trajectory when the participant was facing the entrance to each of the eight arms (in different colors). (**c**) Detail of (b). Rotations without displacement appear as portions of multicolor pinwheels, the differing colored arms of which represent the sighting of different arms as viewed from the participant's position at the center of the pinwheel.

the rewards. The details of that version of the task are reported elsewhere [Marsh et al., 2010] and will not be described here. Participants were then presented with the message, "New Experiment! Find the $$." This signaled the beginning of the "win-stay" task. In this task, four arms were illuminated (lit) and the other four were not. Each lit arm was baited with two rewards. Participants obtained rewards (feedback in the form of a dollar sign, presented for 1 s) after they reached the end of a lit arm, although they were never told where the rewards were located or that multiple rewards were present at the same location. The illumination of a lit arm ceased upon receipt of the second reward located at the end of that arm. Each trial began at the center platform and ended when the participant reached the end of an arm. After reaching the end of an arm, participants reappeared automatically in the middle of the center platform to initiate a new trial, with the initial viewing perspective determined randomly. The run terminated when the participant found all eight hidden rewards (i.e., run duration depended on each participant's performance but in all cases involved eight visits to lit arms) or after 5 min elapsed. Self-timing and other task features were chosen to maximize comparability with the rodent version of the task. Extra-maze cues were pseudorandomly interchanged at the end of each trial, thereby precluding use of a spatial strategy to find the rewards. Rather, finding rewards in this task requires an S-R strategy in which participants have to learn the association between the light stimulus and the response of entering, based on their history of reinforcements.

Behavioral Analyses and RL Model

We reconstructed each participant's trajectory and orientation on the virtual maze. We considered that every sighting of an arm elicited a choice about whether or not to enter that arm. We recorded entering choices directly and a nonentering choice whenever an arm was viewed by the participant but the participant did not enter that arm (Supporting Information). We then fitted a *Q*-learning model to each participant's choice behavior. *Q*-learning [Watkins

and Dayan, 1992] involves learning the value $Q(s, a)$ of an action (or response) a in a state s. In our case, the relevant states were the facing of a lit or an unlit arm, and the relevant actions were entering or not entering that arm. We used standard Q-learning rules to update Q based on the prediction errors δ elicited by the presentation of the outcome (reward or no reward at the end of the arm; Supporting Information).

We tested several RL model variants (Supporting Information). A first version (double-α model) had separate learning rates (α^+ and α^-) for positive and negative prediction errors, thereby allowing learning to differ for reward versus no-reward outcomes [Frank et al., 2007]. A second version (single-α model) had a single learning rate (α) for both positive and negative prediction errors. A third version (zero-α^+ model) explicitly assumed that no learning to enter lit arms could occur ($\alpha^+ = 0$). By comparing the fit to this zero-α^+ model with the fit to the double-α model on a participant-by-participant basis, we determined which participants exhibited any evidence for learning to enter lit arms (i.e., which participants showed more evidence for learning than for no learning given their behavioral data). To do so, we compared models using a goodness-of-fit index penalized by model complexity, namely the Akaike information criterion (AIC), and selected the best-fitting model as that with the minimum AIC value for a given participant (see Supporting Information for a detailed description).

Image Acquisition

Images were acquired on a GE Signa 3T LX scanner (Milwaukee, WI) with a standard quadrature head coil, using a T2*-sensitive gradient-recalled, single-shot, echo-planar pulse sequence with TR = 2800 ms, TE = 25 ms, flip angle = 90°, single excitation per image, FOV = 24 cm × 24 cm, 64 × 64 matrix, 43 slices 3 mm thick, no gap.

Image Analysis

Individual-level analyses were carried out with SPM8 using a General Linear Model (GLM) with a weighted least-squares algorithm, following standard preprocessing steps (Supporting Information). One participant was excluded from the imaging analyses due to extreme head motion in the scanner. Each choice and outcome period was modeled separately as an independent regressor in a GLM. Choice periods (from the beginning of a trial until the participant traversed 10% of the length of an arm, for consistency with the animal literature) were modeled as boxcar functions with length equal to the duration of the corresponding period. Outcome periods (arrival at the end of an arm) were modeled as impulse functions (0 ms), as prediction error signals at outcome are phasic signals with a relatively negligible duration of a few hundred milliseconds [Schultz et al., 1997]. The boxcar and impulse functions were then convolved with a canonical double-gamma hemodynamic response function [Friston et al., 1998]. The resulting series of choice and outcome beta maps from this model, which represent the magnitude of activation during each choice and outcome period, respectively, were Z-normalized and treated as the dependent variables in additional first-level GLMs, using a beta-series analysis described next.

