Neuron
Article

Model-Based Influences on Humans' Choices and Striatal Prediction Errors

Nathaniel D. Daw,[1,*] Samuel J. Gershman,[2] Ben Seymour,[3] Peter Dayan,[4] and Raymond J. Dolan[3]
[1]Center for Neural Science and Department of Psychology, New York University, New York, NY 10012, USA
[2]Department of Psychology and Neuroscience Institute, Princeton University, Princeton, NJ 08540, USA
[3]Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, WC1N 3BG London, UK
[4]Gatsby Computational Neuroscience Unit, University College London, WC1N 3AR London, UK
*Correspondence: daw@cns.nyu.edu
DOI 10.1016/j.neuron.2011.02.027

SUMMARY

The mesostriatal dopamine system is prominently implicated in model-free reinforcement learning, with fMRI BOLD signals in ventral striatum notably covarying with model-free prediction errors. However, latent learning and devaluation studies show that behavior also shows hallmarks of model-based planning, and the interaction between model-based and model-free values, prediction errors, and preferences is underexplored. We designed a multistep decision task in which model-based and model-free influences on human choice behavior could be distinguished. By showing that choices reflected both influences we could then test the purity of the ventral striatal BOLD signal as a model-free report. Contrary to expectations, the signal reflected both model-free and model-based predictions in proportions matching those that best explained choice behavior. These results challenge the notion of a separate model-free learner and suggest a more integrated computational architecture for high-level human decision-making.

INTRODUCTION

A ubiquitous idea in psychology, neuroscience, and behavioral economics is that the brain contains multiple, distinct systems for decision-making (Daw et al., 2005; Kahneman, 2003; Loewenstein and O'Donoghue, 2004; Rangel et al., 2008; Redish et al., 2008; Sloman, 1996). One long-prominent contender, the "law of effect," states that an action followed by reinforcement is more likely to be repeated in the future (Thorndike, 1911). This habit principle is also at the heart of temporal-difference (TD) learning accounts of the dopaminergic system and its action in striatum (Barto, 1995; Schultz et al., 1997). In the actor-critic, for instance, a dopaminergic "reward prediction error" (RPE) signal plays the role of Thorndike's reinforcer, increasing the propensity to take actions that are followed by positive RPEs (Maia, 2010; Suri and Schultz, 1999).

However, it has long been known that the reinforcement principle offers at best an incomplete account of learned action choice. Evidence from reward devaluation studies suggests that animals can also make "goal-directed" choices, putatively controlled by representations of the likely outcomes of their actions (Dickinson and Balleine, 2002). This realizes a suggestion, dating back at least to Tolman (1948), that animals are not condemned merely to repeat previously reinforced actions.

From the perspective of neuroscience, habits and goal-directed action systems appear to coexist in different cortico-striatal circuits. While these systems learn concurrently, they control behavior differentially under alternative circumstances (Balleine and O'Doherty, 2010; Dickinson, 1985; Killcross and Coutureau, 2003). Computational treatments (Balleine et al., 2008; Daw et al., 2005; Doya, 1999; Niv et al., 2006; Redish et al., 2008) interpret these as two complementary mechanisms for reinforcement learning (RL). The TD mechanism is associated with dopamine and RPEs, and is "model-free" in the sense of eschewing the representation of task structure and instead working directly by reinforcing successful actions. The goal-directed mechanism is a separate "model-based" RL system, which works by using a learned "internal model" of the task to evaluate candidate actions (e.g., by mental simulation; Hassabis and Maguire, 2007; Schacter et al., 2007; perhaps implemented by some form of preplay; Foster and Wilson, 2006; Johnson and Redish, 2007).

Barring one recent exception (Gläscher et al., 2010) (which focused on the different issue of the neural substrates of learning the internal model), previous studies investigating the neural substrates of model-free and model-based control have not attempted to detect simultaneous correlates of both as these systems learn concurrently. Thus, the way the controllers interact is unclear, and the prevailing supposition that neural RPEs originate from a distinct model-free system remains untested. Here we exploited the difference between their two types of action evaluation to investigate the interaction of the controllers in humans quantitatively, using functional MRI (fMRI). Model-free evaluation is retrospective, chaining RPEs backward across a sequence of actions. By contrast, model-based evaluation is prospective, directly assessing available future possibilities. Thus, it is possible to distinguish the two using a sequential choice task.

In theory, the choices recommended by model-based and model-free strategies depend on their own, separate valuation

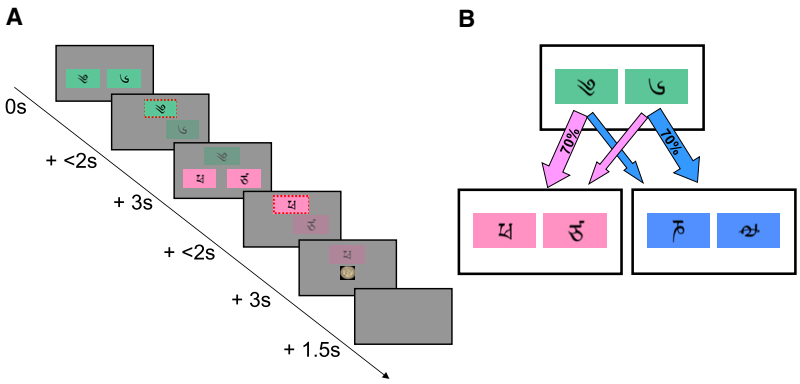

Figure 1. Task Design

(A) Timeline of events in trial. A first-stage choice between two options (green boxes) leads to a second-stage choice (here, between two pink options), which is reinforced with money.

(B) State transition structure. Each first-stage choice is predominantly associated with one or the other of the second-stage states, and leads there 70% of the time.

computations. Thus, if behavior reflects contributions from each strategy, then we can make the clear, testable prediction that neural signals reflecting either valuation should dissociate from behavior (Kable and Glimcher, 2007). Correlates of reward prediction have most repeatedly been demonstrated in fMRI in two areas: the ventromedial prefrontal cortex (vmPFC) and the ventral striatum (ventral putamen and nucleus accumbens) (Delgado et al., 2000; Hare et al., 2008; Knutson et al., 2000, 2007; Lohrenz et al., 2007; O'Doherty, 2004; Peters and Büchel, 2009; Plassmann et al., 2007; Preuschoff et al., 2006; Tanaka et al., 2004; Tom et al., 2007). Of these, value-related signals in mPFC are sensitive to task contingencies, and are thus good candidates for involvement in model-based evaluation (Hampton et al., 2006, 2008; Valentin et al., 2007). Conversely, the ventral striatal signal correlates with an RPE (McClure et al., 2003a; O'Doherty et al., 2003; Seymour et al., 2004), and on standard accounts, is presumed to be associated with dopamine and with a model-free TD system. If so, these signals should reflect *ignorance* of task structure and instead be driven by past reinforcement, even though subjects' behavior, if it is partly under the control of a separate model-based system, may be better informed.

Contrary to this hitherto untested prediction, our results demonstrate that reinforcement-based and model-based value predictions are combined in both brain areas, and more particularly, that RPEs in ventral striatum do not reflect pure model-free TD. These results suggest a more integrated computational account of the neural substrates of valuation.

RESULTS

Behavior

Subjects (n = 17) completed a two-stage Markov decision task (Figure 1) in which, on each trial, an initial choice between two options labeled by (semantically irrelevant) Tibetan characters led probabilistically to either of two, second-stage "states," represented by different colors. In turn, these both demanded another two-option choice, each of which was associated with a different chance of delivering a monetary reward. The choice of one first-stage option led predominantly (70% of the time) to an associated one of the two second-stage states, and this relationship was fixed throughout the experiment. However, to incentivize subjects to continue learning throughout the task, the chances of payoff associated with the four second-stage options were changed slowly and independently, according to Gaussian random walks. Theory (Daw et al., 2005; Dickinson, 1985) predicts that such change should tend to favor the ongoing contribution of model-based evaluation.

