Behavioral/Systems/Cognitive

Human Dorsal Striatal Activity during Choice Discriminates Reinforcement Learning Behavior from the Gambler's Fallacy

Ryan K. Jessup and John P. O'Doherty

Trinity College Institute of Neuroscience, School of Psychology, Trinity College Dublin, Dublin 2, Ireland, and Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, California 91125

Reinforcement learning theory has generated substantial interest in neurobiology, particularly because of the resemblance between phasic dopamine and reward prediction errors. Actor–critic theories have been adapted to account for the functions of the striatum, with parts of the dorsal striatum equated to the actor. Here, we specifically test whether the human dorsal striatum—as predicted by an actor–critic instantiation—is used on a trial-to-trial basis at the time of choice to choose in accordance with reinforcement learning theory, as opposed to a competing strategy: the gambler's fallacy. Using a partial-brain functional magnetic resonance imaging scanning protocol focused on the striatum and other ventral brain areas, we found that the dorsal striatum is more active when choosing consistent with reinforcement learning compared with the competing strategy. Moreover, an overlapping area of dorsal striatum along with the ventral striatum was found to be correlated with reward prediction errors at the time of outcome, as predicted by the actor–critic framework. These findings suggest that the same region of dorsal striatum involved in learning stimulus–response associations may contribute to the control of behavior during choice, thereby using those learned associations. Intriguingly, neither reinforcement learning nor the gambler's fallacy conformed to the optimal choice strategy on the specific decision-making task we used. Thus, the dorsal striatum may contribute to the control of behavior according to reinforcement learning even when the prescriptions of such an algorithm are suboptimal in terms of maximizing future rewards.

Introduction

Reinforcement learning (RL) models have stimulated considerable interest as a framework for understanding reward-related learning in biological organisms (Schultz et al., 1997; Doya, 1999; Daw et al., 2006). The phasic activity of midbrain dopamine neurons resembles a reward prediction error (PE) signal inherent in RL models (Schultz et al., 1997; Hollerman and Schultz, 1998), and human neuroimaging studies have reported PEs in the striatum, a prominent target of such neurons (O'Doherty et al., 2007). Different regions of the striatum show unique correlations with PEs depending on the nature of the task: studies involving action selection for reward implicate the dorsal and ventral striatum, while those involving only stimulus–reward associations report PEs primarily in the ventral striatum (O'Doherty et al., 2004). Such results are best understood in terms of actor–critic variants of RL whereby PEs projecting to a dorsal striatal actor modulate the probability of selecting particular actions, while PEs projecting to a ventral striatal critic update stimulus–reward associations (Montague et al., 1996; Sutton and Barto, 1998; O'Doherty et al., 2007; Schönberg et al., 2007).

Much less is known about striatal contributions in guiding action-selection itself. Schönberg et al. (2007) reported significant differences in reward PEs in the dorsal striatum as a function of between-subject performance differences on a simple reward task. Other work revealed correlations between the magnitude of dorsal striatal activity and the degree to which subjects evidenced behavioral reward-learning, along with the degree to which individuals perceived the existence of a contingency between actions and reinforcement (Haruno et al., 2004; Tricomi et al., 2004). These findings suggest that the dorsal striatum contributes to RL-consistent action selection, yet the extent to which neural activity in the dorsal striatum couples directly to behavior during individual choices remains unaddressed.

Conversely, despite the utility of RL, not all human choice behavior coheres to this framework, particularly, the "gambler's fallacy" (Jarvik, 1951). Individuals adhering to the gambler's fallacy (GF) appear to assume non-independence between sequential outcomes, consistent with the belief that the probability of obtaining a reward decreases after recent reinforcement; consequently, they increase their probability of choosing an action that was not recently rewarded (Estes, 1964; Tversky and Kahneman, 1971). Individuals in a real casino have demonstrated the fallacy (Croson and Sundali, 2005), and when non-humans (e.g., a computer or slot machine) generate sequences of events, the sequences are considered more likely to show the negative recency expected by GF (Ayton and Fischer, 2004). More recent work has

Received Dec. 9, 2010; revised Feb. 7, 2011; accepted March 4, 2011.

Author contributions: R.K.J. and J.P.O. designed research; R.K.J. performed research; R.K.J. analyzed data; R.K.J. and J.P.O. wrote the paper.

This work was funded by Science Foundation Ireland Grant 08/IN.1/B1844 to J.P.O. We thank Donal Cahill for his help recruiting participants and collecting data, and Charlotte Prevost for advice on the manuscript.

Correspondence should be addressed to John P. O'Doherty, MC 228-77, California Institute of Technology, 1200 E. California Blvd., Pasadena, CA 91125. E-mail: jdoherty@caltech.edu.

DOI:10.1523/JNEUROSCI.6421-10.2011

 0270-6474/11/316296-09$15.00/0

examined factors modulating GF and the extent to which it represents rationality (Hahn and Warren, 2009; Barron and Leider, 2010).

Our goal is to ascertain the dorsal striatum's role in choice, hypothesizing a role for this region in mediating RL-consistent behavior. Our task was designed to elicit a variety of choice strategies across trials and human functional magnetic resonance imaging (fMRI; using a restricted field of view encompassing the orbitofrontal cortex ventrally, extending as far as the dorsal borders of the striatum) to measure variation in activity in this area as a function of different choice behavior.

Materials and Methods

Thirty-one participants—18 female and 13 male—completed four 13 min 2 s runs of a roulette wheel task while lying in a whole-body magnetic resonance imaging (MRI) scanner. Three subjects' fMRI data were excluded because of technical problems with the MRI scanner. Each of the four runs consisted of three experimental blocks and one control block. Each experimental block consisted of 16 trials; control blocks contained 4 trials. Each run began with 20 s of a fixation cross and ended with a final fixation period of variable length. On each trial, the participant saw a tricolored roulette wheel with 40% of the area covered by one color (Hi) and the remaining 60% of the area covered by two other colors in equal proportion (Lo options). Participants were clearly and correctly instructed that the amount of area covered by a color indicated the probability that the spinner would land on that color and that the stopping probability of the spinner was completely independent from one trial to the next. If the spinner stopped on the chosen color, participants won €2. On every experimental trial, regardless of whether or not an option was chosen, participants were charged €0.50. The optimal choice method is to choose the color filling a larger proportion of the wheel, resulting in a per trial expected value of $0.40 \cdot 2 - 0.50 = 0.30$ for each of the 192 experimental trials. At the end of the experiment, participants were given the amount that they won over the course of the experiment, in addition to €5 for participation.