Individual-level, beta-series analysis of learning-related changes in activation

We focused our analyses on lit arms because we were interested specifically in the neural correlates of establishing and strengthening of S-R associations, and because, for consistency with the rodent version of the task, participants were simply instructed to find rewards (and not to avoid unrewarded arms). Our design, in fact, would be less suited to study the weakening of S-R associations, for two reasons. First, although such weakening did occur for some participants who clearly learned to avoid unlit arms, the instruction to earn all rewards regardless of visits to unlit arms did not stress speed, and therefore, no explicit incentive was present to learn to avoid unlit arms. Participants thus had a very variable number of visits to unlit arms. Second, the number of visits to unlit arms correlated (inversely) with learning speed, thereby confounding learning and the number of data points available for analysis. We, therefore, restricted our analyses to the δ signals that occurred at the end of lit arms (during outcome periods), henceforward simply referred to as PEs, and to the Q values of entering lit arms (during choice periods), henceforward simply referred to as Q. Note that modeling of Q and PE signals as events occurring at separate times within a trial (choice vs. feedback periods, respectively) circumvents their tendency to correlate (negatively) at the trial level given the mathematics of RL (see eq. 1 in Supporting Information). To identify the neural correlates of Q, we analyzed the beta-map series corresponding to choice periods that terminated when participants chose to enter a lit arm. We built two GLMs—Q-GLM and PE-GLM—each of which contained a model-derived signal (Q and PE, respectively) matched to the corresponding trial period (choice- or outcome-related beta-map series as the dependent variable, respectively) and a global intercept. As a control analysis, we also used an extended Q-GLM model that included linear effects of time at each choice period as a nuisance regressor.

Individual-level, psychophysiological interaction analysis of learning-related changes in connectivity

To test the hypothesis that corticostriatal connections strengthened as participants acquired the S-R association, we assessed whether whole-brain functional connectivity with the functionally defined striatal region-of-interest (ROI) that encoded Q changed as a function of learning.

To account for interindividual differences in brain loci engaged with learning, we extracted functional time courses (i.e., the beta-map series corresponding to choice periods) from participant-specific seeds (ROIs). To select these seeds, we searched for participant-specific maxima related to the effect of interest (Q effect) within a cluster that had positive findings for that effect at the group level, and that fell within the corresponding anatomical ROI of the putamen according to the AAL atlas in the PickAtlas toolbox [Maldjian et al., 2003]. To analyze changes in the connectivity of voxels throughout the brain with each participant's seed during learning, we generated a separate GLM model identical to Q-GLM (above) but with the addition of two regressors: one that corresponded to the seed timecourse and an interaction term calculated by multiplying the seed timecourse by the model-derived Q values. The regression coefficient (beta) map associated with the interaction term represented changes in functional connectivity between each voxel and the seed as a function of learning.

Group-level analyses

We applied a second-level Bayesian analysis to detect a group random effect by estimating the posterior probability that the effect exists based on the observed data [Klein et al., 2007; Neumann and Lohmann, 2003]. This approach to second-level analysis does not require adjustment for multiple comparisons because it has no false positives and does not depend on whether the analysis is performed on a single voxel or the entire brain [Neumann and Lohmann, 2003]. To reduce the number of statistical tests and based on our strong a priori hypothesis that the learning signals of interest would be represented in the striatum, we nonetheless limited our search space for the analysis of learning-related changes in activation to striatal voxels, as defined by the AAL atlas.

To ensure interpretability and comparability of Q signals between learners and nonlearners, we first estimated a $\overline{Q}$-GLM for each participant in which $\overline{Q}$ represented the average Q time series in learners (calculated using the average α in this group), rather than the participant-specific Q time series, and then compared the resulting beta maps across groups. We chose this approach, analogous to that used in prior work [Schonberg et al., 2007], because Q time series in nonlearners by definition show no systematic changes (in the case of no learning, the PE time series would equal obtained outcomes and the Q series would be constant), and therefore, the betas associated with Q in this group are uninterpretable. Individual betas associated with the average-learner Q time series, conversely, can be interpreted as indicating how strongly neural signals relate to a canonical time series that represents average learning. For the same reason, we also based psychophysiological interaction (PPI) comparisons between the groups on a GLM that used the $\overline{Q}$ time series. We considered voxelwise findings as significant whenever posterior probability (PP) ≥ 0.95, which can be considered equivalent to a corrected P-value of 0.05. Post hoc, ROI-based brain-behavior correlations within regions with significant effects of learning (i.e., within striatal voxels) used an uncorrected threshold of $P = 0.01$ ($P = 0.05$ for exploratory analyses).

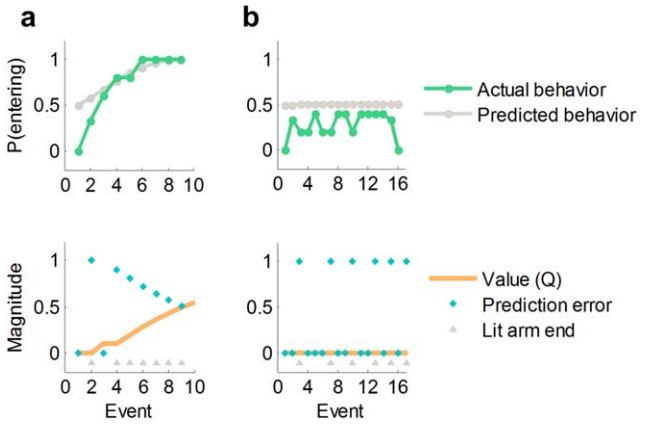

Figure 2.
Examples of behavior and model-derived predictions for a participant who learned the task (**a**) and one who did not (**b**). The top row depicts actual and model-predicted probability of entering a lit arm at its sighting (smoothed with a five-point moving average). The bottom row represents the model predictions concerning values and prediction errors for the same participants (magnitude in a.u.). Participant (a) learns the value of lit arms (Q increases and PEs decrease throughout the task) and learns to enter those arms. Participant (b) shows no evidence of learning, as evidenced by constant PEs and Q and by the fact that the participant continues to enter lit arms less than 50% of the time throughout the task (note that only part of the run is shown for this participant for clarity of display). These participants were classified as learner (a) and nonlearner (b), respectively, by our classification algorithm (Materials and Methods). See Supporting Information for an example computation of Q values based on the model-fitting procedure.

