Published in final edited form as:
J Cogn Neurosci. 2012 January ; 24(1): 106–118. doi:10.1162/jocn_a_00114.

Human Dorsal Striatum Encodes Prediction Errors during Observational Learning of Instrumental Actions

Jeffrey C. Cooper[1,2], **Simon Dunne**[1], **Teresa Furey**[1], and **John P. O'Doherty**[1,2]
[1]Trinity College Dublin

[2]California Institute of Technology

Abstract

The dorsal striatum plays a key role in the learning and expression of instrumental reward associations that are acquired through direct experience. However, not all learning about instrumental actions require direct experience. Instead, humans and other animals are also capable of acquiring instrumental actions by observing the experiences of others. In this study, we investigated the extent to which human dorsal striatum is involved in observational as well as experiential instrumental reward learning. Human participants were scanned with fMRI while they observed a confederate over a live video performing an instrumental conditioning task to obtain liquid juice rewards. Participants also performed a similar instrumental task for their own rewards. Using a computational model-based analysis, we found reward prediction errors in the dorsal striatum not only during the experiential learning condition but also during observational learning. These results suggest a key role for the dorsal striatum in learning instrumental associations, even when those associations are acquired purely by observing others.

INTRODUCTION

Much is known about the neural underpinnings of how associations between stimuli, actions, and rewards are learned through experience and about how those associations guide choices (Montague, King-Casas, & Cohen, 2006; Dayan & Balleine, 2002). Humans and other animals can learn not only by direct experience but also through observing others' experiences (Heyes & Dawson, 1990; Bandura, 1977; Myers, 1970). In spite of the ubiquitous nature of observational learning, its neural substrates remain poorly understood.

Considerable progress has been made in elucidating candidate neural substrates for experiential reward learning. The finding that phasic activity in dopamine neurons resembles a reward prediction error signal from computational models of reinforcement learning (RL) has led to the proposal that these neurons underlie experiential reward learning in dopamine target regions, particularly the striatum (Daw & Doya, 2006; O' Doherty, 2004; Schultz, Dayan, & Montague, 1997). Consistent with this proposal, many human neuroimaging studies have reported activity in the striatum during learning, likely reflecting (at least in part) dopaminergic input (Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006; McClure, Berns, & Montague, 2003; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003). Moreover, different regions of striatum are differentially correlated with reward prediction errors, depending on the nature of the learned association. Although ventral striatum appears to correlate with prediction errors during both Pavlovian and instrumental learning, the

Reprint requests should be sent to Jeffrey C. Cooper, Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125, or via jcooper@caltech.edu..

Europe PMC Funders Author Manuscripts

dorsal striatum appears to be specifically engaged when participants must learn associations between instrumental actions and rewards (Delgado, 2007; O'Doherty et al., 2004; Tricomi, Delgado, & Fiez, 2004). These findings have been interpreted in the context of actor/critic models of RL, whereby a ventral striatal "critic" is hypothesized to mediate learning of stimulus–reward associations, whereas a dorsal striatal "actor" is hypothesized to support learning of instrumental action values (Suri & Schultz, 1999; Sutton & Barto, 1998).

In spite of the large body of research on the dorsal striatum in experiential instrumental reward learning, very little is known about this region's function in observational reward learning. One recent study reported activity in ventral striatum and ventromedial pFC during observational learning (Burke, Tobler, Baddeley, & Schultz, 2010) but did not find dorsal striatal activation. The goal of the present study was to address the role of the dorsal striatum in mediating learning of associations between instrumental actions and rewards when those associations are acquired through observation and to compare and contrast prediction error activity in the dorsal striatum during observational and experiential instrumental learning.

To investigate these questions, we scanned participants using fMRI while they underwent both observational and experiential instrumental conditioning for liquid juice rewards. We used a computational model-based analysis to test for brain regions correlating with reward prediction errors in both observational and experiential cases. To isolate prediction errors specifically related to action learning, we included matched noninstrumental control conditions in which the participant faced reward cues that did not require action selection. We hypothesized that dorsal striatal activity would be correlated with prediction errors during both observational and experiential instrumental conditioning and that the dorsal striatum would be more engaged during observational instrumental conditioning than during noninstrumental conditioning.

METHODS

Participants

Nineteen healthy volunteers from the Trinity College Dublin student population participated; one was excluded for excessive head motion (>6 mm in one run), and two were excluded for showing no evidence of learning, leaving 16 participants for analysis (10 women, 6 men; $M = 22.19$ years). One participant's postexperiment ratings were lost because of technical error, and a different participant's RT data were excluded for being more than 3 *SD* below the grand mean. All were free of current psychiatric diagnoses. Participants were asked to refrain from eating or drinking anything but water for 6 hr before the scan to ensure that they were motivated by the liquid rewards. All participants gave informed consent, and the study was approved by the research ethics committee of the Trinity College School of Psychology.

Task Overview

On each trial, participants encountered a single two-armed slot machine cue, with the trial condition signaled by color and position. There were six randomly intermixed reward cue conditions: two Experienced (instrumental and noninstrumental), two Observed (instrumental and noninstrumental), and two Test (instrumental and noninstrumental). Briefly, on Experienced trials, participants made their own responses and received their own rewards, whereas on Observed trials, participants observed a confederate outside the scanner performing an identical task. On Test trials, participants made responses for the confederate's cues without receiving any outcomes. To enhance the salience of the confederate's live presence and the visual similarity between trials, all trials were shown on a single monitor in the control room outside the scanner, at which the confederate was

seated and to which both the participant's and confederate's response boxes were connected (see Figure 1A for schematic). The participant viewed this monitor and part of the confederate's response box over a live video feed (see Figure 1B for schematic) and was instructed that he or she would be playing on the same computer and responding to the same monitor as the confederate.

Experienced Trials

On Experienced trials, participants were shown a slot machine cue on "their" half of the screen (either top or bottom, counterbalanced by participant), indicating that they were to make a response and receive an outcome on that trial. Two kinds of Experienced trials were presented (see Figure 1B for trial time course). On Experienced instrumental trials, the slot machine was shown with two arms, and participants had 2 sec to select either the left or right arm. Selection was shown on-screen by depressing the selected arm. Following a variable length delay (1 sec followed by a delay randomly drawn from a truncated Poisson distribution; total delay mean = 2.5 sec, range = 1–7 sec) intended to allow separate estimation of the neural response to cues and outcomes, a liquid reward or neutral outcome was delivered by plastic tubing controlled by electronic syringe pump into the participant's mouth during the outcome phase (2 sec). Rewards were 1 ml of a sweet blackcurrant juice (Ribena; GlaxoSmithKline, London, United Kingdom). Neutral outcomes were 1 ml of an affectively neutral tasteless solution consisting of the main ionic components of saliva (25 mM KCl and 2.5 mM $NaHCO_3$), delivered by a separate tube. The two arms had independent probabilities of being rewarded on each trial, each of which varied over the experiment to ensure learning continued throughout the task (see Figure 1C for sample curves). Each arm's reward probability was a sine curve set to drift between 0 and 100% reward probability (with random starting point and half-period randomly set between 0.87 and 1.67 times the number of trials per condition and the machine's arms constrained to be correlated with each other at less than $r = 0.02$) plus a small amount of Gaussian noise on each trial ($M = 0$, $SD = 6\%$; noise was added before scaling sines to have a range of 0–100%). Outcome delivery was also immediately followed by a visual signal for reward (green square) or neutral outcomes (gray square). These signals ensured visual equivalence with Observed trials, where visual cues were required to enable observational learning (see below). The outcome phase was followed by a variable-length intertrial interval (1 sec followed by a delay randomly drawn from a truncated Poisson distribution, total delay mean = 2.5 sec, range = 1–7 sec), for a mean total trial length of 9 sec. Failure to respond resulted in a neutral outcome; missed responses (2.7% of trials) were omitted from analysis.

On Experienced noninstrumental trials, which controlled for reward learning without action selection, a differently colored slot machine cue was presented without arms. After 500 msec, a single indicator was shown on either the left or right sides, and participants had 1.5 sec to indicate, with a button press, which side was selected. Participants, thus, did not select an action but instead passively followed an action selection made by the computer. Trials afterwards followed an identical structure to instrumental trials, including separate independent reward probabilities for each side of the noninstrumental machine, constructed with an identical algorithm. The computer selected the side with higher probability of reward 70% of the time (randomly distributed across trials).