The location (left, right, or top) of each color differed between participants but within participants was constant for the entire task. Between each experimental block, the size of the area covered by each color changed (e.g., on block 1, blue might cover 40% of the area but on block 2, red would cover 40% of the area), but within each block the area remained constant. Within each run, each color covered 40% of the area for one entire block. At the trial onset, the wheel appeared and subjects had 1500 ms to select a color with their right hand (Fig. 1). Afterward, a spinner appeared, spun for 3000 ms, and then stopped abruptly, remaining visible for a further 500 ms. The outcome was then revealed and remained on-screen for 1000 ms. The intertrial interval (ITI) was drawn from a quasi-normal distribution, ranging from 4000 to 12,000 ms with a mean of 8000 ms, during which time a fixation cross was presented on-screen. The outcome image was either a €2 coin (win) or the same coin with a red X covering it (loss). Although suboptimal for enabling estimation of the two time points in the trial (cue and outcome), the spin period was not temporally jittered due to the fact that dopamine neuron PEs projecting to the striatum are known to approximate the properties of a temporal difference learning algorithm. Introducing temporal unpredictability in the time of outcome presentation via spin period jittering could introduce confounding time varying PE responses in the interstimulus interval (O'Doherty et al., 2003).

Control blocks were identical except for the following changes: there were only 4 trials per block, the entire wheel was shown in gray (participants were told to select any of the three available buttons), and the ITI was drawn from a quasi-uniform distribution with a mean of 8000 ms. The outcome image was a scrambled version of the €2 coin. At the start of each block, a screen appeared for 2000 ms, with the following text: "New Round Get Ready."

We were interested in behavior during a truly independent and identically distributed scenario; therefore, no attempt was made to regulate the number of streaks of each length. Had we regulated the number of streaks of each length, the environment would have transitioned to sampling without replacement; in such an environment, GF-consistent behavior is no longer a fallacy but rather is rational (this is because such environments show the negative recency expected by GF). This resulted in differing numbers of streaks of each length. The mean frequencies of win and loss streaks for Hi and Lo outcomes at each streak length are displayed in Table 1.

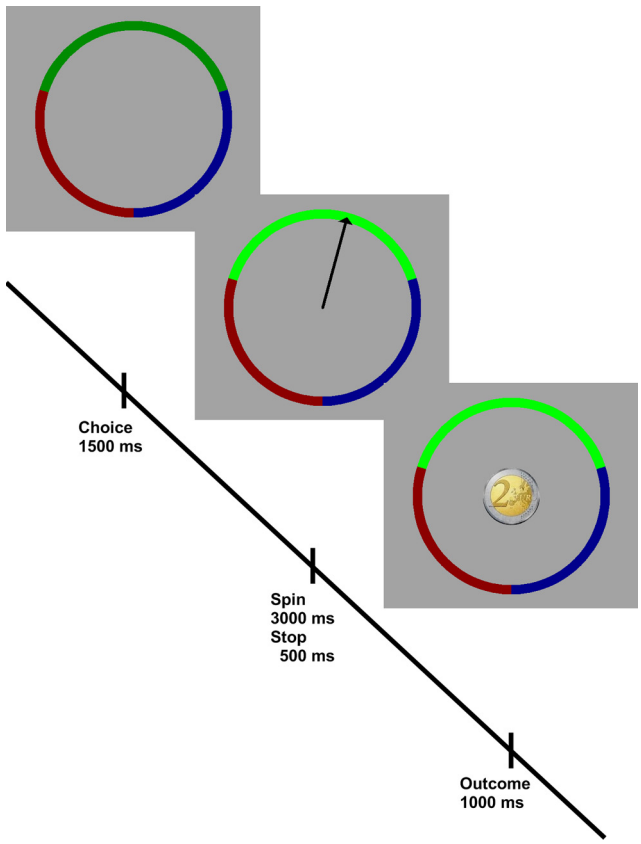

Figure 1. Roulette gambling task. Participants were given 1000 ms to select a colored option. After 1000 ms, if an option was selected (here, as denoted by the lightened color, green was selected), a spinner appeared and spun for 3000 ms, and after stopping, remained on-screen for a further 500 ms. The reward was displayed for 1000 ms. If the spinner stopped on the selected color, as was the case here, the participant received €2; otherwise, he or she won nothing and the same coin superimposed with a red X was displayed. ITI was jittered, with a mean = 8000 ms.

Table 1. Mean frequencies per streak length, separated by outcome

	1	2	3	4	5	6	7
Hi Win	43.65	16.42	6.06	2.45	1.16	0.52	0.10
Hi Lose	39.10	22.97	14.10	8.42	4.77	2.74	1.39
Lo Win	75.81	20.55	6.23	1.52	0.42	0.16	0.03
Lo Lose	80.35	54.06	36.16	23.19	15.39	10.39	6.52

Data represent the mean frequencies across subjects for consecutive identical outcomes; e.g., column 2 of Hi Win shows that the mean number per subject of consecutive wins for the Hi option having a streak of 2 or greater is 16.42.

Informed consent was obtained and the study was approved by the School of Psychology Research Ethics Committee at Trinity College Dublin. Participants received detailed instructions, were guided through 4 practice trials, and then completed 4 practice trials in real time (no feedback regarding the outcomes of their choice was revealed during any of these practice trials).

Behavioral model fitting and comparison. Different variants of RL models were fit to the behavioral data, although for clarity, only one experimental model is presented. This model best fit the individuals who were classified as RL-consistent (described in Results). This model was *SARSA* (state-action-reward-state-action) learning according to

$$Q_i(t) = Q_i(t-1) + \lambda \cdot (R_i(t-1) - Q_i(t-1)), \quad (1)$$

with fictive (also known as full information or foregone payoff) updating. The fictive component allows for individuals to learn from options that were not selected; this is appropriate in our task as it always reveals the winning option via the spinner stopping location. In the above formula, $Q_i(t)$ represents the value of choosing option i on trial t. This quantity is computed by adding the Q value on the previous trial to the product of the updating parameter λ and the (fictive) PE. The PE is the difference between the fictive reward amount R that would have been received had option i been chosen on the preceding trial and the previous Q value for option i. Note that this differs slightly from a traditional PE—in which the Q value for only the chosen option is updated—because the spinner stopping location informs the participant what he or she would have received regardless of his or her selection. $R_i = 1$ if option i won, otherwise 0. Values for $Q_i(0)$ were initialized to the observed win probabilities, as described on the roulette wheel (i.e., [0.4, 0.3, 0.3]).

The softmax choice rule, a specific implementation of the Luce choice axiom (Luce, 1959) was used to generate choice predictions:

$$P_i(t) = e^{Q_i(t)/\theta} \Big/ \sum_{j \in J} e^{Q_j(t)/\theta}. \quad (2)$$

Here, the probability P of choosing option i on trial t is obtained by first exponentiating the quotient of the Q value for option i divided by the inverse temperature parameter θ. This value is then divided by the sum over all options j (out of the total set of options J) of the exponentiated quotient of the Q value for each option j divided by the scaling parameter θ. For each new block, $Q_i(0)$ was reset to the observed win probabilities for each of the options i. There were two free parameters in this RL model: λ and θ.

In addition to RL models, we also fit multiple variants of an existing GF model (Rabin, 2002), adapted for our task, along with other unique instantiations. The best overall version was flexible enough to coarsely account for RL behavioral patterns as well:

$$Q_i(t) = (\chi - \varphi_i(t)) \cdot \lambda + Q_i(0). \quad (3)$$

As above, $Q_i(t)$ represents the value of choosing option i on trial t. This quantity is computed by obtaining the difference between the crossover tolerance parameter χ and the length of the current win streak φ for option i on trial t. This difference is multiplied by the updating parameter λ and then added to the initialized win probability for that option $Q_i(0)$ as presented on the roulette wheel. Values for $Q_i(0)$ were initialized to the observed win probabilities, as described on the roulette wheel (i.e., [0.4, 0.3, 0.3]). The Q value for each option was updated on each trial and all Q values were scaled to sum to unity for each trial. If the minimum Q value was <0, all Q values were shifted up until the $\min(Q) = 0$ and then the Q values were scaled to sum to unity.