RESULTS

Behavioral Analyses and Model Fit

Participants learned to enter lit arms over the course of the experiment (Figs. 2 and 3a). They made an average of 14 choices of whether or not to enter a lit arm during the scan run (s.d. 6.77; mean run duration ± s.d., 117.03 ± 78.68 s). The percentage of entering choices increased in the second half of lit-arm choices relative to the first half ($t_{54} = -6.25$, $P = 6 \times 10^{-8}$, paired t-test), and the time to obtain the last four rewards (out of the total of eight) was shorter than the time to obtain the first four rewards ($t_{54} = 2.03$, $P = 0.0466$, paired t-test). None of these effects were significantly associated with age (Ps > 0.05).

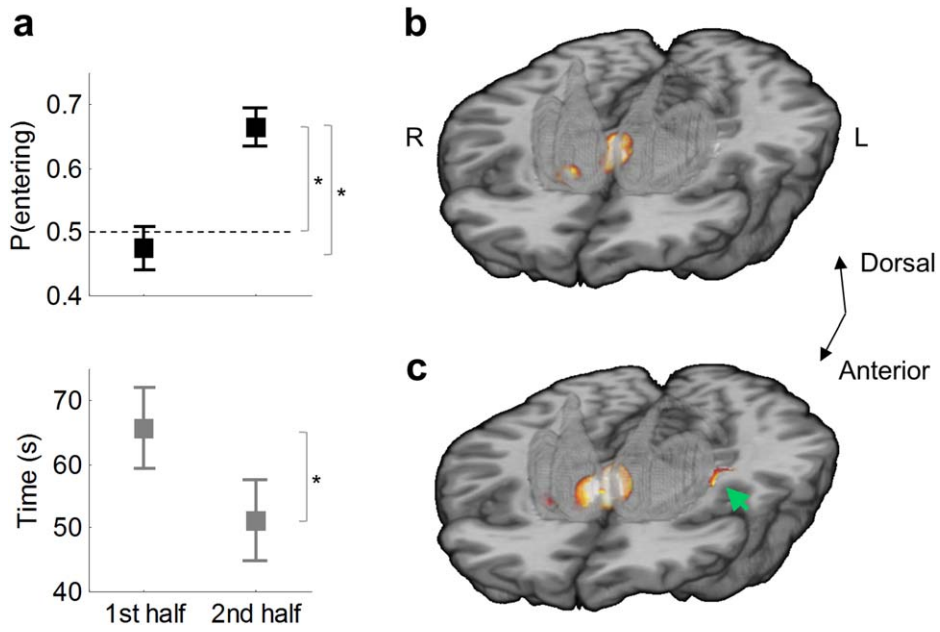

Figure 3.

Behavioral and imaging findings for all participants ($n = 54$). (**a**) Percentage of entering choices for lit arms (top) and time to obtain the rewards (bottom) in the first and second half of the run. Error bars represent s.e.m. Asterisks indicate statistically significant effects at $P < 0.05$. The dashed line indicates chance-level performance. (**b**) Prediction error signal during outcome periods (hot colors) within the striatum in the ventral striatum and anteromedial caudate (caudate head). Three-dimensoinal template of the striatum (semi-transparent gray). (**c**) *Q* signal during choice periods (hot colors) in the ventral and anteromedial regions of the striatum and left posterolateral putamen (green arrow). Maps are thresholded at posterior probability (PP) ≥ 0.95. Supporting Information Figs. S2 and S3 present whole-brain results for the PE and Q signals, respectively.

To determine the best fitting of the RL model variants, we first compared the fit of each model variant to the behavioral data using the AIC (Supporting Information Table S1). The best-fitting model was the one that treated the absence of reward at the end of unlit arms as a negative outcome (coded as -1) rather than as a neutral outcome (coded as 0) and that allowed for different learning rates for positive and negative outcomes (the double-α model), consistent with prior demonstrations that these two types of learning depend on dissociable basal ganglia pathways [Frank et al., 2007; Maia and Frank, 2011]. Participants likely treated the absence of reward as a negative outcome because they learned to expect rewards during the task or generalized some of the positive value associated with lit arms to unlit arms. In either of those cases, the absence of reward at the end of unlit arms would elicit a negative prediction error. Such coding of neutral outcomes relative to overall task expectations has been previously demonstrated [O'Doherty et al., 2003]. Even though our dataset contained fewer data points than those in typical model-based studies, a parameter-recovery analysis showed that the results of our model-fitting procedure were robust (Supporting Information Fig. S1).

Changes in Neural Activity Associated With Learning

Activation in ventral and anteromedial portions of the striatum (nucleus accumbens and caudate head, respectively) in the whole group ($n = 54$) correlated positively with PEs during outcome periods (PP ≥ 0.95; Fig. 3b and Supporting Information Fig. S2), replicating previous fMRI studies [Maia, 2009; Pessiglione et al., 2006]. We found *Q* signals during choice periods in ventral anteromedial regions of the striatum, not only partially overlapping regions that encoded PEs but also in a posterolateral portion of the putamen (PP ≥ 0.95; Fig. 3c and Supporting Information Fig. S3). These effects were independent of age (Ps > 0.05).