Both Experienced instrumental and noninstrumental trials also had a "neutral" control machine intended to control for visuomotor confounds. Both neutral control machines were identical to their respective reward cue machine but were gray in color, and each was always followed by a neutral outcome regardless of the arm or side selected.

Europe PMC Funders Author Manuscripts

Observed Trials

Observed trials were designed to mirror Experienced trials in every way, except instead of the participant making choices and/or responses, the participant observed a confederate outside the scanner perform similar trials. On Observed trials, participants were shown a slot machine on the other half of the screen from the Experienced trials (top or bottom), with a different set of colors from the Experienced machines. Participants never received an outcome themselves on Observed trials. Instead, participants were instructed that the confederate would make responses and receive outcomes on the Observed trials. This single-monitor split-screen design provided an important control by ensuring that Observed trials and Experienced trials were visually as similar as possible.

Two kinds of Observed trials were shown, instrumental and noninstrumental. Both were identical to the respective Experienced trials in their timing, contingencies, and reward probabilities (with independently constructed reward probability curves for each arm of each machine). Participants could thus observe on-screen which arm was chosen or which side was selected and could see whether a reward or neutral outcome was given by the visual signals presented at outcome delivery. Unknown to the participant, the confederate's button box and outcome tubes were actually not connected; instead, the confederate's choices were made by selecting the arm with higher reward probability 70% of the time to approximate average participant likelihood of selecting the arm with higher reward probability in Experienced instrumental trials (based on pilot data). The camera angle was set to obscure the confederate's physical button presses. Participants were instructed to observe the confederate's choices and do their best to learn which arm and which side were rewarded most often.

As with Experienced trials, two Observed neutral control machines were also presented (instrumental and noninstrumental); both were identical to their respective reward cue machine but gray in color (and displayed on the confederate's half of the screen), and both were always followed by a neutral outcome for the confederate regardless of which arm or side was selected.

Test Trials

Test trials gave participants an opportunity to demonstrate their learning for the Observed trials. Two kinds of Test trials were presented, instrumental and noninstrumental. On Test instrumental trials, participants were shown the confederate's Observed instrumental slot machine on their own half of the screen and had 2 sec to select either the left or right arm. Similarly, on Test noninstrumental trials, participants were shown the confederate's Observed noninstrumental slot machine on their own half of the screen, then had a side indicated after 500 msec, and had 1.5 sec to indicate which side was selected. The purpose of these trials was to gain a behavioral measure of the participant's observational learning and provide an incentive for the participant to pay attention to Observed trials. If participants successfully learned through observation, they should have favored arms most recently associated with reinforcement for the confederate. To prevent participants learning about Observed trials through direct experience on Test trials, no outcomes were presented on these trials during the experiment, and Test trials ended after the response (followed by intertrial interval). Instead, participants were told rewarded outcomes on these trials would be put in a "bank," and that "banked" rewards would be multiplied in volume by 10 and given to participants after the scan. Neutral control machines were never shown on Test trials.

Task Procedure

There were 72 trials each of the Experienced instrumental, Experienced noninstrumental, Observed instrumental, and Observed noninstrumental conditions. Within each condition, 48 trials were with reward cue machines and 24 were with neutral control machines. In addition, there were 24 trials of each of the Test instrumental and Test noninstrumental conditions. Each condition's trials were divided equally between four scanner runs (so that each run had 18 Experienced instrumental trials, 6 Test instrumental trials, etc.), and trial types were intermixed randomly throughout each run. Participant screen halves (top or bottom) and machine colors were counterbalanced across participants.

Before the task, participants underwent 10 min of training outside the scanner with no outcomes to ensure they understood each slot machine and the outcomes, including the Test-trial "bank." They also met the confederate and observed the pumps and camera setup before entering the scanner room. After the task, participants made a series of ratings of each machine outside the scanner on 9-point Likert scales, including how much they liked each machine and each outcome.

Scanning

Participants were scanned with a Phillips 3 T MRI scanner using the standard head coil, padded to minimize head motion. Functional images covered the whole brain with 38 contiguous 3.2-mm thick axial slices with gradient-echo T2*-weighted echoplanar imaging (repetition time = 2 sec, echo time = 28 msec, in-plane voxel size = 3 × 3 mm, matrix = 80 × 80). The acquisition plane was tilted about 30° to the anterior–posterior commissure plane to optimize sensitivity in the ventral pFC (Deichmann, Gottfried, Hutton, & Turner, 2003). Each participant's scan consisted of four functional runs of 363 images each; the first four of each run were discarded to account for magnetic equilibration. A high-resolution structural image was also acquired before the task (3-D acquisition; T1-weighted spoiled gradient sequence; voxel size = 0.9 × 0.9 × 0.9 mm, matrix = 256 × 256 × 180).

Stimuli were presented using Cogent 2000 (Wellcome Trust Centre for Neuroimaging, London). To account for swallowing-related motion, pressure data from a sensor attached to the throat was recorded with BioPac hardware and AcqKnowledge software (MP-150 system and TSD160A transducer; BioPac Systems, Inc., Goleta, CA).

Statistical Analysis

fMRI images were analyzed with SPM8 (Wellcome Department of Imaging Neuroscience, London) and MATLAB (The Mathworks, Inc., Natick, MA). Functional images were preprocessed with standard parameters, including slice timing correction (to the center slice), realignment (to each participant's first image), coregistration of the in-plane anatomical image, normalization (to the International Consortium for Brain Mapping (ICBM)/MNI152 template with parameters estimated from each participant's coregistered in-plane image, using SPM8 default normalization parameters), and spatial smoothing (with a 4-mm FWHM Gaussian kernel).

Prediction errors were estimated with a hybrid RL model. Instrumental trials used a SARSA learning model (Sutton & Barto, 1998). The value of the chosen action was updated on each trial according to the rule:

$$\begin{aligned} V_{\mathrm{a}}^{t+1} &= V_{\mathrm{a}}^{t} + \alpha\delta^{t} \\ \delta^{t} &= (O^{t} - V_{\mathrm{a}}^{t}) \end{aligned}$$

where δ^t is that trial's prediction error, O^t is that trial's outcome (set to 1 for *reward* and 0 for *neutral*), and V_a is the value for the chosen action. The probability of selecting left or right on each trial was given by a softmax function of the difference between V_{left} and V_{right}. For Observed cues, values, and prediction errors were based on the observed confederate outcomes. Initial values were set to 0 for all actions. A single pooled learning rate for the group on Experienced instrumental trials was estimated by fitting the parameters α and β (the logistic slope or stochasticity parameter) to participant choices with maximum likelihood estimation; estimating a fixed rate tends to produce better estimates of neural activity due to increased regularization (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006). A separate learning rate for Observed instrumental trials was estimated by fitting the model to participant choices on Test trials, using values based on the observed outcomes. Fitted learning rates are comparable with earlier studies using similar designs (Experienced learning rate = 0.38, Observed learning rate = 0.26; Gläscher, Daw, Dayan, & O'Doherty, 2010; Schönberg et al., 2010; Valentin & O'Doherty, 2009); these parameters are, however, likely to be highly task and domain dependent.

For noninstrumental trials, we used a Rescorla–Wagner learning model (Rescorla & Wagner, 1972), which has an identical learning rule to that used in the RL model described above, but without an action selection component. The model was fit using trial-by-trial RTs (log-transformed) as a measure of conditioning (Bray & O'Doherty, 2007), minimizing the difference between the chosen side's value on each trial and that trial's RT (inverted so that high value corresponded to low RT). This model, therefore, captured within-condition variation in RT, predicting faster RTs for trials when that condition's value was higher. Values and prediction errors for Observed trials again reflected the observed confederate outcomes. As in instrumental trials, learning rates for the group were estimated separately for Experienced and Observed trials, using Experienced and Test RTs. Learning rates were again comparable to earlier studies (Experienced = 0.18, Observed = 0.53); however, because the model fit to Observed noninstrumental trials was nonsignificant (see Results), that condition's rate should be interpreted with caution. Higher learning rates also imply faster forgetting, and thus, the high Observed noninstrumental learning rate may reflect a greater reliance on the recent past for these trials (full RT distributions are provided in Supplementary Figure S1, available at www.odohertylab.org/supplementary/Cooperetal_FigureS1.html.)