Each subject undertook 201 trials, of which 2 ± 2 (mean ± 1 SD) trials were not completed due to failure to enter a response within the 2 s limit. These trials were omitted from analysis.

The logic of the task was that model-based and model-free strategies for RL predict different patterns by which reward obtained in the second stage should impact first-stage choices on subsequent trials. For illustration, consider a trial in which a first-stage choice, uncharacteristically, led to the second-stage state with which it is not usually associated, and in which the choice then made at the second stage was rewarded. The principle of reinforcement would predict that this experience should increase the probability of repeating the first-stage choice because it was ultimately rewarded. However, a subject choosing instead using an internal model of the task's transition structure that evaluates actions prospectively would be expected instead to exhibit a *decreased* tendency to choose that same option. This is because any increase in the value of the rewarded second-stage option will more greatly increase the expected value of the first-stage option that is more likely to lead there. This is actually the first-stage option that was not originally chosen.

Given previous work suggesting the coexistence of multiple valuation processes in the brain (Balleine et al., 2008; Dickinson, 1985), we hypothesized that subjects might exhibit a mixture of both strategies. First, to see learning effects of this sort in a relatively theory-neutral manner, we directly assessed the effect of events on the previous trial (trial *n*) on the choice on the current trial (trial *n+1*). The two key events on trial *n* are whether or not reward was received, and whether the second-stage state presented was common or rare, given the first-stage choice on trial *n*. We evaluated the impact of these events on the chance of repeating the same first-stage choice on trial *n+1*. For reasons outlined above, a simple reinforcement strategy [simulated in Figure 2A using the TD algorithm SARSA(λ) for λ = 1] predicts only a main effect of reward: an ultimately rewarded choice is more likely to be repeated, regardless of whether that reward followed a common or rare transition. Conversely, a model-based strategy (simulated in Figure 2B) predicts a crossover interaction between the two factors, because a rare transition inverts the effect of the subsequent reward.

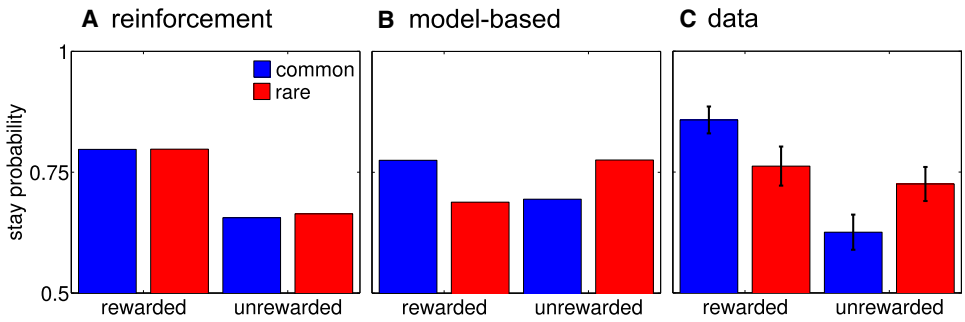

Figure 2. Factorial Analysis of Choice Behavior

(A) Simple reinforcement predicts that a first-stage choice resulting in reward is more likely to be repeated on the subsequent trial, regardless of whether that reward occurred after a common or rare transition.

(B) Model-based prospective evaluation instead predicts that a rare transition should affect the value of the other first-stage option, leading to a predicted interaction between the factors of reward and transition probability.

(C) Actual stay proportions, averaged across subjects, display hallmarks of both strategies. Error bars: 1 SEM.

Figure 2C plots the observed choice proportions as a function of these two factors, in the average across subjects. In order to study effects that were statistically reliable at the level of the population, we quantified the effects using hierarchical logistic regression with all coefficients taken as random effects across subjects. At the population level, the main effect of reward was significantly different from zero ($p < 1e{-}8$, two-tailed), demonstrating a reinforcement effect. However, the interaction between reward and the transition probability was also significant ($p < 5e{-}5$), rejecting a pure reinforcement account and suggesting that subjects take the transition model into account in making their choices. As both theories predict, there was no significant main effect of transition likelihood ($p = 0.5$). Finally, the constant term was significantly positive ($p < 5e{-}12$), suggesting an overall tendency to stick with the same option from trial to trial, reward notwithstanding (Ito and Doya, 2009; Kim et al., 2009; Lau and Glimcher, 2005). We also considered estimates of the effect sizes for each individual within this analysis (conditional on the group-level parameter estimates); the effect of reward was positive (within the 95% confidence interval) for 14/17 subjects, and the interaction was positive for 10/17 individuals, including 7 for whom the main effect of reward was also positive. Together these data suggest that hallmarks of both strategies are seen significantly at the population level and within many individuals, but that there may be between-subject variability in their deployment.

Motivated by these results, we considered the fit of full model-based and model-free [SARSA(λ) TD; Rummery and Niranjan, 1994] RL algorithms to the choice sequences. The former evaluates actions by prospective simulation in a learned model; the latter uses a generalized principle of reinforcement. The generalization, controlled by the reinforcement eligibility parameter λ, is that the estimated value of the second-stage state should act as the same sort of model-free reinforcer for the first-stage choice because the final reward actually received after the second-stage choice. The parameter λ governs the relative importance of these two reinforcers, with $\lambda = 1$ being the special case of Figure 2A in which only the final reward is important, and $\lambda = 0$ being the purest case of the TD algorithm in which only the second-stage value plays a role.

We also considered a hybrid theory (Gläscher et al., 2010) in which subjects could run both algorithms in parallel and make choices according to the weighted combination of the action values that they produce (see Experimental Procedures). We took the relative weight of the two algorithms' values into account in determining the choices to be a free parameter, which we allowed to vary across subjects but assumed to be constant throughout the experiment. Thus, this algorithm contains both the model-based and TD algorithms as special cases, where one or the other gets all weight. We first verified that the model fit significantly better than chance; it did so, at $p < 0.05$ for all 17 subjects (likelihood ratio tests).

We estimated the theory's free parameters individually for each subject by maximum likelihood (Table 1). Such an analysis treats each subject as occupying a point on a continuum trading off the two strategies; tests of the parameter estimates across subjects seek effects that are generalizable to other members of the population (analogous to the random effects level in fMRI; Holmes

Table 1. Best-Fitting Parameter Estimates, Shown as Median Plus Quartiles across Subjects

	β_1	β_2	α_1	α_2	λ	p	w	−LL	$p - r^2$
25th percentile	2.76	2.69	0.46	0.21	0.41	0.02	0.29	167.74	0.17
Median	5.19	3.69	0.54	0.42	0.57	0.11	0.39	200.55	0.26
75th percentile	7.45	5.16	0.87	0.71	0.94	0.22	0.59	228.22	0.40

Also shown are medians and quartiles for the negative log-likelihood (−LL) of the data at the best fitting parameters, and a pseudo-r^2 statistic ($p - r^2$), a normalized measure of the degree to which the model explained the choice data.