Note that while this model shares similarities with the *SARSA* learning rule used above, it also deviates in important ways. First, unlike RL models, on each trial t this model updates $Q_i(0)$ instead of $Q_i(t-1)$. Second, the difference of the crossover tolerance parameter and the streak length can effectively cause the learning rate parameter to switch signs after a fixed number of consecutive identical outcomes (the learning rate parameter was allowed to vary in the range $[-1, 1]$). Hence, when the crossover tolerance parameter exceeded the win streak and the updating parameter was positive $(+ +)$, the likelihood of selecting the streaking option increased, à la RL. Conversely, when either the crossover tolerance parameter value was less than the win streak parameter and the updating parameter was positive $(- +)$ or the crossover tolerance parameter value exceeded the streak length value and the updating parameter value was negative $(+ -)$, the likelihood of selecting the streaking option decreased, à la GF. If a trial was missed, then the probabilities for choosing on the ensuing trial were reset to the initial win probabilities. As with the RL model, the softmax choice rule was used to generate choice predictions. There were three free parameters in this model: χ, λ, θ.

The best fitting parameters were found for each individual by minimizing the negative log likelihood of the predictions, using fminsearch, a hill-climbing algorithm in Matlab (MathWorks). The fit values for each subject were then compared with the fit values from a saturated baseline model which perfectly reproduced the marginal choice probabilities and hence had two free parameters. The only way an experimental model (i.e., RL or GF) could outperform such a baseline model was to account for learning or strategy-related changes in individuals' behavior.

The comparison method used was the Bayesian information criterion (BIC) (Schwarz, 1978):

$$\mathrm{BIC} = 2 \cdot (\mathrm{LL_E} - \mathrm{LL_B}) + k \cdot \ln N. \quad (4)$$

Here, the difference in the log likelihood for the experimental model $\mathrm{LL_E}$ and that of the baseline model $\mathrm{LL_B}$ is doubled and then added to the product of the additional free parameters k for the experimental model and the log of the data points N. As formulated above, this is technically a relative BIC, indicating superior performance of the experimental model when positive, and superior performance of the baseline model when negative.

fMRI acquisition and data preprocessing. The task was conducted in a Philips Achieva 3T scanner using a phased-array eight channel head coil. The imaging data were acquired at a 30° angle from the anterior commissure–posterior commissure line to maximize orbital sensitivity (Deichmann et al., 2003), using a gradient echo T2* weighted echoplanar imaging sequence; 32 2.8 mm ascending slices were acquired with a 0.3 mm slice gap (echo time = 30 ms; repetition time = 2000 ms; voxel size 2 mm, 2 mm, 2.8 mm; matrix 112 × 112 voxels; field of view 224 × 224 mm). The field of view extended from the orbitofrontal cortex and ventral striatum to the superior border of the dorsal striatum. A total of 390 images were collected in each of four runs totaling 1560 functional scans. High-resolution T1 images were collected at the beginning of the session for each participant.

SPM5 (Wellcome Department of Imaging Neuroscience, London, UK; www.fil.ion.ucl.ac.uk/spm) was used to preprocess and analyze the data. The functional data for each participant were slice time corrected to the first slice and then spatially realigned using a six parameter rigid body spatial transformation. The high-resolution structural image was then coregistered to the mean functional image generated by the realignment phase. The mean functional was spatially normalized to the echoplanar imaging template, and the resulting warping parameters were then applied to the realigned functionals and the coregistered structural image. The functionals were then spatially smoothed using an 8 mm Gaussian kernel.

fMRI analysis. We examined changes in blood oxygenation level-dependent (BOLD) response across conditions using a general linear model (GLM) with random effects. For each subject, multiple GLMs were estimated. A 128 s high-pass cutoff filter and a first-order autoregressive correction for serial correlation were applied to each model. Task conditions were convolved with a canonical hemodynamic response function and then entered into the GLM matrix. Each scan of each GLM included a baseline regressor and six movement regressors, generated from the spatial realignment preprocessing step. Additionally, each model had one regressor each for the initial fixation period, final fixation period, at the time of the onset of the roulette wheel for all missed trials, and at the onset of the wheel for neutral trials. The fixation regressors had duration lengths equal to the length of the fixation period.

For statistical inference we used an omnibus height threshold of $p < 0.005$, and then corrected for multiple comparisons at $p < 0.05$ family-wise error (*pFWE*) within small volumes defined on a priori regions of interest in the dorsal striatum, amygdala, and ventromedial prefrontal cortex. For these we took coordinates from relevant prior studies and performed correction for multiple comparison within an 8 mm sphere as described in Results. The extent threshold (*pFWE_cluster*) and cluster size k are reported for each comparison. The MNI (Montreal Neurological Institute) coordinates for the dorsal striatum were taken from a paper which found overlapping PEs within the left dorsal striatum across a variety of rewarding stimuli during an instrumental learning task [−9, 3, 15] (Valentin and O'Doherty, 2009). Because participants in our task used their right hand to make a choice, it was reasonable to expect activation lateralized to the left of the dorsal striatum. The MNI coordinates for the amygdala were from a paper which used high-resolution imaging to examine reward-based activation in an instrumental learning task in bilateral amygdala (Prevost et al., 2011). The left [−23, −7.5, −19] and right [25, −4.5, −19] coordinates were separately averaged across the multiple significant voxels in each hemisphere from that paper. The MNI

Table 2. Regressors used in RL > GF contrast

Category	Previous choice	Previous outcome	Current behavior	Current choice
RL	Hi	Not Hi	Switch	Not Hi
RL	Not Hi	Hi	Switch	Hi
RL	Hi	Hi	Stay	Hi
RL	Not Hi	Not Hi	Stay	Not Hi
GF	Not Hi	Not Hi	Switch	Hi
GF	Hi	Hi	Switch	Not Hi
GF	Not Hi	Hi	Stay	Not Hi
GF	Hi	Not Hi	Stay	Hi

Not Hi refers to either of the two Lo probability options. Switching between Lo options was relatively rare; thus, while separate regressors were created for these events, they were not included in the contrast. Each of the above regressors contained streaks of two or more identical consecutive outcomes.

coordinates for ventromedial prefrontal cortex were obtained from a paper which used instrumental choice to look for goal-directed signals within this region [0, 33, −24] (Valentin et al., 2007). When multiple significant peaks within a single cluster were observed, the statistics for only the maximal peak voxel are reported.