A general linear model was created for each participant to estimate experimental effects. The model included delta function regressors for the cue period (3-sec duration) and outcome period (0-sec duration) for each of the four key conditions: Observed/Experienced × Instrumental/Noninstrumental. Both reward cue and neutral control trials were included in each regressor. Two delta function regressors also modeled the Test instrumental and Test noninstrumental cues. Each cue regressor had a parametric modulator for the trial-specific chosen value, set to 0 for neutral control trials; as our focus was on learning from outcomes, cues were not analyzed here. Each outcome regressor had a parametric modulator for the trial-specific estimated prediction error, set to 0 for neutral control trials. Because neutral control machines were visually similar and had identical motor requirements but led to no learning, this provided an implicit baseline for the prediction error regressors. All regressors of interest were convolved with the SPM8 canonical hemodynamic response function. Regressors of no interest included six for estimated head motion parameters, one indicating scans with greater than 2-mm head motion in any direction, one for throat pressure to account for swallowing, and a constant term for each session.

General linear models included an AR(1) model for temporal autocorrelation and were estimated using maximum likelihood (with autocorrelation hyperparameter estimation restricted to voxels passing a global F-threshold using SPM8 defaults). A high-pass filter

(cutoff, 128 sec) removed low-frequency noise. Beta-weight images for each regressor were combined to form appropriate contrasts within participants, and contrast images were carried forward to group level analyses. Significant effects were tested with one-sample *t* tests across the group.

We examined activations within separate, a priori, dorsal, and ventral ROIs in the striatum. First, an overall striatal region was hand-drawn on the average group anatomical; next, it was divided at the Montreal Neurological Institute (MNI) coordinate $z = 0$ into a bilateral dorsal region encompassing dorsal caudate and putamen and a bilateral ventral region encompassing ventral putamen, ventral anterior caudate, and nucleus accumbens. Activations were thresholded voxelwise at $p < .005$, and significant clusters were identified in each ROI using an extent threshold estimated by Gaussian random field theory to correct for multiple comparisons within each ROI's volume (Worsley et al., 1996). To examine whole-brain activations, we used a voxelwise threshold of $p < .001$ and extent threshold estimated by Gaussian random field theory for each contrast to correct for multiple comparisons across the whole brain. Peaks are reported in ICBM/MNI coordinates.

To examine beta weights across specific clusters, we used leave-one-out extraction to provide an independent criterion for voxel selection (Kriegeskorte, Simmons, Bellgowan, & Baker, 2009): for each participant, beta weights were extracted from significant voxels for that cluster in a group model excluding that participant. For Bayesian model comparison in a cluster, we used SPM8's Bayesian model selection algorithm with beta weights across participants as input (Stephan, Penny, Daunizeau, Moran, & Friston, 2009). We again used a leave-one-out procedure to extract beta weights for the prediction error regressor from the main general linear model (averaged across the cluster within participant). We compared these to beta weights from the same voxels and same regressor in a baseline general linear model that was identical, except every chosen value was set to 0.5 (or 0 for neutral control trials) and every prediction error was set to (outcome – 0.5) (or 0 for neutral control trials). This baseline model provided a conservative test that the signals we observed fit better to an RL model prediction error than any signal related to the outcome itself.

RESULTS

Behavior

Participants rated how much they liked the reward and neutral outcomes following the scan to confirm the motivational value of the rewards. Participants rated the reward outcomes significantly higher than the neutral outcomes, with a mean rating of 6.93 ($SEM = 0.32$) compared with 3.80 for the neutral ($SEM = 0.55$; $t(14) = 5.29$, $p < .001$). The ratings suggest participants still found the reward outcomes valuable through the end of the experiment.

To assess how participants learned from their rewards over time, we examined whether their instrumental choices could be described accurately by a RL model (see Methods for model details and Figure 2B for example of model fits). We fit the model separately to Experienced and Test choices and tested it against several alternatives (Figure 2A). For both Experienced and Test choices, the model fit better than either a baseline model that predicted equal action probabilities on every trial (RL model: Experienced-trial Bayesian Information Criterion (BIC) = 968.26, Test-trial BIC = 508.58; baseline model: Experienced-trial BIC = 1039.72, Test-trial BIC = 512.93) or a saturated model that assigned a separate parameter to every choice (saturated model: Experienced-trial BIC = 4965.06, Test-trial BIC = 2188.0).

For Test trials only, we also compared the RL model to two simple alternatives that would require no learning: one in which participants simply imitated the confederate's last action ("last-action"), and one in which participants remembered only the confederate's last

action–outcome pairing, choosing the same action if the confederate won or the other action if the confederate lost ("last-outcome"). Both alternatives used one fewer parameter than the RL model (eliminating the learning rate while retaining a fitted choice stochasticity). Even after adjusting for the number of free parameters in the models, the RL model fit better than either alternative (last-action model: Testtrial BIC = 513.11; last-outcome model: Test-trial BIC = 511.07). The model fits confirmed that the RL model accounted for instrumental learning whether the outcomes were delivered to the participant or to the confederate and suggested that participants learned about the confederate's outcomes over time rather than adopting a simple imitation strategy.

For noninstrumental trials, we evaluated the RL model fit by examining whether the model explained a significant amount of variance in RTs (after log-transforming RTs to account for their skew); this is equivalent to testing the RL model against a baseline model predicting equal values on every trial (as the baseline would only predict the mean RT). For Experienced noninstrumental trials, the RL model fit was significant ($F(1, 678) = 9.1814$, $p < .005$), confirming that participants learned conditioned responses to the Experienced cues. For Test noninstrumental trials, the model fit was not significant ($F(1, 321) = 0.18$, *ns*), suggesting that participants did not develop strong conditioned responses for Observed noninstrumental cues.

Finally, participants also rated how much they liked each cue, as a self-reported measure of conditioned preference (Figure 2C). Repeated measures ANOVA confirmed that liking differed significantly across cue categories ($F(7, 98) = 15.25$, $p < .001$). In particular, planned comparisons between each reward cue machine and its corresponding neutral control machine found that participants rated the reward cue machine higher for Experienced instrumental ($t(14) = 5.56$, $p < .001$), Experienced noninstrumental ($t(14) = 6.61$, $p < .001$), and Observed instrumental trials ($t(14) = 2.88$, $p = .012$). The comparison for Observed noninstrumental trials was not significant ($t(14) = 0.12$, *ns*). The ratings suggest that for Observed instrumental trials, participants developed conditioned preferences for the Observed cues, although they were never rewarded on them.

Average RTs differed across cue categories (Table 1 and Supplementary Figure S1; $F(5, 2655) = 140.40$, $p < .001$); the key difference was between instrumental and noninstrumental trials, with all noninstrumental trial averages significantly faster than all instrumental trial averages. Experienced instrumental reward cue trials were also significantly faster than Test instrumental trials or Experienced instrumental neutral control trials.

Imaging

Observational Instrumental Prediction Errors—To identify regions underlying prediction errors for observational instrumental learning, we examined activation correlating with the prediction error regressor for Observed instrumental trials (Figure 3A). In the dorsal striatum, a significant cluster in right dorsal caudate was positively correlated with prediction errors ($x, y, z = 21, 17, 7$; $Z = 3.94$; extent = 20 voxels, $p < .05$ corrected for dorsal striatal volume; see also Figure 3C). A smaller cluster in the left dorsal caudate was also detected, although this did not reach the extent threshold for significance ($x, y, z = -15, 14, 7$; $Z = 2.93$; extent = 5 voxels). By contrast, no significant positive clusters were present in the ventral striatum. No significant clusters in either dorsal or ventral striatum were negatively correlated with prediction errors. At the whole-brain threshold, no significant regions were positively or negatively correlated with prediction errors.