Table 2. Model Comparisons between Full (Hybrid) Model and Its Special Cases

	Classical			Bayesian			
	−LL	Number Favoring Hybrid	Aggregate LRT Favoring Hybrid	−log(P(M|D))	Number Favoring Hybrid	Aggregate Log Bayes Factor Favoring Hybrid	Exceedance Probability
hybrid	3364	–	–	3564	–	–	0.92
TD only	3418	5	$\chi^2_{17} = 108$ $p < 5e{-}15$	3594	11	30.0	0.031
model-based only	3501	14	$\chi^2_{51} = 273$ $p < 5e{-}16$	3646	15	82.4	0.0019
$\lambda = 0$	3452	14	$\chi^2_{17} = 176$ $p < 5e{-}16$	3627	16	62.9	0.0012
$\lambda = 1$	3392	4	$\chi^2_{17} = 54.5$ $p < 1e{-}5$	3573	8	8.87	0.049

Shown for each model: raw negative log-likelihood (−LL); the number of subjects favoring the hybrid model on a likelihood ratio test ($p < 0.05$); test statistic and p value for a likelihood ratio test against the hybrid model, aggregated across subjects; the negative log model evidence –log(P(M|D)); the number of subjects favoring the hybrid model according to the model evidence; the log Bayes factor favoring the hybrid model, in the aggregate over subjects; and the Bayesian exceedance probability (Stephan et al., 2009), or probability that each model is the most common among the five over the population.

and Friston, 1998). Due to non-Gaussian statistics (because the parameters are expected to lie in the unit range), we analyzed the estimated parameters' medians using nonparametric tests. Across subjects, the median weighting for model-free RL values was 61% (with model-based RL at 39%), which was significantly different from both 0% and 100% (sign tests, $p < 0.005$), again suggesting that both strategies were mixed in the population. The second important parameter is the reinforcement eligibility parameter λ, which controls the two reinforcement effects in TD, i.e., the relative influence of the estimated value of the second-stage state and the ultimate reward on the model-free value of the first-stage choice. Across subjects, the median estimate for λ was 0.57 (significantly different from 0 and 1; sign tests, $p < 0.05$), suggesting that at the population level, reinforcement occurred in part according to TD-like value chaining ($\lambda < 1$) and in part according to direct reinforcement ($\lambda > 0$).

Since analyzing estimates of the free parameters does not speak to their necessity for explaining data, we used both classical and Bayesian model comparison to test whether these free parameters of the full model were justified by data, relative to four simplifications. We tested the special cases of SARSA(λ) and model-based RL alone, plus the hybrid model, using only direct reinforcement or value chaining (i.e., with λ restricted to 0 or 1). The results in Table 2 show the superiority of the hybrid model both in the aggregate over subjects and also, in most tests, for the majority of subjects considered individually. Finally, we fit the hierarchical model of Stephan et al. (2009) to treat the identity of the best-fitting model as a random effect that itself could vary across subjects. The exceedance probabilities from this analysis, shown in Table 2, indicate that the hybrid model had the highest chance (with probability 92%) of being the most common model in the population. The same analysis estimated the expected proportion of each sort of learner in the population; here the hybrid model was dominant (at 48%), followed by TD at 18%.

Together, these analyses provided compelling support for the proposition that the task exercised both model-free and model-based learning strategies, albeit with evidence for individual variability in the degree to which subjects deploy each of them. Next, armed with the trial-by-trial estimates of the values learned by each putative process from the hybrid algorithm (refit using a mixed-effects model for more stable fMRI estimates; Table 3), we sought neural signals related to these valuation processes.

Neuroimaging

Blood oxygenation level dependent (BOLD) responses in a number of regions—notably the striatum and the mPFC—have repeatedly been shown to covary with subjects' value expectations (Berns et al., 2001; Hare et al., 2008; O'Doherty et al., 2007). The ventral striatum has been closely associated with model-free RL, and so a prime question is whether BOLD signals in this structure indeed reflect model-free knowledge alone, even for subjects whose actual behavior shows model-based influences.

To investigate this question, we sought voxels wherein BOLD activity correlated with two candidate time series. The first time series was the standard RPE based on model-free TD, using just the time points of the transition to the second stage and the delivery of the outcome in order to avoid uncertainty about the appropriate baseline against which to measure the first-stage prediction (see Supplemental Experimental Procedures). The second time series involved subtracting these TD RPEs from the RPEs that would arise if the predictions had been model-based rather than model-free (Daw, in press; Friston et al., 1998; Wittmann et al., 2008).

We adopted this approach (rather than simply including both model-free and model-based RPEs as explanatory variables)

Table 3. Mixed Effects Parameter Estimates Used for fMRI Regressors

β_1	β_2	α_1	α_2	λ	p	w	−LL	$p - r^2$
4.23	2.95	0.70	0.40	0.63	0.17	mean 0.51 SD 0.31	3702	0.22

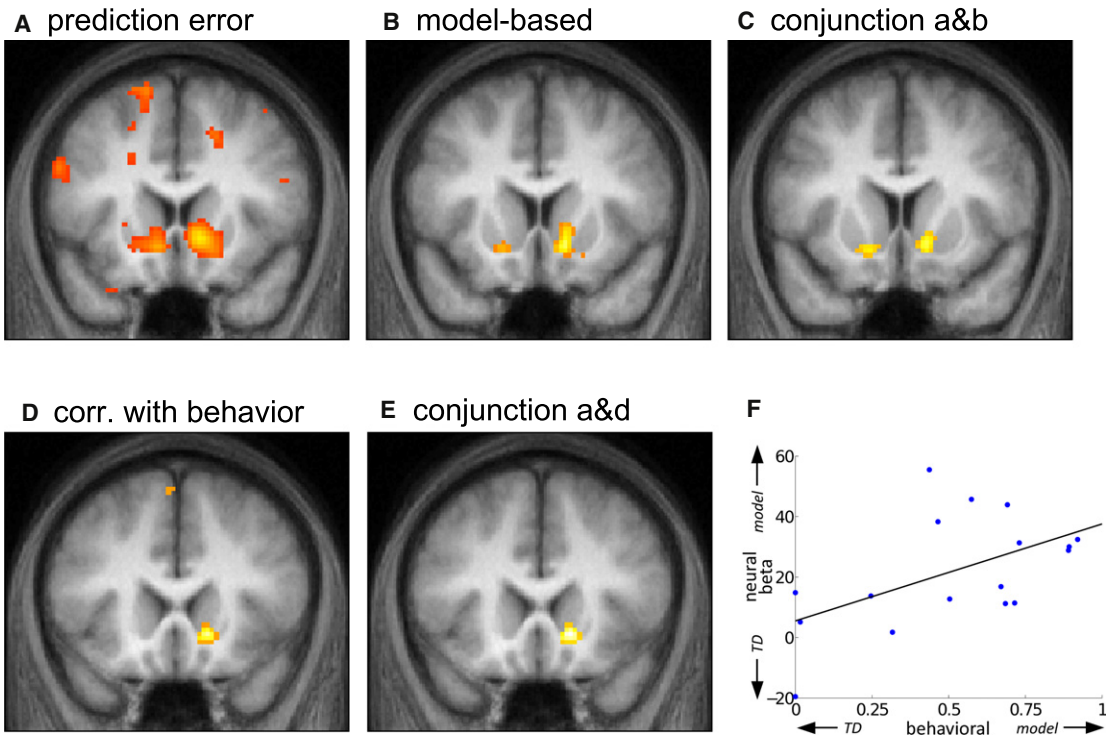

Figure 3. Neural Correlates of Model-free and Model-Based Valuations in RPE in Striatum

All maps are thresholded at $p < 0.001$, uncorrected for display.

(A) Correlates of model-free RPE in bilateral striatum (left peak: −12 10 4, right: 10 12 −4).

(B) RPE signaling in ventral striatum is better explained by including some model-based predictions: correlations with the difference between model-based and model-free RPE signals (left: −10 6 12, right: 12 16 −8).

(C) Conjunction of contrasts from (A) and (B) (left: −12 10 −10, right, 12 16 −6).

(D) Region of right ventral striatum where the weight given to model-based valuations in explaining the BOLD response correlated, across subjects, with that derived from explaining their choice behavior (14 20 −6).