Categorical GLM. The categorical GLM was designed to look for BOLD differences observed when choosing according to RL versus GF. Separate regressors were used to account for neural activity at both the choice and the outcome feedback stage. For the choice phase, one regressor was used for the first trial on each block, as this trial was irrelevant to choosing according to RL or GF. All other choices were defined according to the Hi option because it was expected to draw most of the attention because of its dominance. Regressors were further separated on the basis of multiple factors: whether the Hi option won on the previous trial (Hi Won or Hi Lost), whether the streak length—i.e., how many consecutive identical outcomes for the Hi option—was one or greater (1 or >1), whether the participant chose the Hi option on the current trial (Chose Hi, Chose Not Hi), and whether this choice was identical to the previous choice (stay or switch). Thus, a participant who experienced and engaged in all of these behaviors would have $1 + 2 \cdot 2 \cdot 2 = 9$ regressors beyond the aforementioned regressors. This approach captures nearly all possible choices. On rare occasions a participant switched from one of the low-probability options to the other, adding a maximum of $2 \cdot 2 \cdot 1 = 4$ more regressors. All of these regressors were placed at the onset of the roulette wheel time point. For the feedback stage, one regressor was added for each of the four possible outcomes: win, lose, neutral, and missed choice. These regressors were placed at the time point on which the spinner stopped spinning. Missed responses were registered on only 3 experimental trials and 0.5 control trials on average, across subjects.

The regressors shown in Table 2 were used to test the RL versus GF striatal hypothesis, using outcome streaks >1. Essentially, this contrast compares trials on which a subject chooses (does not choose) the Hi when it is winning (losing) with trials on which the subject chooses (does not choose) the Hi when it is losing (winning). Note that this contrast controls for differences in switch versus stay activity, choice of Hi versus a Lo option, and whether the preceding trial resulted in a win or loss.

RL GLM. The RL GLM used a model-based analysis to reveal BOLD increases correlating with the time series extracted from the predictions using the best fitting parameters from an RL model fit to each participant's behavioral data. The RL model used here differed slightly from the model described above, in that fictive updating was not applied, as this implementation slightly outperformed the model with fictive updating across all subjects. For each trial, the Q(chosen option) was entered into the GLM as a parametric modulator onto a choice regressor. Likewise, an outcome regressor was modulated by the PE extracted from the RL model. As with the categorical model, a separate regressor was added in at the time of feedback for missed choice trials.

Results

Behavioral

Because the known win probabilities were unequal for the three options, the Hi option stochastically dominates. Behavioral violations of stochastic dominance demonstrate irrationality (Diederich and Busemeyer, 1999); hence, individuals should always choose the Hi option, or, allowing for errors, on >95% of the trials. However, across all participants, the mean probability of choosing the Hi option Pr(Choose Hi) = 0.64 and median Pr(Choose Hi) = 0.59. Thus, over one third of all choices can be classified as irrational, significantly >0 ($t_{(30)} = 9.97$, $p < 0.001$).

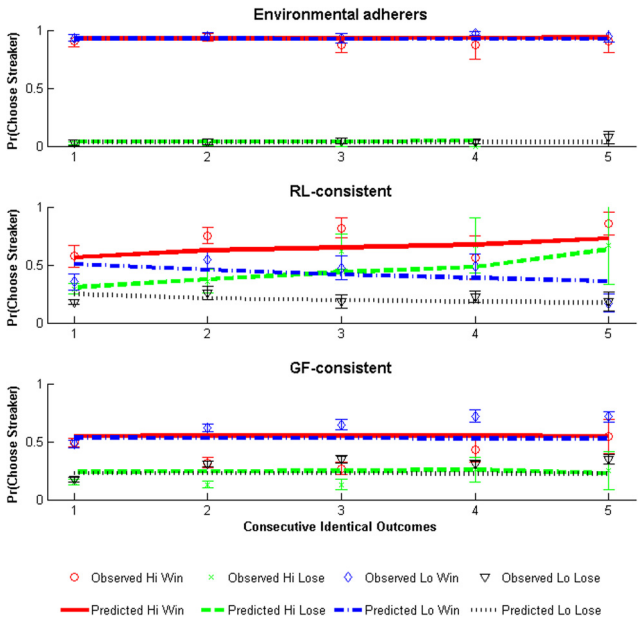

Figure 2. Behavioral results and modeling. The choice data for each subject were separately fit to the RL model described in the text. Subjects were classified according to the weighted RL metric into three patterns of behavior: environmental adherence, RL-consistent, and GF-consistent. The panels show the mean choice behavior (data points) and model predictions (lines), separated by pattern (environmental adherence in the top, RL-consistent in the middle, GF-consistent in the bottom), and plotted according to the number of consecutive identical outcomes (horizontal axis) and the probability of choosing the outcome that is "streaking" (vertical axis). The different possible streaks are Hi Win (red), Lo Win (green), Hi Lose (blue), and Lo Lose (black). Note that the RL model predictions follow the qualitative trend of behavior for the environmental adherer and RL-consistent patterns but not the GF-consistent pattern.

Considerable individual differences in choice behavior were observed and participants naturally separated into two patterns: environmental adherence and feedback responding. Environmental adherers tend to ignore outcomes, adhering only to the described probabilities as presented on the roulette wheel, not the outcome feedback. Feedback responders allow for the roulette outcomes to drive their behavior. These latter individuals separate into two further patterns: RL-consistent and GF-consistent, i.e., those who choose an option that has been winning and those who choose an option that has been losing, respectively. The following metric discriminates between the three patterns:

$$\text{RL metric} = 100 \cdot (\Pr(\text{Choose Hi}(t+2) \mid \text{Hi Won}(t, t+1)) - \Pr(\text{Choose Hi}(t+2) \mid \text{Hi Lost}(t, t+1))). \quad (5)$$

The RL metric indexes whether individuals were more likely to choose the Hi option when it was winning compared with when it was losing. The RL metric is then weighted by each participant's relevant likelihood of choosing a low-probability option (i.e., trials that were relevant for deciphering whether or not they were RL- or GF-consistent; e.g., this contained trials that were either not the first in the block or did not follow a missed trial). Subjects with an absolute weighted RL metric <1 tended toward environmental adherence. RL-consistent participants had a weighted RL metric ≥1, meaning that they were more likely to choose the Hi

Table 3. Median BIC and best fitting parameter values separated by group

Model	Group	BIC	χ	λ	$1/\theta$
RL	Environmental adherers	−0.43	N/A	0.00	0.02
RL	RL-consistent	9.10	N/A	0.03	0.21
RL	GF-consistent	−0.89	N/A	0.00	0.10
GF	Environmental adherers	−0.43	1.82	0.00	0.02
GF	RL-consistent	2.87	2.27	−0.04	0.17
GF	GF-consistent	2.71	5.50	−0.01	0.10

All data were fit at the individual level. BIC is positive where the experimental model outperformed the baseline model. Parameters: χ is the crossover tolerance parameter and is used by the GF model to switch between RL- and GF-consistent behavior; λ is the updating or learning rate parameter; θ is the temperature parameter and controls the extent to which a participant behaves deterministically.