To interpret the neural correlates of *Q* more confidently in terms of learning, we divided participants into learners and nonlearners based on their behavioral performance (Materials and Methods). Our procedure classified 15 (27.8%) participants as learners and 39 (72.2%) as nonlearners. Learners and nonlearners were comparable on sociodemographic characteristics (mean age: 30.6 vs. 26.9; females: 43.7% vs. 43.6%; full-scale IQ: 111.9 vs. 110.8,

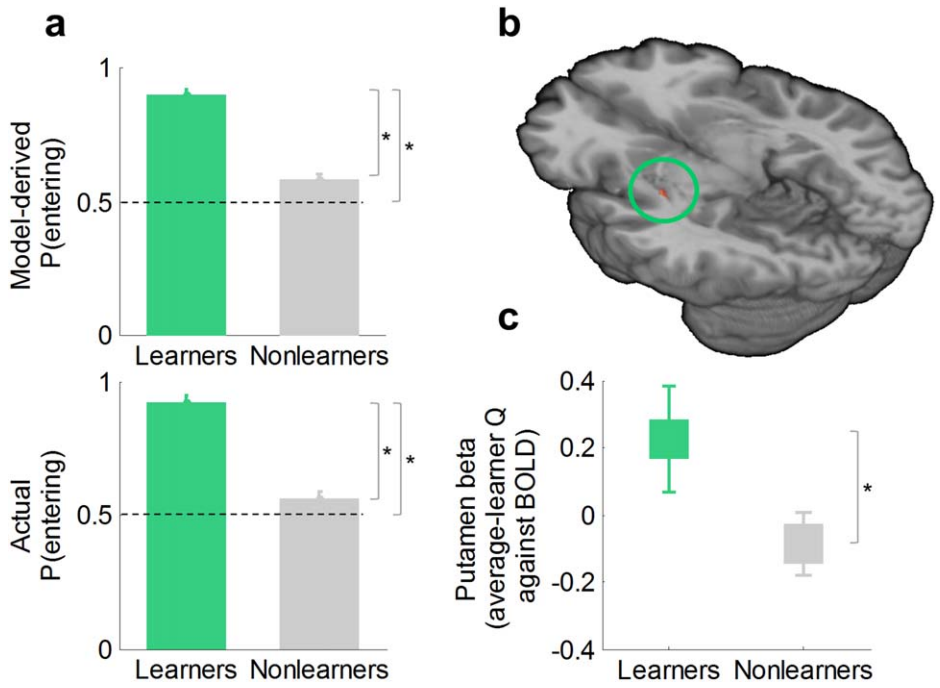

Figure 4.

Behavioral and imaging findings in learners versus nonlearners. (**a**) Model-derived (top) and model-independent (bottom) differences in behavior between learners and nonlearners among the second half of lit-arm choices. Error bars denote s.e.m. Asterisks indicate $P < 0.05$. The mean percentage of "entering" choices (±s.d.) was 92.08% ± 11.15% in learners and 55.99% ± 16.29% in nonlearners ($t_{53} = 8.09$, $P = 7 \times 10^{-12}$). (**b**) Striatal Q effect during choice periods in the left posterolateral putamen in learners only ($n = 15$; peak MNI coordinates [x,y,z]: −33, −15, −2 mm). Map thresholded at $PP \geq 0.95$. (**c**) Q effect in learners versus nonlearners in the right posterolateral putamen (MNI coordinates [x,y,z]: 33, −22, −2 mm, $t_{52} = 1.68$, $P = 0.048$).

respectively; Ps > 0.5) and motion (cumulative motion and motion peaks, Ps > 0.3) but had marked differences in behavioral indices of learning independent from the model (Fig. 4a). Whereas learners clearly performed better than chance in the second half of the run, nonlearners continued to perform at chance level (Wilcoxon signed-rank test: learners $P = 3 \times 10^{-6}$; nonlearners $P = 0.061$). Furthermore, model-derived and model-independent learning indices correlated highly with one another (ΔQ with percentage of entering choices in the second half: Spearman's $\rho = 0.92$, $P = 9 \times 10^{-20}$).

Next, we analyzed Q correlations with striatal BOLD signal during choice periods in the 15 learners and found that only a region in the left posterolateral putamen showed increased activation with increasing values of Q (Fig. 4b). This region overlapped with the region in the putamen that showed Q signals in the entire sample (Fig. 3c). The supplementary motor area (SMA) and premotor cortex (within Brodmann area 6) also displayed Q signals in learners. Using a more lenient threshold ($P = 0.01$, uncorrected), we also identified Q signals in learners in the contralateral posterolateral putamen, but not in other regions of the striatum. All of these neural Q effects in learners persisted after controlling for linear time effects. In contrast, our analyses in the 39 nonlearners found no evidence of progressive activation in cortical or striatal motor regions—either tracking a linear function of time or an average-learner $\overline{Q}$ time series—supporting the interpretation that the neural Q effect in learners represents learning and not another, unspecific process. The direct comparison of $\overline{Q}$ effects in learners versus nonlearners revealed that learners had stronger $\overline{Q}$ signals in the right posterolateral putamen (Fig. 4c), as well as in other regions within the striatum (anteromedial and posterolateral caudate). Furthermore, post hoc exploratory analyses using a lenient threshold of uncorrected $P = 0.05$ showed that participants who demonstrated the most learning had stronger correlations with Q in the posterolateral putamen ($t_{52} = 2.40$, $P = 0.043$, uncorrected, MNI coordinates [x,y,z]: 33, −16, −5 mm, in the whole group), even when the analysis was restricted to learners (conjointly significant group effect and correlation with participants' ΔQ, within learners, $t_{13} = 2.40$, $P = 0.016$, uncorrected, [x, y, z]: 33, −22, 1 mm). Even though the latter set of results are simply