In a post hoc analysis, to directly examine the selectivity of the dorsal caudate for observational instrumental prediction errors, we extracted average beta weights for each trial type from the significant right caudate cluster (using a leave-one-out extraction procedure to

avoid nonindependence issues [Kriegeskorte et al., 2009]; see Methods). A 2 × 2 repeated measures ANOVA revealed a main effect for instrumental trials greater than noninstrumental trials ($F(1, 63) = 12.65$, $p < .005$), but no other main effect or interaction (Figure 3B). This cluster was significantly more active for prediction errors in Observed instrumental trials than baseline ($t(15) = 3.29$, $p < .01$); no other condition's activation, however (including Experienced instrumental trials), was significantly different from baseline (all $ts < 1.71$, *ns*). This region of dorsal caudate, then, was selectively active for instrumental as compared with noninstrumental Observed prediction errors but relatively similarly active for Experienced and Observed instrumental prediction errors.

To examine whether activation in this cluster truly reflected reinforcement learning prediction errors, we used Bayesian model comparison within the significant right caudate cluster to compare whether its activation in each condition was better fit by prediction errors or by a baseline model that responded to outcome valence but did not learn over time (see Methods). For instrumental Observed trials, the exceedance probability of the prediction error model compared with baseline was 99.67%; that is, activation in this region was 99.67% likely to have been produced by prediction errors from the RL model as opposed to a static response to instrumental Observed outcomes over time.

To examine whether learning performance related to dorsal caudate activation, we examined learning performance (i.e., RL model fit, indexed by each participant's model log-likelihood for Test instrumental trials) in a simple Condition × Performance ANCOVA; this analysis revealed no main effect of Learning Performance ($F(1, 56) = 0.1$, *ns*) and no interaction with Condition ($F(3, 56) = 1.51$, *ns*). This finding, however, should be qualified by the relatively low variation in learning performance (after screening out nonlearners).

Experiential Instrumental Prediction Errors—For comparison, we next examined activation for prediction errors in the Experienced instrumental trials (Figure 4A). In the ventral striatum, a significant cluster in right nucleus accumbens was positively correlated with prediction errors ($x, y, z = 9, 11, -4$; $Z = 3.19$; extent = 16 voxels, $p < .05$ corrected for striatal volume). In the dorsal striatum, a cluster near the corrected threshold in left dorsal putamen was also positively correlated with prediction errors ($x, y, z = -24, -1, 10$; $Z = 3.87$; extent = 14 voxels, $p = .055$ corrected for striatal volume). This cluster's location was similar to that in a recent study of instrumental prediction error activation (Schönberg et al., 2010) and met a small-volume corrected cluster threshold centered on that study's coordinates (extent = 8 voxels within 8-mm-radius sphere centered at −27, 6, 9, $p < .05$ corrected for small volume). By contrast to the dorsal caudate cluster from the Observed instrumental trials, a 2 × 2 ANOVA on the leave-one-out extracted beta weights from the dorsal putamen cluster revealed a main effect of Experienced greater than Observed trials (Figure 4B; $F(1, 63) = 11.64$, $p < .005$), but no other main effect or interaction. This cluster's activation for prediction errors in Experienced instrumental trials was significantly greater than baseline ($t(15) = 2.92$, $p < .05$); no other condition's activation (including Experienced noninstrumental trials) was significant different from baseline (all $ts < 1.77$, *ns*). Together, the dorsal and ventral clusters are consistent with earlier studies that find prediction error activation in both dorsal and ventral striatum for experiential prediction errors (O'Doherty et al., 2004).

No significant clusters in dorsal or ventral striatum were negatively correlated with prediction errors. At the whole-brain threshold, no significant regions were positively or negatively correlated with prediction errors; in addition, in direct whole-brain *t* tests between Experienced and Observed instrumental prediction error regressors, no clusters were significantly different between conditions. Finally, a conjunction analysis over prediction

error activations for Experienced and Observed instrumental trials revealed no significant overlap between activated regions for the two trial types.

Noninstrumental Prediction Errors—To examine whether prediction error activations were specific to instrumental learning, we also examined activation for prediction errors in Observed and Experienced noninstrumental trials. For Observed noninstrumental trials, no significant clusters were positively or negatively correlated with prediction errors in either dorsal or ventral striatum or at the whole-brain threshold. For Experienced noninstrumental prediction errors, by contrast, a significant cluster in left ventral putamen was positively correlated with prediction errors (Figure 5; $x, y, z = -21, 5, -4$; $Z = 3.24$; extent = 13 voxels, $p < .05$ corrected for ventral striatal volume). No significant clusters were present in dorsal striatum, and no significant clusters were negatively correlated in either striatal ROI. At the whole-brain threshold, no regions were positively or negatively correlated nor were any significantly different in direct tests between Experienced and Observed noninstrumental prediction errors.

DISCUSSION

Prediction error signals have been reported throughout the human striatum during experiential reward learning, with the dorsal striatum playing a specific role during instrumental experiential learning. In the present study, we extend these findings by showing that activation in parts of the human dorsal striatum correlates with reward prediction errors even when those error signals are computed while observing another person perform an instrumental learning task. A Bayesian model comparison suggested these signals were better explained by prediction errors than merely observed outcomes. The dorsal striatum is, thus, involved in encoding prediction errors during learning of instrumental associations whether those associations are acquired through experience or purely through observation. This involvement is selective for prediction errors during instrumental learning, as shown by comparison with a noninstrumental reward learning control. Observed instrumental and noninstrumental trials both required only passive observation by the participant, yet only the instrumental condition recruited the dorsal striatum.

These findings provide new insights into the functions of dorsal striatum in instrumental reward learning. The selectivity for instrumental prediction errors fits with theories that propose a role for this region in updating action values, as opposed to updating stimulus-specific cue values (Balleine, Delgado, & Hikosaka, 2007; Samejima, Ueda, Doya, & Kimura, 2005; Haruno et al., 2004; O'Doherty et al., 2004). This study extends that role to updating even observed action values but also advances our understanding of how action values are represented in the dorsal striatum. Participants made no motor responses on Observed trials and did not receive any outcomes. The updating process in this region cannot therefore operate on an action representation as simple as a reflexive linkage between an instrumental cue and a personal motor response. Instead, it must operate on a more abstract representation that can be affected by how an individual interprets an observed outcome and that can be translated at the proper moment (when that instrumental response can be performed by the individual herself) into a specific personal motor action.

By contrast, observational reward prediction errors did not activate the ventral striatum in either instrumental or noninstrumental trials. Reward prediction errors were found for both instrumental and noninstrumental experienced outcomes in ventral and posterior striatum, replicating earlier findings (Schultz, 2006; McClure et al., 2003; O'Doherty et al., 2003). As the experienced rewards involved greater sensory involvement, as well as higher subjective rewards (as measured by the liking ratings), one possibility is that the ventral striatum is more sensitive than the dorsal striatum to these sensory and subjective factors. Prior

evidence suggests ventral striatum is critical for encoding subjective reward value per se in stimulus–outcome associations (Cardinal, Parkinson, Hall, & Everitt, 2002). This possibility is consistent with the ideas suggested by actor/critic accounts of ventral striatum: namely, that ventral striatum is selectively involved in updating stimulus- or state-based subjective values, which are used in both instrumental and noninstrumental conditioning, as opposed to the action values updated by dorsal striatum, which are used only during instrumental conditioning. Observing another person's reward after she selects a given action might increase the odds that you will produce that action given the chance; this increase corresponds to an increased action value. The same observation, though, may not improve your estimation of how rewarded you will feel for the next observed reward; this corresponds to an unchanged stimulus value for the observed cues.

One caveat to the interpretation that dorsal striatum is selective for instrumental versus noninstrumental observational learning (as it is in experiential learning) is that participants did not exhibit strong behavioral evidence of learning in the noninstrumental Observed condition (as measured by the nonsignificant model fit to RTs and the liking ratings compared with neutral cues). It may simply be the case that passive observation of cues predicting reward in others is an insufficiently salient event to drive learning. Perhaps including the affective responses generated by the observed partner in response to the cues or outcomes in the observational paradigm could enhance behavioral learning in the noninstrumental condition; this is an important direction for future research on observational conditioning. Because learning was not clearly established in this noninstrumental observational condition we cannot conclude anything about the representation of noninstrumental observational prediction errors in the present study. However, including the noninstrumental observational condition in our analysis does allow us to exclude the contribution of stimulus- or motor-related confounds in the instrumental observational condition, as these features were similar between instrumental and noninstrumental conditions. Thus, the noninstrumental condition still acts as a useful control condition in the present task.