Table 4. Descriptive statistics for choosing the Hi option, conditioned on previous outcomes

	N	Choose Hi | Hi Won	Choose Hi | Hi Lost
Environmental adherers	8	0.92 (0.12)	0.94 (0.11)
RL-consistent	6	0.77 (0.16)	0.47 (0.17)
GF-consistent	17	0.33 (0.16)	0.65 (0.13)
Total	31	0.57 (0.31)	0.69 (0.21)

The mean (standard deviation) conditional probabilities for choosing the Hi option conditioned on it winning or losing consecutively 2 or more times are given, separated by group. High variance within subject choice is characterized by choice probabilities within a cell that fail to load on either 0 or 1.

when it was winning rather than losing. GF-consistent participants have a weighted RL metric ≤ -1. The three patterns of behavior and the best fits of their data to the RL model are presented in Figure 2. While environmental adherers ($n = 8$) and RL-consistent subjects ($n = 6$) were well fit by the RL model, it is unable to capture the qualitative patterns shown in the GF-consistent subjects ($n = 17$), as they tended to choose a losing option, in contradiction to standard RL theory. Table 3 presents the median best fitting parameters and BIC value across subjects separated by choice pattern for both models. Positive BIC values indicate that the model was superior to the baseline model. For the RL model, only the RL-consistent subjects had a positive BIC score as environmental adherers showed no learning-related changes over time and GF-consistent subjects' behavior defied RL assumptions. For the GF model, both RL- and GF-consistent subjects had positive BIC scores, but the model failed to fit either of these groups as well as the RL model fit the RL-consistent participants. Interestingly, more than half of all subjects are classified as GF-consistent.

While there are overall differences in the propensity of individuals to choose according to GF or RL, considerable trial-by-trial variation was found within participants concerning whether individual choices conformed to either strategy. Table 4 demonstrates the considerable variance on a trial-to-trial basis. The fact that the choice probabilities do not all load on either 0 or 1 reveals this variance (the mean choice probabilities are shown separated by group to demonstrate that this considerable variance is not merely an artifact of averaging over participants with widely disparate choice strategies). Based on this observation, we first describe analyses of within-subject variation in strategy in the neuroimaging data before moving on to explore effects of variation in overall strategy used between subjects.

Neuroimaging

Within-subject analyses

Of primary interest is the extent to which neural activity in the dorsal striatum at the time of choice discriminates—on a trial-to-trial basis—whether participants choose consistent with RL or GF. All participants who were categorized as environmental adherers—because they ignored observed outcomes—were excluded and the analysis was run on the feedback responders. [To maintain a balanced design, only the subjects with all regressors represented were entered in the contrast reported here ($n = 17$); however, the same pattern of results at the same statistical threshold is found when all of the feedback responders are included.]

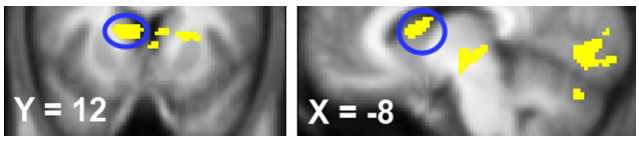

Figure 3. Contrast of RL- with GF-consistent choices. Coronal (left) and sagittal (right) views of the tmap of activation in the RL–GF contrast, thresholded at $pUnc < 0.001$, $k > 100$, and overlaid onto the mean of the participants' structural images. Using a small volume correction, the left caudate activity was significant at $pFWE < 0.05$.

The results indicated that on a trial-to-trial basis, differences in neural activations in the dorsal striatum existed when choosing according to RL as opposed to GF. Figure 3 shows a voxel by voxel map of the *t* statistics (tmap) overlaid onto the mean of the participants' structural images. The significant increase in activity was observed in the left dorsal caudate ($z = 3.83$, $pFWE = 0.003$; $pFWE\_cluster = 0.006$; $k = 147$; peak voxel MNI coordinates: [−8, 10, 12]). When choosing consistently with RL principles, individuals showed increased dorsal striatal activity compared with when choosing consistently with GF. Using the RL GLM, we then tested whether this dorsal striatal difference was due to a difference in the expected reward corresponding to the chosen action, as generated by the RL model, as opposed to being related specifically to a discrimination between RL-consistent and RL-inconsistent behavior. However, no significant correlations were found with the *Q* value regressor generated by the RL model in this region [even at uncorrected p ($pUnc$) < 0.05], suggesting that the dorsal striatal activity at the time of choice is unlikely to be driven simply by a difference in RL-generated estimates of expected reward.

To further delineate the role of the dorsal striatum in RL, we next tested for a corresponding relationship between dorsal caudate activation and PE. This would demonstrate that this region is not only used for RL-consistent choices, as shown above, but also updated by outcome feedback. We tested this with the RL GLM by looking for a significant parametric modulation by PE on the outcome regressor using the data from all participants. We found that left caudate neural activity positively correlated with PE (Fig. 4*A*). Three peaks were found in a single cluster, all of which exceeded a familywise error correction ($z = 3.07$, $pFWE = 0.012$; $pFWE\_cluster = 0.020$; $k = 63$; maximal peak MNI voxel coordinates [−6, 2, 10]). Hence, left dorsal caudate activity is modulated by PE at the time of outcome. Additionally, we observed whole-brain corrected significant activation in the left ventral striatum ($z = 4.49$, $pFWE = 0.038$; $pFWE\_cluster = 0.011$; $k = 1689$; maximal peak MNI voxel coordinates [−14, 0, −16]). The significant dorsal striatal cluster revealed by this contrast overlaps with the significant cluster from the previous contrast (Fig. 4*C*), suggesting that the same dorsal striatal region reports PE on all trials but then is only used when choosing according to RL principles.

We also set out to address whether PE activity observed in the dorsal striatum was best accounted for in terms of the RL model as opposed to that elicited by the GF model. An important feature of the GF model (which is a weakness as well as a strength) is that it is extremely flexible: it can capture both GF- and RL-consistent

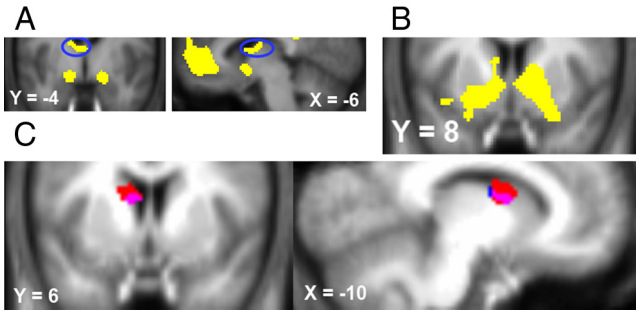

Figure 4. Striatal activity at reward and choice. ***A***, Coronal (left) and sagittal (right) views of the tmap of activation from the correlation of BOLD activation with reward PE from an RL model, thresholded at $pUnc < 0.005$, $k > 60$, and overlaid onto the mean of the participants' structural images. Using a small volume correction, the left caudate activity was significant at $pFWE < 0.05$. ***B***, Coronal view of the tmap of activation in the ventral striatum from the same contrast and threshold as shown in ***A***. The left ventral striatal activity was significant at $pFWE < 0.05$, corrected for the whole brain. ***C***, Coronal (left) and sagittal (right) views of the significant dorsal striatal clusters that show increased activity during RL- compared with GF-consistent choice at the time of stimulus onset (red), a positive correlation with PE at the time of reward (blue), and where the two clusters overlap (magenta).