(E) Conjunction of contrasts from (A) and (D) (14 20 −6).

(F) Scatterplot of the correlation from (D), from average activity over an anatomically defined mask of right ventral striatum. ($r^2 = 0.28$, $p = 0.027$.)

to reduce the correlation between the regressors of interest, and also because it encompassed the test of the null hypothesis that RPE signaling in striatum was purely model-free. If so, then the signal would be accounted for entirely by the model-free regressor, and the difference time series should not correlate significantly. If, however, the BOLD signal reflected pure model-based values, or any combination of both, then it would be best described by some weighted combination of the two regressors; that is, the difference regressor would account for residual BOLD activity in addition to that accounted for by the model-free RPE. We tested the conjunction of the two regressors to verify whether BOLD activity in a voxel was indeed significantly correlated with the weighted sum of both (Nichols et al., 2005).

Figure 3A shows that BOLD activity correlated significantly with the model-free RPE time series in left and right ventral striatum (both $p < 0.001$; except where noted, all reported statistics are corrected at the cluster level for familywise error due to whole-brain multiple comparisons). Moreover, this activity was better characterized, on average, as including some model-based valuation: the model-based difference regressor loaded significantly (right, $p < 0.005$, left, $p < 0.05$; Figure 3B) in the same area (conjunction; right, $p < 0.01$, whole-brain corrected; left, $p < 0.01$, small-volume corrected within an anatomically defined mask of the bilateral nucleus accumbens; Figure 3C). Similar results, though less strong, were also observed in medial/vmPFC, where both model-free RPE ($p < 0.001$; Figure 4A) and the difference regressor indicating model-based valuation ($p < 0.01$; Figure 4B) correlated significantly with BOLD activity. However, although the conjunction between these two maps showed voxels significant at $p < 0.001$ uncorrected, it survived whole-brain multiple comparison correction for cluster size (at $p < 0.005$ corrected; Figure 4C) only when the threshold on the conjunction map was relaxed to $p < 0.005$ uncorrected. (Note that cluster size correction is valid independent of the threshold on the underlying uncorrected map, although examining additional thresholds implies additional multiple comparisons; Friston et al., 1993.)

These results suggested that RPE-related BOLD signals in ventral striatum, and also in vmPFC, reflected valuations computed at least in part by model-based methods rather than pure TD. To investigate this activity further, we compared across subjects neural and behavioral estimates of the degree of reliance on model-based valuation. The neural and behavioral

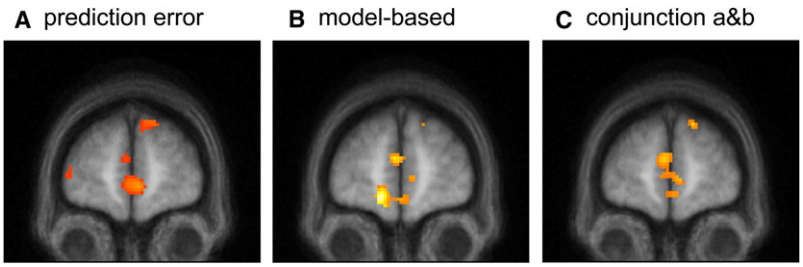

Figure 4. Neural Correlates of Model-free and Model-Based Valuations in RPE in mPFC
Maps have been thresholded at $p < 0.001$ uncorrected (A and B) or $p < 0.005$ uncorrected (C) for display. (A) Correlates of model-free RPE in mPFC (−4 66 14). (B) RPE signaling in mPFC is better explained by including some model-based predictions: correlations with the difference between the two RPE signals (−4 56 14). (C) Conjunction of contrasts from (A) and (B) (−4 62 12).

estimates should correlate if, though computed using different observables, they were measuring the same phenomenon, and if RPE activity in striatum were related to a behaviorally relevant mixture of model-based and model-free values, rather than to one or the other. We measured the degree of model-based valuation in the neural signal by the effect size estimated for the model-based difference regressor (with a larger weighting indicating that the net signal represented an RPE more heavily weighted toward model-based values). Behaviorally, we assessed the degree of model-based influence on choices by the fit of the weighting parameter *w* in the hybrid algorithm. Significant correlation between these two estimates was indeed detected in right ventral striatum ($p < 0.0,1$ small-volume corrected within an anatomical mask of bilateral nucleus accumbens; Figure 3D); and the site of this correlation overlapped the basic RPE signal there ($p < 0.01$, small-volume corrected; Figure 3E). Figure 3F illustrates a scatterplot of the effect, here independently re-estimated from BOLD activity averaged over an anatomically defined mask of right nucleus accumbens. The finding of consistency between both these estimates helps to rule out unanticipated confounds specific to either analysis.

All together, these results suggested that BOLD activity in striatum reflected a mixture of model-free and model-based evaluations, in proportions matching those that determine choice behavior. Finally, in order to characterize more directly this activity and to interrogate this conclusion via an analysis using different data points and weaker theoretical assumptions, we subjected BOLD activity in ventral striatum to a factorial analysis of its dependence on the previous trial's events, analogous to that used for choice behavior in Figure 2. In particular, the TD RPE when a trial starts reflects the value expected during the trial (as in the anticipatory activity of Schultz et al., 1997), which can be quantified as the predicted value of the top-level action chosen (Morris et al., 2006). For reasons analogous to those discussed above for choice behavior, learning by reinforcement as in TD(λ) (for $\lambda > 0$) predicts that this value should reflect the reward received following the same action on the previous trial. However, a model-based valuation strategy instead predicts that this previous reward effect should interact with whether the previous choice was followed by a common or rare transition.

We therefore examined BOLD activity at the start of trials in right ventral striatum (defined anatomically) as a function of the reward and transition on the previous trial. For reasons mentioned above, these signals did not form part of the previously described parametric RPE analyses. In order to isolate activity specifically related to the same action that had been learned about on the previous trial, we restricted our assessment to those trials in which the same action was chosen twice in a row (Morris et al., 2006). As seen in Figure 5A, there was a main effect of reward ($p < 0.005$), consistent with TD-like valuation. This, to our knowledge, is the first time that RPEs in BOLD signal have been directly shown to exhibit learning through an explicit dependence on previous-trial outcomes (Bayer and Glimcher, 2005). Across subjects, the interaction with the transition probability—the marker for model-based evaluation—was not significant ($p > 0.4$), but the size of the interaction per subject (taken as another neural index of the per-subject model-based effect) correlated with the behavioral index of model-based valuation ($p < 0.02$; Figure 5B). This last result further confirmed that striatal BOLD signal reflected model-based valuation to the extent that choice behavior did. Indeed, speaking to the consistency of the results, although the two neural estimates reported here for the extent of model-based valuation in the striatal BOLD signal (Figures 3F and 5B) were generated from different analytical approaches, and based on activity modeled at different time points within each trial, they significantly correlated with one another ($r^2 = 0.37$; $p < 0.01$).

DISCUSSION

We studied human choice behavior and BOLD activity in a two-stage decision task that allowed us to disambiguate model-based and model-free valuation strategies through their different claims about the effect of second-stage reinforcement on first-stage choices and BOLD signals. Here, ongoing adjustments in the values of second-stage actions extended the one-shot reward devaluation challenge often used in animal conditioning studies (Dickinson, 1985) and also the introduction of novel goals as in latent learning (Gläscher et al., 2010): they continually tested whether subjects prospectively adjusted their preferences for actions leading to a subsequent incentive (here, the second-stage state) when its value changed. Following Daw et al. (2005), we see such reasoning via sequential task structure as the defining feature that distinguishes model-based from model-free approaches to RL (although Hampton et al., 2006, and Bromberg-Martin et al., 2010 hold a somewhat different view: they associate model-based computation with learning nonsequential task structure as well).