These findings are consistent with a growing literature that suggests processing other people's outcomes involves some, but not all, of the same neural systems as processing the same outcome when it happens to oneself (Singer & Lamm, 2009; Keysers & Gazzola, 2006). Observing others in pain, for example, activates some of the regions involved in experiencing pain (like the anterior insula and anterior cingulate), but other regions are not necessarily activated by that observation—specifically those thought to encode the primary sensory experience of pain (like sensorimotor cortex and posterior insula; Decety & Lamm, 2009; Singer et al., 2004). Observing errors in others can activate both regions involved in personal errors (such as cingulate cortex) and those involved in personal rewards (such as ventral striatum), depending on context (de Bruijn, de Lange, von Cramon, & Ullsperger, 2009). Observed rewards can also activate regions that process personal rewards, like the ventral striatum or ventromedial pFC (Burke et al., 2010), but these regions tend to be more active for personal than observed rewards (Mobbs et al., 2009). The current work extends these findings into the dorsal striatum and not only highlights that the regions involved can be consistent between experience and observation but also identifies their specific computational roles.

These results differ somewhat from a recent study of observational learning, which found that the ventromedial pFC encoded a positive reward prediction error and the ventral striatum encoded a negative reward prediction error for others' outcomes during an instrumental learning task that also included experienced rewards (Burke et al., 2010). Social context plays an especially important role in the response to others' outcomes (Cooper, Kreps, Wiebe, Pirkl, & Knutson, 2010; Mobbs et al., 2009; Singer et al., 2006;

Delgado, Frank, & Phelps, 2005), and variations between the tasks may help explain the differences. Burke et al. suggest their findings in the ventral striatum may have been driven by comparison effects between the observed outcomes and the participants' own; our experiment used liquid instead of monetary outcomes, which may lend themselves less to social comparisons. In addition, in their task, participants received experienced outcomes for choices between the same stimuli as the observed partners, immediately following the observed outcomes. Regions like the ventromedial pFC and ventral striatum are more engaged in updating stimulus-based values, which were likely more relevant in their task. In our task, by contrast, participants never received experienced outcomes following the Observed cues, and so updating subjective stimulus values for those cues was likely less salient in our study. Finally, although their study did not find dorsal striatal activation, it had participants choose between stimuli, whereas ours had participants make instrumental choices with an explicitly action-based cue (a slot machine arm); this may have made encoding action values more salient. Future studies should test how these task variations affect the computations supporting observational learning.

Although this study focused on dorsal striatal function, another important question for future research is how dorsomedial pFC (including anterior cingulate) may be involved in observational learning. These regions are thought to be involved both in representing others' mental states (Amodio & Frith, 2006; Mitchell, Macrae, & Banaji, 2005) and in personal action-based decision-making (Jocham, Neumann, Klein, Danielmeier, & Ullsperger, 2009; Rushworth, Mars, & Summerfield, 2009). The current study did not find prediction error activation in this region, perhaps because of the incentive domain (liquid rewards compared with symbolic or monetary rewards), task design (comparing rewards to neutral instead of to punishments), or task demands (e.g., ACC response to prediction errors is thought in part to encode uncertainty or volatility in expected rewards, which changed only slowly in this study). These regions may also be promising targets for study with more complex computational models than the simple RL model employed here, such as models explicitly modeling beliefs about others' mental states (Behrens, Hunt, Woolrich, & Rushworth, 2008; Hampton, Bossaerts, & O'Doherty, 2008).

Another important direction for future study will be to understand the contributions of different subregions of the dorsal striatum to different learning processes. In this study, Experienced prediction errors (instrumental and noninstrumental) were found in a lateral, posterior region of dorsal putamen, whereas instrumental prediction errors (Observed and Experienced) were found in an anterior region of dorsal caudate. These findings might relate to accumulating evidence that suggests neuroanatomical and functional differentiation within the dorsal striatum in both rodents and humans, with dorsomedial striatum being closely connected to dorsal prefrontal regions and supporting goal-directed behavior and dorsolateral putamen being more closely connected to ventral and motor cortex and supporting habitual behavior (Balleine & O'Doherty, 2010; Haber & Knutson, 2010; Tricomi, Balleine, & O'Doherty, 2009; Tanaka, Balleine, & O'Doherty, 2008; Yin, Ostlund, & Balleine, 2008; Yin, Ostlund, Knowlton, & Balleine, 2005). In the current findings, the more anterior medial activation may thus be more related to more abstract goal-directed processes within reward learning, which may be relatively more important without primary reward (as in Observed learning). The more posterior activation may be more related to more habitual or sensory processes within reward learning, which would be more important for Experienced than Observed trials and present to some extent on noninstrumental trials. More research in humans is needed, however, to clearly distinguish the functions of these distinct areas of dorsal striatum.

The finding of observational reward prediction error signals in the dorsal striatum may also relate to previous findings reporting prediction error signals in the dorsal striatum during

learning about counterfactual events (Lohrenz, McCabe, Camerer, & Montague, 2007). As in the observational case, counterfactual prediction errors are elicited in relation to an outcome that is not directly experienced; instead, these errors correspond to what would have been obtained had an alternative action been selected. The finding that counterfactual prediction errors and observational prediction errors both recruit dorsal striatum could suggest that these two signals may depend on similar underlying computational processes, although future work will be necessary to directly compare and contrast these two types of learning signal.

In summary, we found that a region of dorsal caudate was activated by instrumental reward prediction errors during observational learning, indicating that this region may update an abstract representation of action values even when the actions and outcomes are observed. These results extend our understanding of the functions of the dorsal striatum in instrumental reward learning and in RL by showing that this region can be engaged during learning of instrumental actions without either direct experience of the outcomes or requiring the actions to be personally performed.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors gratefully acknowledge financial support from the Wellcome Trust (grant WT087388AIA to J. O. D.) and the Irish Research Council for Science, Engineering, and Technology (to J. C. C.). We thank Sojo Joseph for technical support and the members of the O'Doherty laboratory for helpful comments.

REFERENCES

Amodio DM, Frith CD. Meeting of minds: The medial frontal cortex and social cognition. Nature Reviews Neuroscience. 2006; 7:268–277.

Balleine BW, Delgado MR, Hikosaka O. The role of the dorsal striatum in reward and decision-making. Journal of Neuroscience. 2007; 27:8161–8165. [PubMed: 17670959]

Balleine BW, O'Doherty JP. Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. Neuropsychopharmacology. 2010; 35:48–69. [PubMed: 19776734]

Bandura, A. Social learning theory. Prentice-Hall; Englewood Cliffs, NJ: 1977.

Behrens TE, Hunt LT, Woolrich MW, Rushworth MF. Associative learning of social value. Nature. 2008; 456:246–249.

Bray S, O'Doherty J. Neural coding of reward prediction error signals during classical conditioning with attractive faces. Journal of Neurophysiology. 2007; 97:3036–3045. [PubMed: 17303809]

Burke CJ, Tobler PN, Baddeley M, Schultz W. Neural mechanisms of observational learning. Proceedings of the National Academy of Sciences, U.S.A. 2010; 107:14431–14436.

Cardinal RN, Parkinson JA, Hall J, Everitt BJ. Emotion and motivation: The role of the amygdala, ventral striatum, and prefrontal cortex. Neuroscience and Biobehavioral Reviews. 2002; 26:321–352. [PubMed: 12034134]

Cooper JC, Kreps TA, Wiebe T, Pirkl T, Knutson B. When giving is good: Ventromedial prefrontal cortex activation for others' intentions. Neuron. 2010; 67:511–521. [PubMed: 20696386]

Daw ND, Doya K. The computational neurobiology of learning and reward. Current Opinion in Neurobiology. 2006; 16:199–204. [PubMed: 16563737]

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ. Cortical substrates for exploratory decisions in humans. Nature. 2006; 441:876–879. [PubMed: 16778890]

Dayan P, Balleine BW. Reward, motivation, and reinforcement learning. Neuron. 2002; 36:285–298. [PubMed: 12383782]

de Bruijn ER, de Lange FP, von Cramon DY, Ullsperger M. When errors are rewarding. Journal of Neuroscience. 2009; 29:12183–12186. [PubMed: 19793976]

Decety, J.; Lamm, C. The biological basis of empathy. In: Cacioppo, JT.; Berntson, GG., editors. Handbook of neuroscience for the behavioral sciences. John Wiley and Sons; New York: 2009.