behavior, depending on the specific parameter values that are used. Indeed, the PE signal generated by the GF model will mimic the RL PE under conditions where a participant's behavior is RL-like. Thus, although they are different in formulation, the RL and GF models can be considered nested: the GF model essentially accommodates everything captured in the RL model, and has the additional flexibility to accommodate GF behavior. We therefore conducted a nested model comparison of PE-related activity starting with the most parsimonious model, the RL model. Then we tested whether the PE signal engendered by the GF model captures any additional variance in the dorsal striatum above and beyond that which is captured by the RL model. To do this, we slightly modified the RL GLM, entering in the PE signals from both models as parametric regressors. Critically, we entered the RL PE signal first in the design matrix, and the GF PE second. We then tested whether additional activity in the dorsal striatum could be captured by the GF model once the RL PE signal had accounted for all of the common variance. We found no significant activity in dorsal striatum loading on the GF regressor, even at $pUnc < 0.01$. Thus, we can conclude that the signal we observe in the dorsal striatum is most parsimoniously accounted for by RL-based models.

Between-subject analyses

In addition to testing for strategy-dependent activity within subjects, we also tested for brain regions showing modulation at different time points as a function of the extent to which a subject's overall choice behavior was consistent with GF or RL. To do this, we used the categorical GLM to analyze activity in the feedback responders group during the gambling task compared with the control task at two time points: the time of choice and the time of outcome receipt. We then tested for areas responding in this contrast at these two time points that were also modulated as a function of the RL metric described in the behavioral results. While no significant between-subject correlations with the RL metric were found at the time of choice, at the time of outcome, we found significant effects in two areas known to play a role in encoding rewarding and punishing outcomes: the ventromedial prefrontal cortex and bilateral amygdala (Fig. 5). For the ventromedial prefrontal cortex, two peaks were found in a single cluster, both of which exceeded a familywise error correction ($z = 3.71$, $pFWE = 0.002$; $pFWE\_cluster = 0.011$; $k = 144$; maximal peak MNI voxel coordinates [4, 28, −24]). For the left amygdala, two clusters, one with four peaks and the other with one peak, were found, all exceeding a familywise error correction ($z = 3.02$, $pFWE = 0.013$; $pFWE\_cluster = 0.023$; $k = 39$; maximal peak MNI voxel coordinates [−22, −4, −26]), and one peak in one cluster was found for the right amygdala ($z = 3.75$, $pFWE = 0.001$; $pFWE\_cluster = 0.012$; $k = 143$; peak MNI voxel coordinates [24, −2, −18]). Activity in both of these regions in response to outcomes—whether a rewarding or a punishing outcome compared with a control outcome—was positively correlated with the RL metric, such that the more an individual conformed to an RL strategy, the greater the activity in these regions in response to outcomes.

Discussion

Our findings implicate the mid-caudate nucleus in the dorsal striatum in mediating RL-consistent behavior. When subjects made RL-consistent choices, activity in this region increased relative to trials on which subjects made RL-inconsistent choices. We excluded other possible explanations for the activity, including whether the preceding trial resulted in a gain or loss, whether subjects maintained their existing choice or switched to a new choice, and whether they chose the Hi or a Lo option. The finding that the dorsal striatum is involved in the expression of RL-consistent behavior builds on a large previous literature implicating the striatum in learning associations via RL mechanisms (Wickens et al., 2003; Frank et al., 2004; Haruno et al., 2004; Roitman et al., 2004; Yin et al., 2004; Knutson and Cooper, 2005; Samejima et al., 2005; Atallah et al., 2007; Balleine et al., 2007; Lau and Glimcher, 2008). This finding also resonates with previous results indicating greater RL-related activity in the dorsal striatum in subjects who learned to choose rewarding actions compared with those who failed to learn (Schönberg et al., 2007). The present results extend those findings, by demonstrating that on a trial-by-trial basis—when subjects choose in an RL-consistent manner—the dorsal striatum is more engaged compared with situations when alternate choice strategies are deployed.

The present results are pertinent to proposals regarding multiple mechanisms for guiding choice in the human brain (Shiffrin and Schneider, 1977; Balleine and Dickinson, 1998; Doya, 1999, 2002; Kahneman and Frederick, 2002; McClure et al., 2004; Daw et al., 2005; Atallah et al., 2007; Balleine and O'Doherty, 2010). One proposal postulates two competing systems: a model-based system wherein choices are computed online using a "model" or cognitive map of the decision problem, and a "model-free" method, in which learned values acquired through trial and error RL are used to generate choices (Daw et al., 2005). Our findings could be interpreted as suggesting a role for the mid-dorsal striatum in model-free RL. Simply choosing recently rewarded actions and avoiding those that were not could be seen as model-free RL, whereas adherence to GF could be construed as "model-based," in that adherence likely requires a richer representation of the task structure—such as how many rewards or losses were obtained and assumptions about the dependency between successive choices of the same option.

It is notable that the same region of dorsal striatum engaged during RL-consistent choice was also correlated with an RL-generated PE signal at outcome time. These findings suggest that the same area of dorsal striatum recruited for learning stimulus–response associations might also be involved during choice when those learned stimulus–response relationships are used to drive behavior. Thus, neural systems involved in controlling choice may substantially overlap with those involved in learning associations underpinning such behavior in the first place. These find-