We recently used a similar task in a complementary study (Gläscher et al., 2010) that minimized learning about the rewards (by reporting them explicitly and keeping them stable) to isolate learning about the state transition contingencies. Here, in

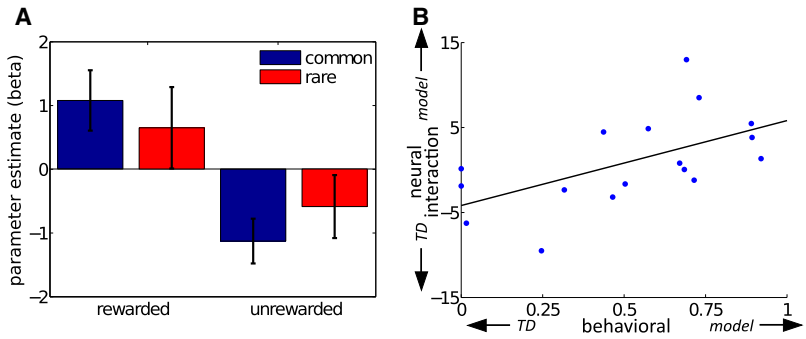

Figure 5. Factorial Analysis of BOLD Signal at Start of Trial, from Average Activity over an Anatomical Mask of Right Nucleus Accumbens

(A) Signal change (relative to mean) as a function of whether the choice on the previous trial was rewarded or unrewarded, and whether that occurred after a common or rare transition (compare Figure 2C). Error bars: 1 SEM.

(B) Scatterplot of the correlation, across subjects, between the contrast measuring the size of the interaction between reward and transition probability (an index of model-based valuation), and the weight given to model-based versus model-free valuations in explaining choice behavior. ($r^2 = 0.32$, $p = 0.017$).

contrast, we minimized transition learning (by partly instructing subjects) and introduced dynamic rewards to allow us to study the learning rules by which neural signals tracked them. This, in turn, allowed us to test an uninvestigated assumption of the analysis in the previous paper, i.e., the isolation of model-free value learning as expressed in the striatal PE.

Our previous computational theory of multiple RL systems in the brain (Daw et al., 2005) focused on a dynamic mechanism for trading off the reliance on model-based and model-free valuations based on their relative uncertainties. In the current task, the ever-changing rewards should keep the tradeoff roughly constant over time, allowing us to focus on the broader two-system structure of this theory. Rather than confronting the many (unknown) factors that determine the uncertainties of each system within each subject, we treated the balance between the two processes as exogenous, controlled by a constant free parameter (w) whose value we could estimate. Indeed, consistent with our intent, there was no significant trend (analyses not presented) toward progressive habit formation (Adams, 1982; Gläscher et al., 2010).

Nevertheless, consistent with findings from animal learning (Balleine and O'Doherty, 2010; Balleine et al., 2008; Dickinson, 1985; Dickinson and Balleine, 2002), we found clear evidence for both TD- and model-like valuations, suggesting that the brain employs a combination of both strategies. The standard view is that the two putative systems work separately and in parallel, a view reinforced by the strong association of the mesostriatal dopamine system with model-free RL, and the fact that, in animal studies, each system appears to operate relatively independently when brain areas associated with the other are lesioned (Killcross and Coutureau, 2003; Yin et al., 2004; Yin et al., 2005). Also consistent with this idea, previous work (Hampton et al., 2006, 2008) suggested that model-based influences on the vmPFC expected value signal, but did not test for additional model-free influences there, nor conversely, whether model-based influences also affected striatal RPEs. Here we found that even the signal most associated with model-free RL, the striatal RPE, reflects both types of valuation, combined in a way that matches their observed contributions to choice behavior. The finding that a similar result in vmPFC was weaker may reflect the fact that neural signaling there is, in some studies, better explained by a correlated variable, expected future value, and not RPE per se (Hare et al., 2008); residual error due to such a discrepancy could suppress effects there. However, in a sequential task these two quantities are closely related, thus, unlike Hare's, the present study was not designed to dissociate them.

Our ventral striatal finding invites a reevaluation of the standard account of RPE signaling in the brain, because it suggests that even a putative TD system does not exist in isolation from model-based valuation. One possibility about what might replace this account is suggested by contemplating an infelicity of the algorithm used here for data analysis. In order to reject the null hypothesis of purely model-free RPE signaling, we defined a generalized RPE with respect to model-based predictions as well. However, this augmented signal was nugatory, in the sense that model-based RPEs played no role in our account of choice behavior. Indeed, model-based learners do not rely on model-based RPEs: the learning problem they face—tracking state transition probabilities and immediate rewards rather than cumulative future rewards—demands different training signals (Gläscher et al., 2010).

This apparent mismatch encourages consideration of a hybrid of a different sort. We have so far examined theories in which model-based and model-free predictions compete directly to select actions (Daw et al., 2005). However, model-based and model-free RPEs could also usefully be integrated for training. For instance, consider the standard actor-critic account (Barto et al., 1983; Barto, 1995). This uses RPEs derived from model-free predictions (the critic) to reinforce action selection policies (the actor). Errors in model-based predictions, if available, could serve the same purpose. A model-free actor trained, in part, by such a model-based critic would, in effect, cache (Daw et al., 2005) or memorize the recommendations of a model-based planner, and could execute them subsequently without additional planning.

The computational literature on RL includes some related ideas in algorithms, such as prioritized sweeping (Moore and Atkeson, 1993), which caches the results of model-based evaluation (albeit without a model-free component), and Dyna (Johnson and Redish, 2005; Sutton, 1990), which trains a model-free algorithm (though offline) using simulated experiences generated from a world model. In neuroscience, various theories have been proposed in which a world model impacts the input to the model-free system (Bertin et al., 2007; Daw et al., 2006a; Doya, 1999; Doya et al., 2002). The architecture suggested here more closely resembles the "biased" learning hypothesized by Doll et al. (2009), according to which top-down information (there

provided by experimenter instructions rather than a learned world model) modifies the target of model-free RL. Outside the domain of learning, striatal BOLD responses are indeed affected by values communicated by instruction rather than experience (Fitzgerald et al., 2010; Tom et al., 2007) and also by emotional self-regulation (Delgado et al., 2008).

Further theoretical work is needed to characterize the different algorithms suggested by this general architecture. However, in general, by preserving the overall structure of parallel model-based and model-free systems—albeit systems that would exchange information at an earlier level—the proposal of a model-based critic would appear to remain consistent with the lesion data suggesting that the systems can function in isolation (Killcross and Coutureau, 2003; Yin et al., 2004, 2005), and with behavioral data demonstrating that distinct decision systems may have different properties and can be differentially engaged in different circumstances (Doeller and Burgess, 2008; Frank et al., 2007; Fu and Anderson, 2008). It also remains consistent with other fMRI studies (Doeller et al., 2008; Poldrack et al., 2001; Venkatraman et al., 2009) suggesting that overall activity in different brain systems associated with either system can modulate with time or circumstances, presumably in relation to the extent that either process is engaged.

Apart from training, a different use for model-based RPEs would be for online action evaluation and selection. In particular, Doya (1999) proposed that a world model could be used to predict the next state following a candidate action, and that a dopaminergic RPE with respect to that projected state could then be used to evaluate whether the action was worth taking. (A related scheme was suggested by McClure et al., 2003b; Montague et al., 1995, 1996.) RPEs for planning would appear to be categorically different in timing and content than RPEs for learning, in that the former are triggered by hypothetical state transitions and the latter by actual ones, as in the effects reported here. The Doya (1999) circuit also differs from a full model-based planner in that it envisions only a single step of model-based state lookahead; however, to test this limitation would require a task with longer sequences.