Deichmann R, Gottfried JA, Hutton C, Turner R. Optimized EPI for fMRI studies of the orbitofrontal cortex. Neuroimage. 2003; 19:430–441. [PubMed: 12814592]

Delgado MR. Reward-related responses in the human striatum. Annals of the New York Academy of Sciences. 2007; 1104:70–88. [PubMed: 17344522]

Delgado MR, Frank RH, Phelps EA. Perceptions of moral character modulate the neural systems of reward during the trust game. Nature Neuroscience. 2005; 8:1611–1618.

Gläscher J. Visualization of group inference data in functional neuroimaging. Neuroinformatics. 2009; 7:73–82. [PubMed: 19140033]

Gläscher J, Daw N, Dayan P, O'Doherty JP. States vs. rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron. 2010; 66:585–595. [PubMed: 20510862]

Haber SN, Knutson B. The reward circuit: Linking primate anatomy and human imaging. Neuropsychpharmacology. 2010; 35:4–26.

Hampton AN, Bossaerts P, O'Doherty JP. Neural correlates of mentalizing-related computations during strategic interactions in humans. Proceedings of the National Academy of Sciences, U.S.A. 2008; 105:6741–6746.

Haruno M, Kuroda T, Doya K, Toyama K, Kimura M, Samejima K, et al. A neural correlate of reward-based behavioral learning in caudate nucleus: A functional magnetic resonance imaging study of a stochastic decision task. Journal of Neuroscience. 2004; 24:1660–1665. [PubMed: 14973239]

Heyes CM, Dawson GR. A demonstration of observational learning in rats using a bidirectional control. The Quarterly Journal of Experimental Psychology, Series B, Comparative and Physiological Psychology. 1990; 42:59–71.

Jocham G, Neumann J, Klein TA, Danielmeier C, Ullsperger M. Adaptive coding of action values in the human rostral cingulate zone. Journal of Neuroscience. 2009; 29:7489–7496. [PubMed: 19515916]

Keysers C, Gazzola V. Towards a unifying neural theory of social cognition. Progress in Brain Research. 2006; 156:379–401. [PubMed: 17015092]

Kriegeskorte N, Simmons WK, Bellgowan PS, Baker CI. Circular analysis in systems neuroscience: The dangers of double dipping. Nature Neuroscience. 2009; 12:535–540.

Lohrenz T, McCabe K, Camerer CF, Montague PR. Neural signature of fictive learning signals in a sequential investment task. Proceedings of the National Academy of Sciences, U.S.A. 2007; 104:9493–9498.

McClure SM, Berns GS, Montague PR. Temporal prediction errors in a passive learning task activate human striatum. Neuron. 2003; 38:339–346. [PubMed: 12718866]

Mitchell JP, Macrae CN, Banaji MR. Forming impressions of people versus inanimate objects: Social-cognitive processing in the medial prefrontal cortex. Neuroimage. 2005; 26:251–257. [PubMed: 15862225]

Mobbs D, Yu R, Meyer M, Passamonti L, Seymour B, Calder AJ, et al. A key role for similarity in vicarious reward. Science. 2009; 324:900. [PubMed: 19443777]

Montague PR, King-Casas B, Cohen JD. Imaging valuation models in human choice. Annual Review of Neuroscience. 2006; 29:417–448.

Myers WA. Observational learning in monkeys. Journal of the Experimental Analysis of Behavior. 1970; 14:225–235. [PubMed: 16811470]

O'Doherty J, Dayan P, Schultz J, Deichmann R, Friston KJ, Dolan RJ. Dissociable roles of the ventral and dorsal striatum in instrumental conditioning. Science. 2004; 304:452–454. [PubMed: 15087550]

O'Doherty JP. Reward representations and reward-related learning in the human brain: Insights from neuroimaging. Current Opinion in Neurobiology. 2004; 14:769–776. [PubMed: 15582382]

O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. Temporal difference models and reward-related learning in the human brain. Neuron. 2003; 38:329–337. [PubMed: 12718865]

Pessiglione M, Seymour B, Flandin G, Dolan RJ, Frith CD. Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature. 2006; 442:1042–1045. [PubMed: 16929307]

Rescorla, RA.; Wagner, AR. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In: Black, AH.; Prokasy, WF., editors. Classical conditioning II: Current research and theory. Appleton Century Crofts; New York: 1972. p. 65-99.

Rushworth MF, Mars RB, Summerfield C. General mechanisms for making decisions? Current Opinion in Neurobiology. 2009; 19:75–83. [PubMed: 19349160]

Samejima K, Ueda Y, Doya K, Kimura M. Representation of action-specific reward values in the striatum. Science. 2005; 310:1337–1340. [PubMed: 16311337]

Schönberg T, O'Doherty JP, Joel D, Inzelberg R, Segev Y, Daw ND. Selective impairment of prediction error signaling in human dorsolateral but not ventral striatum in Parkinson's disease patients: Evidence from a model-based fMRI study. Neuroimage. 2010; 49:772–781. [PubMed: 19682583]

Schultz W. Behavioral theories and the neurophysiology of reward. Annual Review of Psychology. 2006; 57:87–115.

Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275:1593–1599. [PubMed: 9054347]

Singer T, Lamm C. The social neuroscience of empathy. Annals of the New York Academy of Sciences. 2009; 1156:81–96. [PubMed: 19338504]

Singer T, Seymour B, O'Doherty J, Kaube H, Dolan R, Frith CD. Empathy for pain involves the affective but not sensory components of pain. Science. 2004; 303:1157–1162. [PubMed: 14976305]

Singer T, Seymour B, O'Doherty J, Stephan K, Dolan R, Frith CD. Empathic neural responses are modulated by the perceived fairness of others. Nature. 2006; 439:466–469. [PubMed: 16421576]

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. Neuroimage. 2009; 46:1004–1017. [PubMed: 19306932]

Suri RE, Schultz W. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. Neuroscience. 1999; 91:871–890. [PubMed: 10391468]

Sutton, RS.; Barto, AG. Reinforcement learning: An introduction. MIT Press; Cambridge, MA: 1998.

Tanaka SC, Balleine BW, O'Doherty JP. Calculating consequences: Brain systems that encode the causal effects of actions. Journal of Neuroscience. 2008; 28:6750–6755. [PubMed: 18579749]

Tricomi E, Balleine BW, O'Doherty JP. A specific role for posterior dorsolateral striatum in human habit learning. European Journal of Neuroscience. 2009; 29:2225–2232. [PubMed: 19490086]

Tricomi EM, Delgado MR, Fiez JA. Modulation of caudate activity by action contingency. Neuron. 2004; 41:281–292. [PubMed: 14741108]

Valentin VV, O'Doherty JP. Overlapping prediction errors in dorsal striatum during instrumental learning with juice and money reward in the human brain. Journal of Neurophysiology. 2009; 102:3384–3391. [PubMed: 19793875]

Worsley KJ, Marrett S, Neelin P, Vandal AC, Friston KJ, Evans AC. A unified statistical approach for determining significant voxels in images of cerebral activation. Human Brain Mapping. 1996; 4:58–73. [PubMed: 20408186]

Yin HH, Ostlund SB, Balleine BW. Reward-guided learning beyond dopamine in the nucleus accumbens: The integrative functions of cortico-basal ganglia networks. European Journal of Neuroscience. 2008; 28:1437–1448. [PubMed: 18793321]

Yin HH, Ostlund SB, Knowlton BJ, Balleine BW. The role of the dorsomedial striatum in instrumental conditioning. European Journal of Neuroscience. 2005; 22:513–523. [PubMed: 16045504]