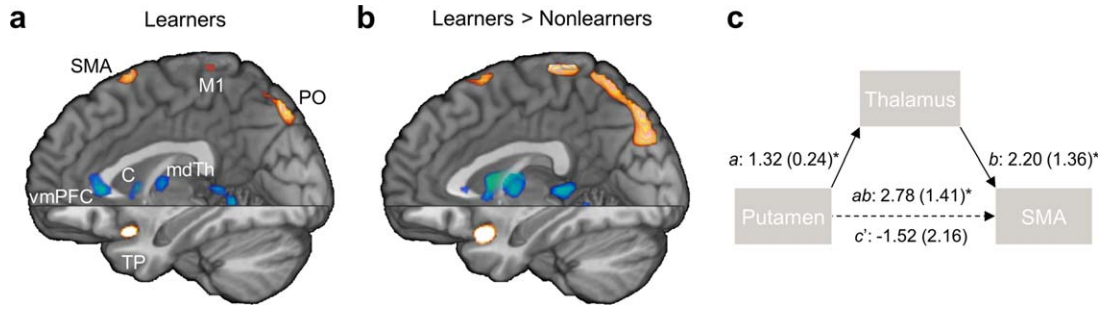

Figure 5.

Learning-related changes in connectivity with the posterolateral putamen ($Q \times$ posterolateral putamen PPI). (**a**) PPI map in learners. (**b**) Difference map for the PPI between learners and nonlearners. (**c**) Path diagram of intrinsic connections between the putamen and the premotor cortex-SMA (SMA) via the thalamus supporting a full mediation effect in learners (path coefficients are shown [s.e. in parenthesis]; *a*: $P = 10^{-6}$, *b*: $P = 0.001$, putamen-SMA path *c*: $P = 0.04$, *ab*-controlled putamen-SMA path *c'*: $P = 0.8$, *ab*: $P = 4 \times 10^{-7}$, bootstrap test). C: caudate; M1: motor cortex; mdTh: mediodorsal thalamus; PO: parieto-occipital region; SMA: supplementary motor area; TP: temporal pole; vmPFC: ventromedial prefrontal cortex. Note that spatial smoothing with a Gaussian kernel (full-width-at-half-maximum = 2 mm) was applied to reduce edge effects for visualization purposes. Supporting information Fig. S5 presents whole-brain axial sections of the nonsmoothed difference map for PPI between learners and nonlearners.

presented as support for our interpretation of the group differences in $\overline{Q}$ signals in terms of learning, note that these exploratory brain-behavior analyses used a lenient statistical threshold and thus need to be interpreted with caution. Finally, the group comparison revealed that activation in limbic and paralimbic regions, including the hippocampus, decreased progressively in learners, an effect that was absent in nonlearners (Supporting Information Fig. S4).

Changes in Neural Connectivity Associated With Learning

Having determined that the posterolateral putamen tracks the Q value of entering lit arms in learners, we next examined changes in the functional connectivity of this region associated with learning. To do so, we computed the PPI [Friston et al., 1997] of the psychometric estimate Q by the physiological activation during choice in a seed point placed within the posterolateral putamen. By regressing this PPI term against whole-brain activation maps, we identified regions in which coupling with the posterolateral putamen changed as a function of Q (i.e., regions where functional connectivity with the posterolateral putamen changed with learning). Consistent with the organization of CBGTC loops [Draganski et al., 2008; Haber et al., 2000; Lehericy et al., 2004; Postuma and Dagher, 2006] and predictions from RL models, in learners, the premotor-motor, somatosensory, visual, and superior parietal cortices increased their connectivity with the posterolateral putamen as participants learned (Fig. 5a and Supporting Information Fig. S5). These learning-related changes in connectivity were specific to the posterolateral putamen, as they were not present for the ventral striatum (Supporting Information Fig. S6). Conversely, multiple regions within the limbic CBGTC loop—including the anteromedial ventral striatum, ventromedial prefrontal cortex, mediodorsal thalamus, and amygdala-hippocampus—as well as the midbrain decreased their connectivity with the posterolateral putamen with learning (Fig. 5a). As connections within the sensorimotor CBGTC loop strengthened with learning, the connections between the limbic and sensorimotor CBGTC loops weakened with learning. Indeed, whereas in learners the ventral anteromedial striatum was functionally connected with the posterolateral putamen in the first half of the experiment (mean $r = 0.615$, $P = 0.003$), such connectivity disappeared in the second half of the experiment ($P = 0.1$), although this difference was nonsignificant. Most of the changes in connectivity with the posterolateral putamen that we identified in learners differed significantly from those in nonlearners (Fig. 5b), suggesting that the changes in putaminal coupling in learners likely represented a learning-related process rather than nonspecific changes in connectivity.