In the present study, as in most fMRI studies of RPEs, our effects focused on ventral striatum, and we did not see any correlates of the organization of striatum into components associated with different learning strategies as suggested by the rodent literature (Yin et al., 2004, 2005). Furthermore, although there is evidence suggesting that RPE effects in the ventral striatal BOLD signal reflect, at least in part, dopaminergic action there (Knutson and Gibbs, 2007; Pessiglione et al., 2006; Schönberg et al., 2010), the BOLD signal in striatum likely conflates multiple causes, including cortical input and local activity, and it is thus not possible to identify it uniquely with dopamine. Indeed, it is possible that, even if the effects attributed to our model-free RPE regressor are dopaminergic in origin, the residual effects captured by the model-based difference regressor in the same voxels arise from other sources. The questions raised by the present study thus invite resolution by testing a similar multistep task in animals using dopamine unit electrophysiology or voltammetry. In this respect, recent results by Bromberg-Martin et al. (2010) showing that, in a serial reversal task (albeit nonsequential), a dopaminergic RPE response is more sophisticated than a basic TD theory would predict, provide a tantalizing clue that our results might hold true of dopaminergic spiking as well.

Overall, by demonstrating that it is feasible to detect neural and behavioral signatures of both learning strategies, the present study opens the door to future within-subject studies targeted at manipulating and tracking the tradeoff dynamically, and thence, at uncovering the computational mechanisms and neural substrates for controlling it. Such metacontrol of decision systems is of particular practical importance, because, for instance, the compulsive nature of drug abuse has been proposed to result from aberrant expression of habitual control (Everitt and Robbins, 2005), and similar mechanisms have also, plausibly, been linked to other serious issues of self-control, including undersaving and overeating (Loewenstein and O'Donoghue, 2004).

EXPERIMENTAL PROCEDURES

Participants and Behavioral Task

Seventeen healthy adults (five female; mean age 25.8 years) participated in this study. All participants gave written informed consent, and the study was conducted in accordance with the guidelines of the local ethics committee.

The task consisted of 201 trials, in three blocks of 67, separated by breaks. The events in the trial are sketched in Figure 1A. Each trial consisted of two stages. In the first stage, subjects used an MRI compatible button box to choose between two options, represented by Tibetan characters in colored boxes. If subjects failed to enter a choice within 2 s, the trial was aborted. The chosen option rose to the top of the screen, while the option not chosen faded and disappeared. At the second stage, subjects were presented with either of two more choices between two options ("states"), and entered another choice. The second choice was rewarded with money (depicted by a pound coin, though subjects were paid 20% of this amount), or not (depicted by a zero). Trials were separated by an intertrial interval of randomized length, on average about 1 TR.

Which second-stage state was presented depended, probabilistically, on the first-stage choice, according to the transition scheme shown in Figure 1B. The assignment of colors to states was counterbalanced across subjects, and the two options at each state were permuted pseudorandomly between left and right from trial to trial. Each bottom-stage option was rewarded according to a probability associated with that option. In order to encourage ongoing learning, these reward probabilities were diffused at each trial by adding independent Gaussian noise (mean 0, SD 0.025), with reflecting boundaries at 0.25 and 0.75.

In a computerized training session prior to the fMRI task, subjects were instructed that the reward probabilities would change, but those controlling the transitions from the first to the second stage would remain fixed. They were also instructed about the overall structure of the transition matrix, specifically, that each first-stage option was primarily associated with one or the other of the second-stage states, but not which one. Prior to the scanning session, to familiarize themselves with the structure of the task, subjects played 50 trials on a practice task using a different stimulus set.

Behavioral Analyses

We first conducted a logistic regression in which the dependent variable was the first-stage choice (coded as stay versus switch), and the explanatory variables were the reward received on the previous trial, a binary indicator variable indicating whether the previous trial's transition was common or rare, and the interaction of the two. We took all coefficients as random effects across subjects, and estimated this multilevel regression using the lme4 linear mixed effects package (Bates and Maechler, 2010) in the R statistical language (R Development Core Team, 2010). We also extracted posterior effect size estimates (conditional on the estimated population-level prior) and confidence intervals from the posterior covariance for each of the individuals from this fit. The predictions in Figures 2A and 2B are derived from simulations of SARSA(1)

and model-based algorithms (below), using the parameters best fit to the subjects' data within each class of algorithm.

Computational Model of Behavior

In a second set of analyses, we fit choice behavior to an algorithm that is similar to the hybrid algorithm of Gläscher et al. (2010). In particular, it learned action values via both model-based RL (explicit computation of Bellman's equation) and by model-free SARSA(λ) TD learning (Rummery and Niranjan, 1994), and assumed choices were driven by the weighted combination of these two valuations. The relative weighting was controlled by a free parameter w, which we assumed to be constant across trials. We also computed TD RPEs with respect to both the model-free and model-based valuations, and, for fMRI analysis, defined a difference regressor as the difference between them. Full equations are given in Supplemental Experimental Procedures.

Behavioral Estimation

For behavioral analysis, we estimated the free parameters of the algorithm separately for each subject, to maximize the log-likelihood of the data (from the log of Equation 2 summed over all trials; see Supplemental Information), for the choices actually made conditioned on the states and rewards previously encountered. We constrained the learning rates to lie between zero and one, but allowed λ and w (which also nominally range between zero and one) to float arbitrarily beyond these boundaries, so as to make meaningful the tests of whether the median estimates were different from the nominal boundaries across the population.

For classical model comparison, we repeated this procedure for the nested subcases, and tested the null hypothesis of the parametric restriction (either individually per subject or for likelihoods aggregated over the population) using likelihood ratio tests. For Bayesian model comparison, we computed a Laplace approximation to the model evidence (MacKay, 2003) integrating out the free parameters; this analysis requires a prior over the parameters, which we took to be Beta(1.1,1.1) for the learning rates, λ and w, Normal(0,1) for p, and Gamma(1.2,5) for the softmax temperatures, selected so as to be uninformative over the parameter ranges we have seen in previous studies, and to roll off smoothly at parametric boundaries. We also fit the model of Stephan et al. (2009), which takes model identity as a random effect, by submitting the Laplace-approximated log model evidences to the spm_BMS routine from SPM8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/).

Thus, we performed all behavioral analyses assuming the parameters (and in some cases the model identity) to be random effects across subjects. However, to generate regressors for neural analyses on a common scale, we refit the algorithm to the choices, taking only w as a random effect, instantiated once per subject, and assuming common values for the other parameters. This is because in these sorts of algorithms, noise and variation in parameter estimates from subject to subject results, effectively, in a rescaling of regressors between subjects, which suppresses the significance of neural effects in a subsequent second-level fMRI analysis, producing poor results (Daw, in press; Daw et al., 2006b; Gershman et al., 2009; Schönberg et al., 2007, 2010).

fMRI Procedures

Functional imaging was conducted using a 1.5T Siemens Sonata MRI scanner to acquire gradient echo T2*-weighted echo-planar images (EPI) with BOLD contrast. Standard preprocessing was performed; see Supplemental Experimental Procedures for full details of preprocessing and acquisition.

fMRI Analysis

The fMRI analysis was based around the time series of model-free and model-based RPEs as generated from the simulation of the model over each subject's experiences. We defined two parametric regressors—the model-free RPE, and the difference between the model-free and model-based RPEs. The latter regressor characterizes how net BOLD activity would differ if it were correlated with model-based RPEs or any weighted mixture of both. For each trial, the RPE time series were entered as parametric regressors modulating impulse events at the second-stage onset and reward receipt. To test the correspondence between behavioral and neural estimates of the model-based effect, we also included the per-subject estimate of the model-based effect (w, above) from the behavioral fits as a second-level covariate for the difference regressor. A full description of the analysis is given in Supplemental Experimental Procedures.