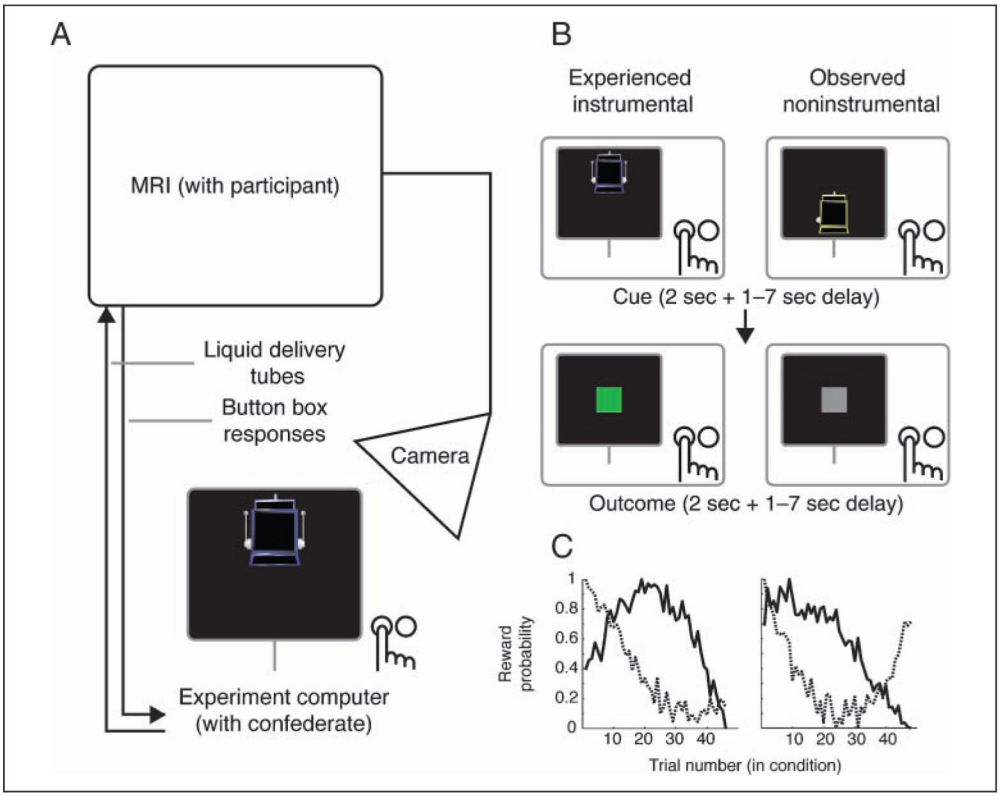

Figure 1.

Experiment setup and trial structure. (A) Schematic of experimental setup. Participants lay in the MRI scanner while a confederate partner sat at the experiment computer outside (shown as hand with buttons, not to scale; confederate button presses were obscured from view). Participants viewed the entire experiment via a live video camera feed aimed at the outside monitor. Responses from inside the scanner were connected to the experiment computer, and liquid outcomes were delivered by pumps controlled by the experiment computer. (B) Trial structure and schematic of participant view. Two trial types are shown, as viewed on participant screen. Participant screen showed experiment computer and confederate hand (with button presses obscured). In Experienced trials, participants saw a slot machine cue on their half of the screen (e.g., top) and made their response. After a variable delay, the cue was followed by a liquid reward or neutral outcome, as well as a colored indicator (reward shown). In Observed trials, participants saw the slot machine cue on the confederate's half of the screen and observed her response; the liquid outcome was then nominally delivered to the confederate instead of the participant, whereas the visual indicator was identical to Experienced trials (neutral outcome shown). On instrumental trials, the participant or confederate selected between two arms; on noninstrumental trials, the computer chose one side or the other (left indicator shown) and the participant or confederate responded to the chosen side with a button press. (C) Representative reward probability curves for two conditions. Each line indicates the probability of reward for one arm or side of a given condition's slot machine over the experiment for a single participant.

Europe PMC Funders Author Manuscripts

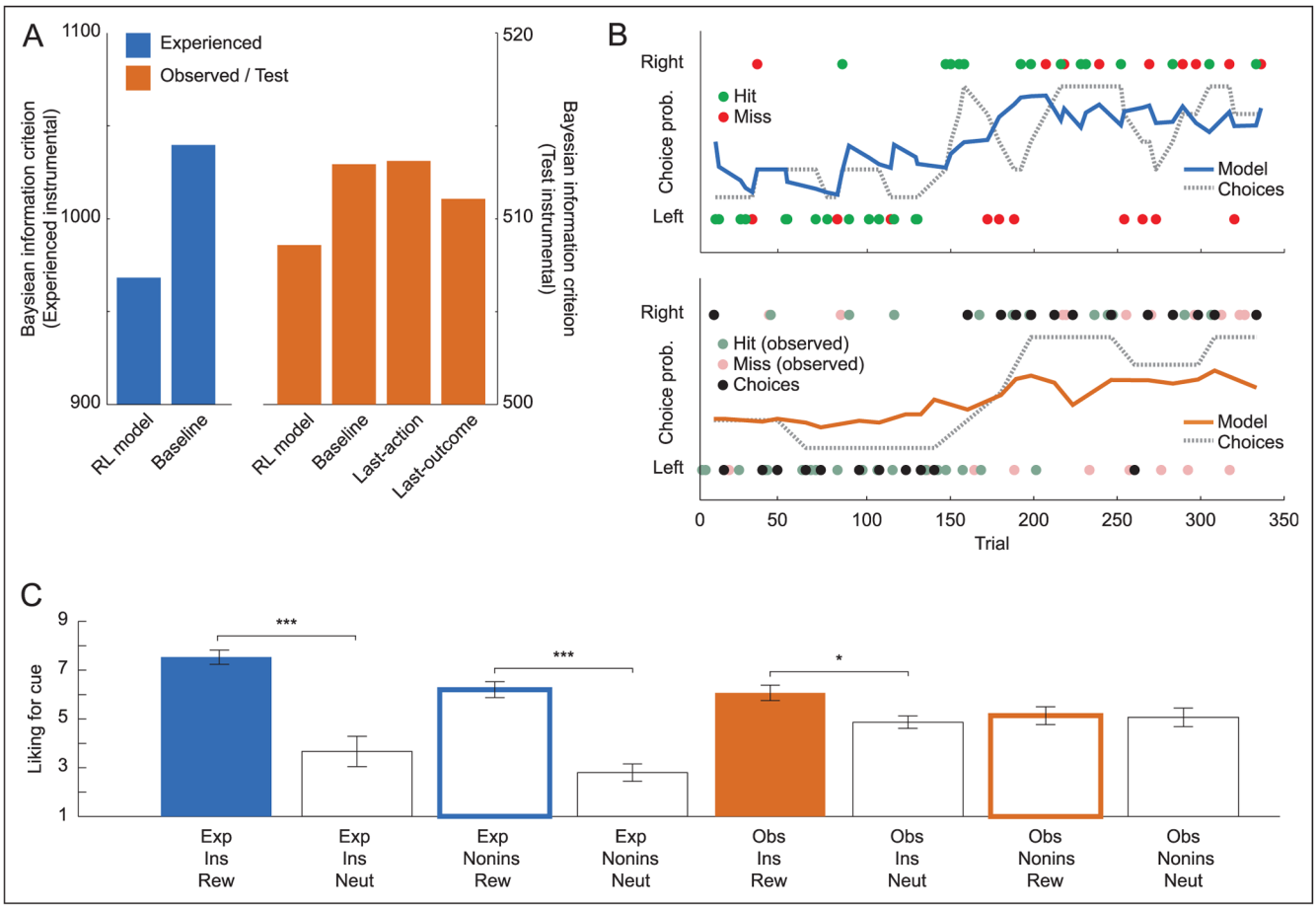

Figure 2.

Conditioning for experienced and observed outcomes. (A) Model fit for RL model of instrumental learning compared with alternative models. Fit is measured by BIC (smaller indicates better fit). See Results for model details. Saturated model not shown for clarity. (B) Time course of instrumental choices and RL model predictions for single representative participant. Top: Experienced trials. Circles indicate choice of left or right action (top or bottom of *y* axis); color indicates reward or neutral outcome. Dashed line indicates average choice probability over previous four trials at each time point (a smoothed measure of behavior). Solid line indicates model-predicted choice probability at each time point. Bottom: Observed and Test trials. Colored circles indicate Observed confederate choices; color indicates reward or neutral outcome for confederate. Dark circles indicate participant Test-trial choices, which were performed in extinction (without any outcome during scan). Lines indicate average Test-trial choice probability and model predictions. (C) Average self-reported liking for cues after experiment. Exp = Experienced, Obs = Observed, Ins = instrumental, Nonins = Noninstrumental, Rew = reward cue, Neut = neutral control. Error bars indicate *SEM* across participants. Only significant differences between reward cue and neutral control machines within condition are shown. ***$p < .001$, *$p < .05$.