Functional Anatomy of Corticostriatal Connections

Finally, we examined whether intrinsic connectivity between motor cortices and the putamen was mediated by the basal-ganglia output pathways through their influence on the thalamus, or whether these functional connections instead represented direct corticostriatal projections (Supporting Information). (We use the term "intrinsic connectivity" to refer to learning-unrelated connectivity throughout the task, as opposed to the learning-related

connectivity assessed in PPI analyses; see the Supporting Information.) A mediation analysis [Wager et al., 2009] based on the choice-related activations within the posterolateral putamen, premotor cortex-SMA, and lateral thalamus ROIs in learners showed that the thalamus fully mediated the functional connections between putamen and premotor cortex-SMA ($Z_{ab} = 4.61$, $P = 4 \times 10^{-7}$, bootstrap test; Fig. 5c). This finding, consistent with the CBGTC anatomy, suggests that response execution in this task depends on striato-thalamo-cortical pathways.

DISCUSSION

Using computational fMRI, we found that activation in the posterolateral putamen (along with other brain regions) tracks the strength of a putative S-R association on a trial-by-trial basis. Consistent with the known anatomy of the CBGTC sensorimotor loop [Draganski et al., 2008; Haber et al., 2000; Lehericy et al., 2004], we found that the posterolateral putamen is selectively connected to the premotor cortex-SMA, lateral thalamus, and sensory cortex. Critically, we showed for the first time that corticostriatal connections between sensorimotor areas and the posterolateral putamen (but not between sensorimotor areas and ventral striatum) strengthened as individuals learned the task based on reinforcement history, supporting a central assumption of RL models [Frank, 2005; Maia, 2009]. Finally, the thalamus mediated intrinsic connections between the posterolateral putamen and cortical motor areas, implicating the striato-thalamo-cortical pathway in response execution.

The fMRI paradigm used in the present study is directly analogous to the radial-arm maze task that was used originally to identify a specific role for the dorsolateral striatum (equivalent to the putamen in primates) in the formation of S-R associations in rodents [McDonald and Hong, 2004; Packard, 1999; Packard et al., 1989]. Our approach allowed us to uncover within-session changes in putaminal activation and connectivity among learners that occurred within a span of a few minutes. Despite the limitations discussed below, we suspect that our finding of progressive putaminal engagement reflects the formation of S-R associations rather than other types of associations. Previous human studies linked the posterolateral putamen [Knowlton et al., 1996] to habit formation, showing that putaminal activation at the onset of task blocks increased over the course of each training day and across days of training [Tricomi et al., 2009], and to valuation following extensive training [Wunderlich et al., 2012]. We instead focused on rapidly acquired associations that likely represent an early phase of habit learning. Our findings, together with those prior ones, suggest that the role of the posterior putamen begins at, but is not restricted to, the early phases of S-R learning.

We found that only learners modified their behavior to act optimally in response to lit arms and encoded the value (Q) of entering those arms in the putamen during learning. PE signals in the dorsal striatum have previously been shown to differ between those individuals who learn to select an optimal action versus those who do not [Schonberg et al., 2007]. Action selection, however, does not rely directly on PEs, but rather on value signals (see Supporting Information eq. 3). Our findings, therefore, more directly link learning to choose an action to the representations that are hypothesized to support such choice. More generally, our findings support the parallels between the dorsal striatum and the "actor" in the actor-critic model of action selection [Barto, 1995; O'Doherty et al., 2004]. This model proposes that the current state (i.e., the current situation or stimuli) is represented in cortex and that the basal ganglia implements two computational modules: the critic, which learns the values of states, and the actor, which learns and stores S-R associations [Barto, 1995; Maia, 2009; Maia and Frank, 2011]. Central to this model, the strength of an S-R association, which corresponds to the preference for (or value of) a given action in a given state, is assumed to be stored in the synaptic weights—the connections—between state units in the cortex and action units in the striatum, an assumption that finds strong support in our connectivity findings.

We also observed a progressive disengagement of the limbic circuit and a decoupling of this circuit from the sensorimotor circuit as individuals learned, compatible with a shift from an evaluative to a habitual mode of behavior. This interaction between corticostriatal circuits, possibly instantiated via the spiraling striato-nigro-striatal connections [Haber et al., 2000], might represent an active process driving the dynamic transition toward habitual responding. This shift in control from the limbic to the sensorimotor circuit, which we observed here at a relatively short timescale, is hypothesized to play a crucial role in the establishment of pathological habits, such as those involved in drug addiction, at longer timescales [Belin et al., 2009].

An alternative or complementary explanation for the progressive decoupling of limbic areas—particularly the ventromedial prefrontal cortex—from the putamen with learning is that, early in training, values might be communicated by the putamen to the ventromedial prefrontal cortex, which, possibly together with other areas in the limbic loop, might compare the values stored in the putamen-based habit system with those of a forward-planning, model-based system [Wunderlich et al., 2012]. With habit strengthening, the model-based system might tend to disengage, and values represented in the putamen might directly affect behavior, with less need for mediation by the ventromedial prefrontal cortex.