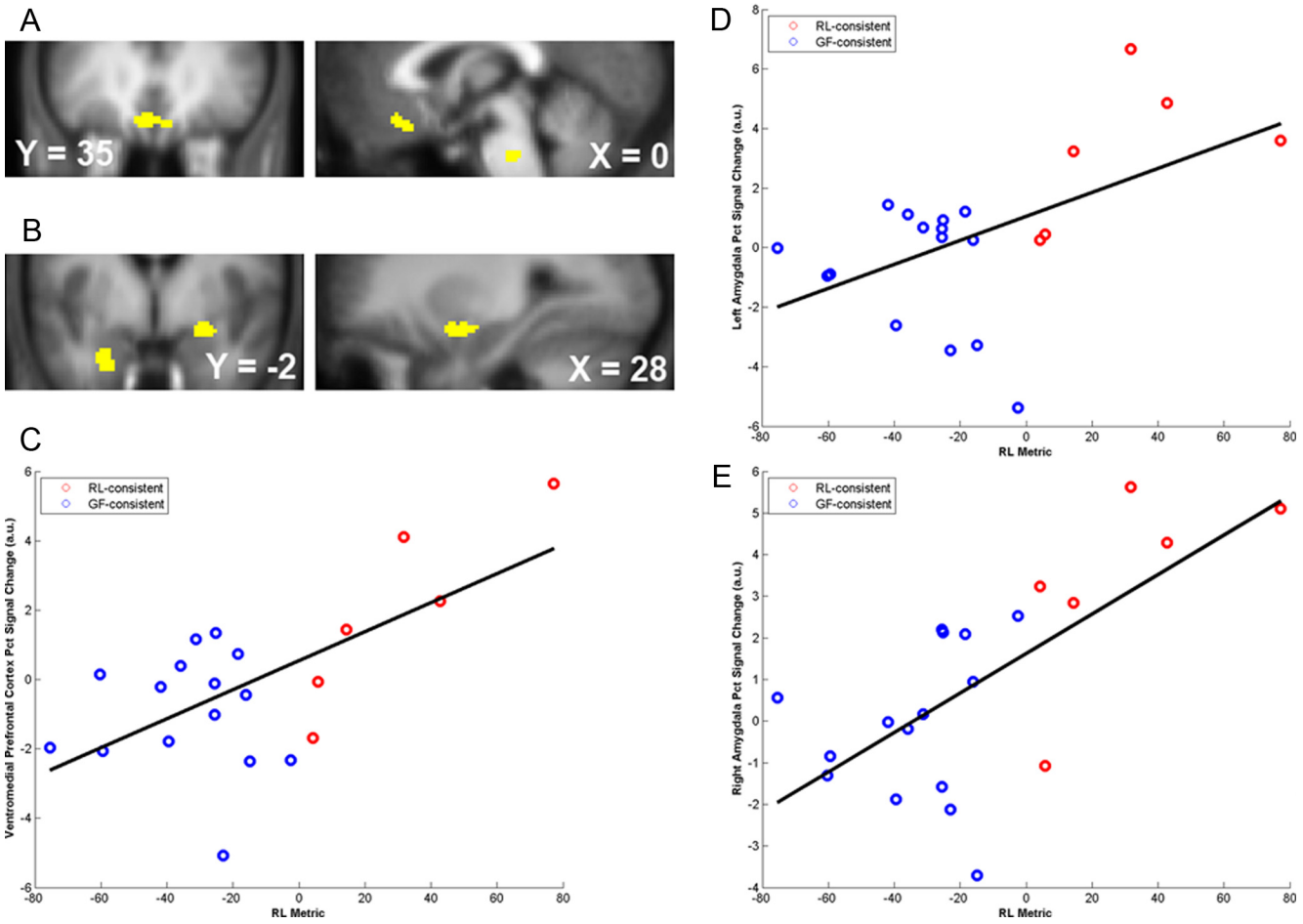

Figure 5. Correlation of BOLD activity with RL metric. ***A***, Coronal (left) and sagittal (right) views of the tmap of activation in the ventromedial prefrontal cortex from the second-level correlation of BOLD activity with the RL metric in the gambling– control contrast, thresholded at $pUnc < 0.001$, $k > 60$, and overlaid onto the mean of the participants' structural images. Using a small volume correction, the cluster was significant at $pFWE < 0.05$. ***B***, Coronal (left) and sagittal (right) views of the tmap of activation in bilateral amygdala from the same contrast and threshold as shown in ***A***. ***C–E***, Scatterplot showing individual data points for the RL metric plotted against the BOLD activity in the ventromedial prefrontal cortex (***C***), left amygdala (***D***), and right amygdala (***E***) clusters, together with regression lines of best fit through the data. As indicated by the weighted RL metric, circles are colored to denote group membership: GF-consistent (blue), RL-consistent (red). The BOLD data in each plot were independently obtained by taking an 8 mm sphere centered on the same small volume coordinates for each region that were used to test for significant contrasts.

ings cohere with actor–critic models of the striatum, providing further evidence that the dorsal striatum resembles the actor (Montague et al., 1996; Balleine and O'Doherty, 2010). It is also possible that another region mediates RL-consistent choice and that the dorsal striatum is more active in such choices because it reports the stimulus–response associations pertinent to the present choice, although that view is less parsimonious than the one presented here. Further research will be needed to tease apart this possibility from the present one.

We also tested for BOLD differences on a between-subject basis as a function of the degree to which individual behavior was overall biased toward exhibiting RL- or GF-like behavior. Although we found no between-subject differences in the striatum, we did find positive correlations between the degree of activity in ventromedial prefrontal cortex and bilateral amygdala in response to outcomes and the propensity for an individual to choose according to RL. This indicates that RL adherents may show enhanced processing of outcomes in these brain regions relative to GF adherents. Increased sensitivity to outcomes could potentially lead to a greater propensity for behavior to be controlled by RL mechanisms compared with alternative strategies, such as GF, that may instead require increased attention to other task components.

It is notable that no brain regions were found to have increased activity when subjects chose according to GF, nor did any regions show enhanced activity across subjects as a function of the propensity to exhibit GF behavior. However, we did not obtain fMRI coverage over dorsal cortical regions including the dorsolateral prefrontal or parietal cortex, and it is conceivable that these areas might have greater involvement in the initiation of GF behavior (Huettel et al., 2002; Akitsuki et al., 2003), especially if they are more relevant to model-based as opposed to model-free learning (Gläscher et al., 2010). For example, when individuals observed violations of patterns—regardless of whether those patterns were of consecutively identical or alternating events—increased activations were observed in medial frontal gyrus, inferior frontal gyrus and anterior cingulate cortex (Huettel et al., 2002). Further studies will be needed to compare and contrast the role of these dorsal cortical brain areas in mediating GF behavior.

Beyond the neural results, the present behavioral findings also have implications for understanding choice behavior. Despite receiving complete information about the nature of the task, the majority of subjects failed to pursue the optimal strategy of always choosing the dominant option (the option yielding a 40% chance of reward). Instead, most subjects varied their choices

according to a mixture of RL and GF, both of which are suboptimal and therefore irrational in this context. This finding builds on previous evidence that individuals tend to either ignore or fail to adhere to described information in a task—even when that information is available on each and every trial—and instead let their choices be driven by either the experience of obtained feedback (Jessup et al., 2008) or perceived structure—such as interdependencies between trials—even in direct contradiction to the explicit instructions. Together, our behavioral results suggest that two "irrational" though competing choice strategies observed in this task tend to dominate choice behavior, even where there is a clear rational alternative. One feasible counterargument to the claim of "irrationality" is that it is possible that the behavioral expression of GF found here could possibly be interpreted as "rational" if volunteers harbored doubts about the veridical nature of the experimenter instructions (Grether and Plott, 1979). Rationality considerations notwithstanding, it is tempting to conclude that these two distinct strategies may be ubiquitous in human choice behavior.

An alternative interpretation is that since stochastic choice functions such as the softmax rule predict variability in choice behavior, GF-consistent choices could be generated by a stochastic choice function, while the overall policy coheres with RL. However, this possibility ignores the theoretical underpinnings of stochastic choice functions such as the softmax rule. Typical explanations for invoking the softmax rule include the notions of environmental uncertainty, exploration, and between-subject analyses (Sutton and Barto, 1998; Daw et al., 2006). However, in the present task, the independence of outcomes was declared at the task outset and the payoff distributions are fully described on each trial; so there is no environmental uncertainty. Moreover, because information about alternate outcomes is always provided, there is no need for exploration: an individual can exploit while simultaneously learning about other options. Also, because this interpretation relates to a within-subject analysis, the between-subject explanation is irrelevant. Hence, while the softmax choice rule can describe choice behavior in this task, it is unable to explain it. Thus, explanations beyond the softmax are required to make sense of the data. Furthermore, and perhaps most importantly, if the observed variability in responses was purely due to noisy or disinterested responding, there should be (1) little discernible pattern in behavioral choice and (2) no difference in dorsal striatal activity at the time of choice when choosing according to RL as opposed to GF. However, as Figures 2 and 3 attest, both of these notions are ruled out. Likewise, more complex decision strategies could be used to recategorize some of the individual trials. However, for the same reason as above, this critique can also be reasonably ruled out due to the observation of reliable differences in brain activity using the current, more parsimonious, manner of categorization.