Europe PMC Funders Author Manuscripts

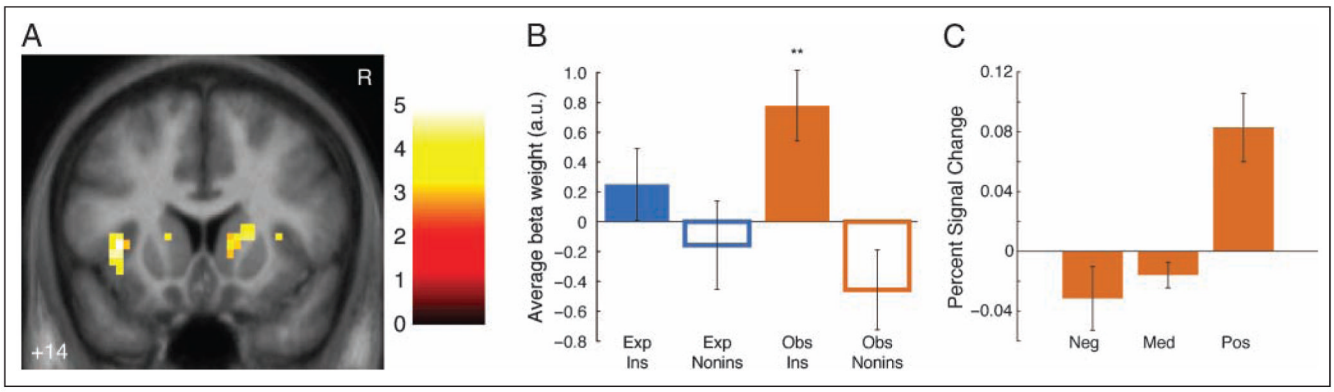

Figure 3.

Dorsal caudate activation for Observed instrumental prediction errors. (A) Activation for Observed instrumental prediction error regressor. Maps are thresholded at $p < .005$ voxelwise with 5 voxel extent threshold for display; cluster in right dorsal caudate meets extent threshold corrected for multiple comparisons across dorsal striatum. Coordinates are in ICBM/MNI space. Color bar indicates *t* statistic. R indicates right. (B) Average beta weights (calculated with leave-one-out extraction; see Methods) in significant dorsal caudate cluster. Error bars indicate *SEM*s across participants. Only significant differences from baseline shown; between-condition tests show only main effect of instrumental versus noninstrumental conditions ($F(1, 63) = 12.65$, $p < .005$). $^{**}p < .01$. (C) Dorsal caudate activation by prediction error size. Bars indicate estimated effect size (in percent signal change) in significant dorsal caudate cluster (calculated with leave-one-out extraction) for outcomes on instrumental Observed trials by prediction error size and valence. Effect sizes estimated as canonical hemodynamic response peak, adjusted for all other conditions (Gläscher, 2009). Neg = prediction error < -0.33. Med = prediction error ≥ -0.33 and < 0.33. Pos = prediction error ≥ 0.33. Significant differences are not shown. Error bars are *SEM* across participants.

Europe PMC Funders Author Manuscripts

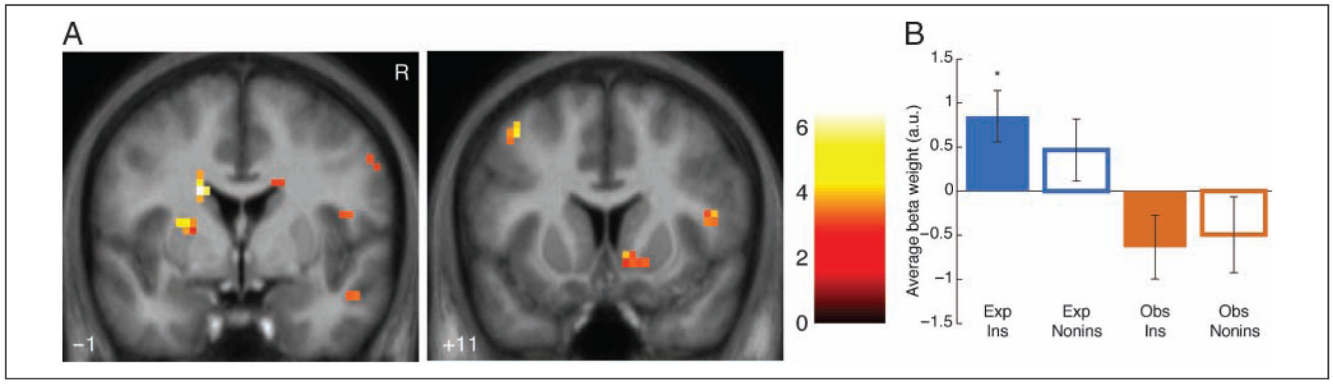

Figure 4.

Dorsal and ventral striatum activation for Experienced instrumental prediction errors. (A) Activation for Experienced instrumental prediction error regressor. Ventral striatal cluster (right) meets extent threshold corrected for multiple comparisons across ventral striatum. Maps are thresholded at $p < .005$ voxelwise with 5 voxel extent threshold for display. Coordinates are in ICBM/MNI space. Color bar indicates t statistic. R indicates right. (B) Average beta weights (calculated with leave-one-out extraction; see Methods) in significant dorsal putamen cluster. Error bars indicate standard errors of the mean across participants. Only significant differences from baseline shown; between-condition tests show only main effect of Experienced vs. Observed conditions ($F(1, 63) = 11.64$, $p < .005$). *$p < .05$.

Europe PMC Funders Author Manuscripts

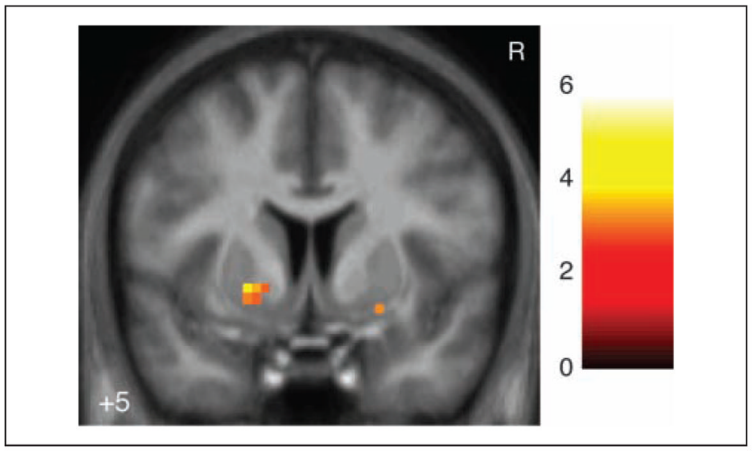

Figure 5.
Ventral striatum activation for Experienced noninstrumental prediction errors. Left cluster meets extent threshold corrected for multiple comparisons across ventral striatum. Map is thresholded at $p < .005$ voxelwise with 5 voxel extent threshold for display. Coordinates are in ICBM/MNI space. Color bar indicates *t* statistic. R indicates right.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Table 1

Average Reaction Time

Condition	Mean Reaction Time, msec (SEM)
Experienced instrumental reward cue	973.93_a (39.87)
Experienced instrumental neutral control	1039.64_b (44.28)
Test instrumental	1041.13_b (41.18)
Experienced noninstrumental reward cue	613.89_c (23.94)
Experienced noninstrumental neutral control	606.39_c (29.55)
Test noninstrumental	637.68_c (24.99)

$n = 15$ (one participant's data excluded as an outlier). Entries that do not share subscripts differ at $p < .05$ by Tukey's honestly significant difference; entries that share subscripts do not significantly differ. Differences were calculated on log-transformed RTs; original data are shown for clarity. *SEM*s are calculated across participants within condition. Only reward cues were shown in Test trials. See Supplementary Figure S1 for full distributions.