For display purposes, we render activations at an uncorrected threshold of $p < 0.001$ (except relaxing this in one case to $p < 0.005$), overlaid on the average of subjects' normalized structural images. For all reported statistics, we subjected these uncorrected maps to cluster-level correction for family-wise error due to multiple comparisons over the whole brain, or, in a few cases (noted specifically), over a small volume defined by an anatomical mask of bilateral nucleus accumbens. This mask was hand-drawn on the subject-averaged structural image, according to the guidelines of Breiter et al. (Ballmaier et al., 2004; Breiter et al., 1997; Schönberg et al., 2010)—notably, defining the nucleus' superior border by a line connecting the most ventral point of the lateral ventricle to the most ventral point of the internal capsule at the level of the putamen. Conjunction inference was by the minimum t statistic (Nichols et al., 2005) using the conjunction null hypothesis. The difference regressor was orthogonalized against the RPE regressor, so that up to minor correlation that can be reintroduced by whitening and filtering, it captured only residual variation in BOLD activity not otherwise explained by the model-free RPE. However, note that conjunction inference via the minimum t statistic is valid even when the conjoined contrasts are not independent (Nichols et al., 2005).

ROI Analyses

We also used the right-hemisphere portion of the mask of nucleus accumbens (right being the side on which we have previously observed stronger RPE activity; e.g., Daw et al., 2006b; Wittmann et al., 2008) to define the ROI for two analyses conducted with the MarsBaR ROI toolbox (Brett et al., 2002). First, average activity from the region was extracted and subjected to the same analysis as described above, to produce Figure 3F. Second, the activity from the region was subject to a second regression analysis using a different design, which tagged the first-stage onset of each trial with an impulse regressor of one of five types: switches (trials on which the opposite first-stage choice from the one on the previous trial was made) and stays (four types of events modeling all combinations of the factors reward versus nonreward and common versus rare transition in the previous trial). An additional nuisance regressor was included at the time of outcomes. Per-subject effect sizes for the four stay regressors were subject to a 2 × 2 repeated-measure ANOVA, and, additionally, the value for each subject of the contrast measuring the interaction of the two factors ([reward/common minus nonreward/common] minus [reward/rare minus nonreward/rare]) was correlated with the weight given to model-based values (the estimated parameter w) from the behavioral fit.

SUPPLEMENTAL INFORMATION

Supplemental Information for this article includes Supplemental Experimental Procedures and can be found with this article online at doi:10.1016/j.neuron.2011.02.027.

ACKNOWLEDGMENTS

The authors are grateful to Yael Niv, Dylan Simon, Aaron Bornstein, Seth Madlon-Kay, Bianca Wittmann, Bernard Balleine, Jan Gläscher, and John O'Doherty for helpful conversations and advice. This work was in part supported by a McKnight Scholar Award (N.D.), NIMH grant 1R01MH087882-01, part of the CRCNS program (N.D.), a NARSAD Young Investigator Award (N.D.), the Gatsby Charitable Foundation (P.D.), and a Wellcome Trust Programme Grant to R.J.D.

Accepted: January 10, 2011
Published: March 23, 2011

REFERENCES

Adams, C.D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. Q. J. Exp. Psychol. *33b*, 77–98.

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

Balleine, B.W., Daw, N.D., and O'Doherty, J.P. (2008). Multiple forms of value learning and the function of dopamine. In Neuroeconomics: Decision Making and the Brain, P.W. Glimcher, C. Camerer, R.A. Poldrack, and E. Fehr, eds. (San Diego, CA: Academic Press).

Ballmaier, M., Toga, A.W., Siddarth, P., Blanton, R.E., Levitt, J.G., Lee, M., and Caplan, R. (2004). Thought disorder and nucleus accumbens in childhood: A structural MRI study. Psychiatry Res. *130*, 43–55.

Barto, A.G. (1995). Adaptive Critics and the Basal Ganglia. In Models of Information Processing in the Basal Ganglia, J.L. Davis, J.C. Houk, and D.G. Beiser, eds. (Cambridge, MA: MIT Press), pp. 215–232.

Barto, A., Sutton, R., and Anderson, C. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Trans. Syst. Man Cybern. *13*, 834–846.

Bates, D., and Maechler, M. (2010). lme4: Linear mixed effects models using S4 classes. R package version 0.999375 33. http://CRAN.R-project.org/package=lme4.

Bayer, H.M., and Glimcher, P.W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron *47*, 129–141.

Berns, G.S., McClure, S.M., Pagnoni, G., and Montague, P.R. (2001). Predictability modulates human brain response to reward. J. Neurosci. *21*, 2793–2798.

Bertin, M., Schweighofer, N., and Doya, K. (2007). Multiple model-based reinforcement learning explains dopamine neuronal activity. Neural Netw. *20*, 668–675.

Breiter, H.C., Gollub, R.L., Weisskoff, R.M., Kennedy, D.N., Makris, N., Berke, J.D., Goodman, J.M., Kantor, H.L., Gastfriend, D.R., Riorden, J.P., et al. (1997). Acute effects of cocaine on human brain activity and emotion. Neuron *19*, 591–611.

Brett, M., Anton, J.-L., Valabregue, R., and Poline, J.-B. (2002). Region of interest analysis using an SPM toolbox. In 8th International Conference on Functional Mapping of the Human Brain (Sendai, Japan).

Bromberg-Martin, E., Matsumoto, M., Hong, S., and Hikosaka, O. (2010). A pallidus-habenula-dopamine pathway signals inferred stimulus values. J. Neurophysiol. *104*, 1068–1076.

Daw, N.D. (in press). Trial-by-trial data analysis using computational models. In Affect, Learning and Decision Making, Attention and Performance XXIII, E.A. Phelps, T.W. Robbins, and M. Delgado, eds. (New York: Oxford University Press).

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

Daw, N.D., Courville, A.C., and Touretzky, D.S. (2006a). Representation and timing in theories of the dopamine system. Neural Comput. *18*, 1637–1677.

Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., and Dolan, R.J. (2006b). Cortical substrates for exploratory decisions in humans. Nature *441*, 876–879.

Delgado, M.R., Nystrom, L.E., Fissell, C., Noll, D.C., and Fiez, J.A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. J. Neurophysiol. *84*, 3072–3077.

Delgado, M.R., Gillis, M.M., and Phelps, E.A. (2008). Regulating the expectation of reward via cognitive strategies. Nat. Neurosci. *11*, 880–881.

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. Philos. Trans. R. Soc. Lond. B Biol. Sci. *308*, 67–78.

Dickinson, A., and Balleine, B. (2002). The role of learning in the operation of motivational systems. Stevens' Handbook of Experimental Psychology, Third Edition, Vol.3: Learning, Motivation, and Emotion, H. Paschler and R. Gallistel, eds. (New York: John Wiley & Sons), pp. 497–534.

Doeller, C.F., and Burgess, N. (2008). Distinct error-correcting and incidental learning of location relative to landmarks and boundaries. Proc. Natl. Acad. Sci. USA *105*, 5909–5914.

Doeller, C.F., King, J.A., and Burgess, N. (2008). Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. Proc. Natl. Acad. Sci. USA *105*, 5915–5920.

Doll, B.B., Jacobs, W.J., Sanfey, A.G., and Frank, M.J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. Brain Res. *1299*, 74–94.

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

Doya, K., Samejima, K., Katagiri, K., and Kawato, M. (2002). Multiple model-based reinforcement learning. Neural Comput. *14*, 1347–1369.

Everitt, B.J., and Robbins, T.W. (2005). Neural systems of reinforcement for drug addiction: From actions to habits to compulsion. Nat. Neurosci. *8*, 1481–1489.