To conclude, our findings implicate a part of the human dorsal striatum in choice behavior. These findings build on burgeoning evidence implicating the dorsal striatum in the selection of actions leading to reward, and in encoding of stimulus–response or stimulus–response–outcome associations (Lauwereyns et al., 2002; Hikosaka et al., 2006; Balleine et al., 2007). At the point of choice, the mid-caudate is involved in driving choice behavior only when that behavior coheres with the predictions of RL and not alternative, model-based strategies. Intriguingly, in the present study, the most common choice strategies—whether RL- or GF-like—deviate from rationality. Thus, the dorsal striatum may contribute to the control of behavior according to RL even when the prescriptions of such an algorithm are patently suboptimal.

References

Akitsuki Y, Sugiura M, Watanabe J, Yamashita K, Sassa Y, Awata S, Matsuoka H, Maeda Y, Matsue Y, Fukuda H, Kawashima R (2003) Context-dependent cortical activation in response to financial reward and penalty: an event-related fMRI study. Neuroimage 19:1674–1685.

Atallah HE, Lopez-Paniagua D, Rudy JW, O'Reilly RC (2007) Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. Nat Neurosci 10:126–131.

Ayton P, Fischer I (2004) The hot hand fallacy and the gambler's fallacy: two faces of subjective randomness? Mem Cognit 32:1369–1378.

Balleine BW, Dickinson A (1998) Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology 37:407–419.

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.

Balleine BW, Delgado MR, Hikosaka O (2007) The role of the dorsal striatum in reward and decision-making. J Neurosci 27:8161–8165.

Barron G, Leider S (2010) The role of experience in the gambler's fallacy. J Behav Decis Mak 23:117–129.

Croson R, Sundali J (2005) The gambler's fallacy and the hot hand: empirical data from casinos. J Risk Uncertain 30:195–209.

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

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

Deichmann R, Gottfried JA, Hutton C, Turner R (2003) Optimized EPI for fMRI studies of the orbitofrontal cortex. Neuroimage 19:430–441.

Diederich A, Busemeyer JR (1999) Conflict and the stochastic-dominance principle of decision making. Psychol Sci 10:353–359.

Doya K (1999) What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Netw 12:961–974.

Doya K (2002) Metalearning and neuromodulation. Neural Netw 15:495–506.

Estes WK (1964) Probability learning. In: Categories of human learning (Melton AW, ed), pp 89–128. New York: Academic.

Frank MJ, Seeberger LC, O'reilly RC (2004) By carrot or by stick: cognitive reinforcement learning in parkinsonism. Science 306:1940–1943.

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

Grether DM, Plott CR (1979) Economic theory of choice and the preference reversal phenomenon. Am Econ Rev 69:623–638.

Hahn U, Warren PA (2009) Perceptions of randomness: why three heads are better than four. Psychol Rev 116:454–461.

Haruno M, Kuroda T, Doya K, Toyama K, Kimura M, Samejima K, Imamizu H, Kawato M (2004) A neural correlate of reward-based behavioral learning in caudate nucleus: a functional magnetic resonance imaging study of a stochastic decision task. J Neurosci 24:1660–1665.

Hikosaka O, Nakamura K, Nakahara H (2006) Basal ganglia orient eyes to reward. J Neurophysiol 95:567–584.

Hollerman JR, Schultz W (1998) Dopamine neurons report an error in the temporal prediction of reward during learning. Nat Neurosci 1:304–309.

Huettel SA, Mack PB, McCarthy G (2002) Perceiving patterns in random series: dynamic processing of sequence in prefrontal cortex. Nat Neurosci 5:485–490.

Jarvik ME (1951) Probability learning and a negative recency effect in the serial anticipation of alternative symbols. J Exp Psychol 41:291–297.

Jessup RK, Bishara AJ, Busemeyer JR (2008) Feedback produces divergence from prospect theory in descriptive choice. Psychol Sci 19:1015–1022.

Kahneman D, Frederick S (2002) Representativeness revisited: attribute substitution in intuitive judgment. In: Heuristics and biases: the psychology of intuitive judgment (Gilovich T, Griffin D, Kahneman D, eds), pp 49–81. New York: Cambridge UP.

Knutson B, Cooper JC (2005) Functional magnetic resonance imaging of reward prediction. Curr Opin Neurol 18:411–417.

Lau B, Glimcher PW (2008) Value representations in the primate striatum during matching behavior. Neuron 58:451–463.

Lauwereyns J, Watanabe K, Coe B, Hikosaka O (2002) A neural correlate of response bias in monkey caudate nucleus. Nature 418:413–417.

Luce RD (1959) Individual choice behavior: a theoretical analysis. New York: Wiley.
McClure SM, Laibson DI, Loewenstein G, Cohen JD (2004) Separate neural systems value immediate and delayed monetary rewards. Science 306:503–507.
Montague PR, Dayan P, Sejnowski TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J Neurosci 16:1936–1947.
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.
Prevost C, McCabe JA, Jessup RK, Bossaerts P, O'Doherty JP (2011) Differentiable contributions of human amygdalar subregions in the computations underlying reward and avoidance learning. Eur J Neurosci, In press.
Rabin M (2002) Inference by believers in the law of small numbers. Q J Econ 117:775–816.
Roitman MF, Stuber GD, Phillips PE, Wightman RM, Carelli RM (2004) Dopamine operates as a subsecond modulator of food seeking. J Neurosci 24:1265–1271.
Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of action-specific reward values in the striatum. Science 310:1337–1340.
Schönberg T, Daw ND, Joel D, O'Doherty JP (2007) Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. J Neurosci 27:12860–12867.
Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. Science 275:1593–1599.
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464.
Shiffrin RM, Schneider W (1977) Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. Psychol Rev 84:127–190.
Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Cambridge, MA: MIT.
Tricomi EM, Delgado MR, Fiez JA (2004) Modulation of caudate activity by action contingency. Neuron 41:281–292.
Tversky A, Kahneman D (1971) Belief in the law of small numbers. Psychol Bull 76:105–110.
Valentin VV, O'Doherty JP (2009) Overlapping prediction errors in dorsal striatum during instrumental learning with juice and money reward in the human brain. J Neurophysiol 102:3384–3391.
Valentin VV, Dickinson A, O'Doherty JP (2007) Determining the neural substrates of goal-directed learning in the human brain. J Neurosci 27:4019–4126.
Wickens JR, Reynolds JN, Hyland BI (2003) Neural mechanisms of reward-related motor learning. Curr Opin Neurobiol 13:685–690.
Yin HH, Knowlton BJ, Balleine BW (2004) Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur J Neurosci 19:181–189.