A somewhat surprising finding in our study was the relatively large percentage of subjects who failed to learn the task. Prior to performing this task, all participants performed a similar but spatially based version of the task [Marsh et al., 2010]. We suspect that participants who failed to learn the present version of the task persisted in using a spatial strategy that precluded them from adopting the S-R strategy required in the present version of the

task. Although we did not collect subjective reports regarding strategy use that would directly support this interpretation, the stronger deactivation of the medial temporal lobe in learners than in nonlearners during training (Supporting Information Fig. S4) and data suggesting that continued use of a spatial strategy impaired performance on the win-stay task (Supporting Information) do provide some support for this interpretation. Furthermore, the existence of a substantial percentage of both learners and nonlearners is actually an advantage of the present experiment, as it allowed us to more firmly establish the involvement of the patterns of activity and of changes in connectivity that we identified in learning.

Another limitation of our experimental design is that we did not include experimental manipulations such as outcome devaluation that would allow us to more firmly establish that participants' learning was based on S-R associations and not on alternative representations (e.g., situation-action-outcome or stimulus-outcome associations). The relevance of our findings remains even if one interprets them more broadly in terms of an unspecified RL mechanism rather than specifically in terms of S-R learning. Nonetheless, the known involvement of the posterolateral putamen and its homologue region in rodents (the dorsolateral striatum) in S-R learning [McDonald and Hong, 2004; Packard et al., 1989], as well as the remarkable fit between our findings and the predictions of "model-free RL" [Daw et al., 2005] models such as the actor-critic (which work by S-R learning), provide some support our interpretation in terms of S-R learning. Furthermore, early learning on the win-stay task in rats is insensitive to outcome devaluations, suggesting that performance on this task does rely on S-R learning (even from early stages of learning) [Sage and Knowlton, 2000]. Finally, independent manipulation of actions and rewards in a recent study revealed anticipatory signals in the putamen that represent action and not state value [Guitart-Masip et al., 2011]. Nonetheless, we acknowledge that our findings may represent forms of RL other than S-R learning, including stimulus-outcome and stimulus-response-outcome learning, a possibility that warrants further examination with additional experimental manipulations.

Our analyses are obviously limited by the information that can be obtained in fMRI. We interpreted changes in functional connectivity as a systems-level index of short-term synaptic plasticity, but we cannot exclude other explanations for such changes, such as the emergence of synchronous oscillatory activity. However, synaptic plasticity may control synchronous oscillatory activity [Paik and Glaser, 2010], suggesting that both phenomena may not even be independent. Lastly, our striatum-focused analyses of learning signals have limitations common to any ROI study. In particular, our results regarding PE and Q signals in the striatum do not imply that these signals are not also represented elsewhere in the brain. In fact, we observed that other brain regions also represented these learning signals (Supporting Information Figs. S2 and S3).

CONCLUSION

In summary, our findings suggest a direct link between strengthening of corticostriatal connections within the sensorimotor circuit and learning of S-R associations in humans. In addition, our results suggest that habit learning involves a disengagement of the limbic loop and a reduction in its influence over the sensorimotor loop, with control transitioning to the latter. This transition had previously been hypothesized to underlie the formation of pathological habits [Belin et al., 2009], which brings substantial translational potential to this work.

REFERENCES

Balleine BW, O'Doherty JP (2010): Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology 35: 48–69.

Barto AG (1995): Adaptive critics and the basal ganglia. In: Houk JC, Davis JL, Beiser DG, editors. Models of Information Processing in the Basal Ganglia. Cambridge, MA: MIT Press.

Belin D, Jonkman S, Dickinson A, Robbins TW, Everitt BJ (2009): Parallel and interactive learning processes within the basal ganglia: Relevance for the understanding of addiction. Behav Brain Res 199:89–102.

Centonze D, Picconi B, Gubellini P, Bernardi G, Calabresi P (2001): Dopaminergic control of synaptic plasticity in the dorsal striatum. Eur J Neurosci 13:1071–1077.

Charpier S, Deniau JM (1997): In vivo activity-dependent plasticity at cortico-striatal connections: Evidence for physiological long-term potentiation. Proc Natl Acad Sci USA 94: 7036–7040.

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

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011): Model-based influences on humans' choices and striatal prediction errors. Neuron 69:1204–1215.

de Wit S, Watson P, Harsay HA, Cohen MX, van de Vijver I, Ridderinkhof KR (2012): Corticostriatal connectivity underlies individual differences in the balance between habitual and goal-directed action control. J Neurosci 32:12066–12075.

den Ouden HE, Friston KJ, Daw ND, McIntosh AR, Stephan KE (2009): A dual role for prediction error in associative learning. Cereb Cortex 19:1175–1185.

den Ouden HE, Daunizeau J, Roiser J, Friston KJ, Stephan KE (2010): Striatal prediction error modulates cortical coupling. J Neurosci 30:3210–3219.

Draganski B, Kherif F, Kloppel S, Cook PA, Alexander DC, Parker GJ, Deichmann R, Ashburner J, Frackowiak RS (2008): Evidence for segregated and integrative connectivity patterns in the human Basal Ganglia. J Neurosci 28:7143–7152.

Frank MJ (2005): Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. J Cogn Neurosci 17:51–72.

Frank MJ, Moustafa AA, Haughey HM, Curran T, Hutchison KE (2007): Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc Natl Acad Sci USA 104:16311–16316.

Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ (1997): Psychophysiological and modulatory interactions in neuroimaging. NeuroImage 6:218–229.

Friston KJ, Josephs O, Rees G, Turner R (1998): Nonlinear event-related responses in fMRI. Magn Reson Med 39:41–52.

Glimcher PW (2011): Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. Proc Natl Acad Sci USA 108(Suppl 3):15647–15654.

Guitart-Masip M, Fuentemilla L, Bach DR, Huys QJ, Dayan P, Dolan RJ, Duzel E (2011): Action dominates valence in anticipatory representations in the human striatum and dopaminergic midbrain. J Neurosci 31:7867–7875.

Haber SN, Fudge JL, McFarland NR (2000): Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. J Neurosci 20:2369–2382.

Klein TA, Neumann J, Reuter M, Hennig J, von Cramon DY, Ullsperger M (2007): Genetically determined differences in learning from errors. Science 318:1642–1645.

Knowlton BJ, Mangels JA, Squire LR (1996): A neostriatal habit learning system in humans. Science 273:1399–1402.

Lehericy S, Ducros M, Van de Moortele PF, Francois C, Thivard L, Poupon C, Swindale N, Ugurbil K, Kim DS (2004): Diffusion tensor fiber tracking shows distinct corticostriatal circuits in humans. Ann Neurol 55:522–529.

Maia TV (2009): Reinforcement learning, conditioning, and the brain: Successes and challenges. Cogn Affect Behav Neurosci 9:343–364.

Maia TV, Frank MJ (2011): From reinforcement learning models to psychiatric and neurological disorders. Nat Neurosci 14:154–162.

Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH (2003): An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. NeuroImage 19:1233–1239.

Marsh R, Hao X, Xu D, Wang Z, Duan Y, Liu J, Kangarlu A, Martinez D, Garcia F, Tau GZ, Yu S, Packard MG, Peterson BS (2010): A virtual reality-based FMRI study of reward-based spatial learning. Neuropsychologia 48:2912–2921.

McDonald RJ, Hong NS (2004): A dissociation of dorso-lateral striatum and amygdala function on the same stimulus-response habit task. Neuroscience 124:507–513.

McDonald RJ, White NM (1993): A triple dissociation of memory systems: Hippocampus, amygdala, and dorsal striatum. Behav Neurosci 107:3–22.

Neumann J, Lohmann G (2003): Bayesian second-level analysis of functional magnetic resonance images. NeuroImage 20:1346–1355.

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, Dayan P, Friston K, Critchley H, Dolan RJ (2003): Temporal difference models and reward-related learning in the human brain. Neuron 38:329–337.

O'Doherty JP, Hampton A, Kim H (2007): Model-based fMRI and its application to reward learning and decision making. Ann N Y Acad Sci 1104:35–53.

Packard MG (1999): Glutamate infused posttraining into the hippocampus or caudate-putamen differentially strengthens place and response learning. Proc Natl Acad Sci USA 96:12881–12886.

Packard MG, Knowlton BJ (2002): Learning and memory functions of the Basal Ganglia. Annu Rev Neurosci 25:563–593.

Packard MG, Hirsh R, White NM (1989): Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: Evidence for multiple memory systems. J Neurosci 9:1465–1472.

Paik SB, Glaser DA (2010): Synaptic plasticity controls sensory responses through frequency-dependent gamma oscillation resonance. PLoS Comput Biol 6:e1000927.

Pawlak V, Kerr JN (2008): Dopamine receptor activation is required for corticostriatal spike-timing-dependent plasticity. J Neurosci 28:2435–2446.

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.

Postuma RB, Dagher A (2006): Basal ganglia functional connectivity based on a meta-analysis of 126 positron emission tomography and functional magnetic resonance imaging publications. Cereb Cortex 16:1508–1521.

Sage JR, Knowlton BJ (2000): Effects of US devaluation on win-stay and win-shift radial maze performance in rats. Behav Neurosci 114:295–306.

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

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

Tricomi E, Balleine BW, O'Doherty JP (2009): A specific role for posterior dorsolateral striatum in human habit learning. Eur J Neurosci 29:2225–2232.

van den Bos W, Cohen MX, Kahnt T, Crone EA (2012): Striatum-medial prefrontal cortex connectivity predicts developmental changes in reinforcement learning. Cereb Cortex 22:1247–1255.

Wager TD, Waugh CE, Lindquist M, Noll DC, Fredrickson BL, Taylor SF (2009): Brain mediators of cardiovascular responses to social threat: Part I: Reciprocal dorsal and ventral sub-regions of the medial prefrontal cortex and heart-rate reactivity. NeuroImage 47:821–835.

Watkins CJCH, Dayan P (1992): Q-Learning. Mach Learn 8:279–292.

Wimmer GE, Daw ND, Shohamy D (2012): Generalization of value in reinforcement learning by humans. Eur J Neurosci 35: 1092–1104.

Wunderlich K, Dayan P, Dolan RJ (2012): Mapping value based planning and extensively trained choice in the human brain. Nat Neurosci 15:786–791.

Xu T, Yu X, Perlik AJ, Tobin WF, Zweig JA, Tennant K, Jones T, Zuo Y (2009): Rapid formation and selective stabilization of synapses for enduring motor memories. Nature 462:915–919.

Yin HH, Knowlton BJ (2006): The role of the basal ganglia in habit formation. Nat Rev Neurosci 7:464–476.