Fitzgerald, T.H., Seymour, B., Bach, D.R., and Dolan, R.J. (2010). Differentiable neural substrates for learned and described value and risk. Curr. Biol. *20*, 1823–1829.

Foster, D.J., and Wilson, M.A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. Nature *440*, 680–683.

Frank, M.J., Moustafa, A.A., Haughey, H.M., Curran, T., and Hutchison, K.E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proc. Natl. Acad. Sci. USA *104*, 16311–16316.

Friston, K.J., Worsley, K.J., Frackowiak, R.S.J., Mazziotta, J.C., and Evans, A.C. (1993). Assessing the significance of focal activations using their spatial extent. Hum. Brain Mapp. *1*, 210–220.

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

Fu, W.T., and Anderson, J.R. (2008). Solving the credit assignment problem: Explicit and implicit learning of action sequences with probabilistic outcomes. Psychol. Res. *72*, 321–330.

Gershman, S.J., Pesaran, B., and Daw, N.D. (2009). Human reinforcement learning subdivides structured action spaces by learning effector-specific values. J. Neurosci. *29*, 13524–13531.

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

Hampton, A.N., Bossaerts, P., and O'Doherty, J.P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J. Neurosci. *26*, 8360–8367.

Hampton, A.N., Bossaerts, P., and O'Doherty, J.P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc. Natl. Acad. Sci. USA *105*, 6741–6746.

Hare, T.A., O'Doherty, J., Camerer, C.F., Schultz, W., and Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. J. Neurosci. *28*, 5623–5630.

Hassabis, D., and Maguire, E.A. (2007). Deconstructing episodic memory with construction. Trends Cogn. Sci. *11*, 299–306.

Holmes, A., and Friston, K. (1998). Generalisability, random effects and population inference. Neuroimage *7*, S754.

Ito, M., and Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. J. Neurosci. *29*, 9861–9874.

Johnson, A., and Redish, A.D. (2005). Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. Neural Netw. *18*, 1163–1171.

Johnson, A., and Redish, A.D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. J. Neurosci. *27*, 12176–12189.

Kable, J.W., and Glimcher, P.W. (2007). The neural correlates of subjective value during intertemporal choice. Nat. Neurosci. *10*, 1625–1633.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. Am. Psychol. *58*, 697–720.

Killcross, S., and Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. Cereb. Cortex *13*, 400–408.

Kim, H., Sul, J.H., Huh, N., Lee, D., and Jung, M.W. (2009). Role of striatum in updating values of chosen actions. J. Neurosci. *29*, 14701–14712.

Knutson, B., and Gibbs, S.E. (2007). Linking nucleus accumbens dopamine and blood oxygenation. Psychopharmacology (Berl.) *191*, 813–822.

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

Knutson, B., Rick, S., Wimmer, G.E., Prelec, D., and Loewenstein, G. (2007). Neural predictors of purchases. Neuron *53*, 147–156.

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

Loewenstein, G., and O'Donoghue, T. (2004). Animal spirits: Affective and deliberative processes in economic behavior. Working Paper 04-14, Center for Analytic Economics, Cornell University.

Lohrenz, T., McCabe, K., Camerer, C.F., and Montague, P.R. (2007). Neural signature of fictive learning signals in a sequential investment task. Proc. Natl. Acad. Sci. USA *104*, 9493–9498.

MacKay, D.J.C. (2003). Information Theory, Inference, and Learning Algorithms (New York: Cambridge University Press).

Maia, T.V. (2010). Two-factor theory, the actor-critic model, and conditioned avoidance. Learn. Behav. *38*, 50–67.

McClure, S.M., Berns, G.S., and Montague, P.R. (2003a). Temporal prediction errors in a passive learning task activate human striatum. Neuron *38*, 339–346.

McClure, S.M., Daw, N.D., and Montague, P.R. (2003b). A computational substrate for incentive salience. Trends Neurosci. *26*, 423–428.

Montague, P.R., Dayan, P., Person, C., and Sejnowski, T.J. (1995). Bee foraging in uncertain environments using predictive Hebbian learning. Nature *377*, 725–728.

Montague, P.R., Dayan, P., and Sejnowski, T.J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. *16*, 1936–1947.

Moore, A.W., and Atkeson, C.G. (1993). Prioritized sweeping: Reinforcement learning with less data and less time. Mach. Learn. *13*, 103–130.

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

Nichols, T., Brett, M., Andersson, J., Wager, T., and Poline, J.B. (2005). Valid conjunction inference with the minimum statistic. Neuroimage *25*, 653–660.

Niv, Y., Joel, D., and Dayan, P. (2006). A normative perspective on motivation. Trends Cogn. Sci. *10*, 375–381.

O'Doherty, J.P. (2004). Reward representations and reward-related learning in the human brain: Insights from neuroimaging. Curr. Opin. Neurobiol. *14*, 769–776.

O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H., and Dolan, R.J. (2003). Temporal difference models and reward-related learning in the human brain. Neuron *38*, 329–337.

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

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., and Frith, C.D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature *442*, 1042–1045.

Peters, J., and Büchel, C. (2009). Overlapping and distinct neural systems code for subjective value during intertemporal and risky decision making. J. Neurosci. *29*, 15727–15734.

Plassmann, H., O'Doherty, J., and Rangel, A. (2007). Orbitofrontal cortex encodes willingness to pay in everyday economic transactions. J. Neurosci. *27*, 9984–9988.

Poldrack, R.A., Clark, J., Paré-Blagoev, E.J., Shohamy, D., Creso Moyano, J., Myers, C., and Gluck, M.A. (2001). Interactive memory systems in the human brain. Nature *414*, 546–550.

Preuschoff, K., Bossaerts, P., and Quartz, S.R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. Neuron *51*, 381–390.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Rangel, A., Camerer, C., and Montague, P.R. (2008). A framework for studying the neurobiology of value-based decision making. Nat. Rev. Neurosci. *9*, 545–556.

Redish, A.D., Jensen, S., and Johnson, A. (2008). A unified framework for addiction: Vulnerabilities in the decision process. Behav. Brain Sci. *31*, 415–437, discussion 437–487.

Rummery, G., and Niranjan, M. (1994). On-line Q-learning using connectionist systems. Technical Report CUED/F-INFENG/TR 166, Engineering Department (Cambridge University).

Schacter, D.L., Addis, D.R., and Buckner, R.L. (2007). Remembering the past to imagine the future: The prospective brain. Nat. Rev. Neurosci. *8*, 657–661.

Schönberg, T., Daw, N.D., Joel, D., and O'Doherty, J.P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. J. Neurosci. *27*, 12860–12867.

Schönberg, T., O'Doherty, J.P., Joel, D., Inzelberg, R., Segev, Y., and Daw, N.D. (2010). 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 *49*, 772–781.

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

Seymour, B., O'Doherty, J.P., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., Friston, K.J., and Frackowiak, R.S. (2004). Temporal difference models describe higher-order learning in humans. Nature *429*, 664–667.

Sloman, S.A. (1996). The empirical case for two systems of reasoning. Psychol. Bull. *119*, 3–22.

Stephan, K.E., Penny, W.D., Daunizeau, J., Moran, R.J., and Friston, K.J. (2009). Bayesian model selection for group studies. Neuroimage *46*, 1004–1017.

Suri, R.E., and Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. Neuroscience *91*, 871–890.

Sutton, R. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In Proceedings of the Seventh International Conference on Machine Learning (San Diego, California).

Tanaka, S.C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., and Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. Nat. Neurosci. *7*, 887–893.

Thorndike, E.L. (1911). Animal Intelligence; Experimental Studies (New York: The Macmillan Company).

Tolman, E.C. (1948). Cognitive maps in rats and men. Psychol. Rev. *55*, 189–208.

Tom, S.M., Fox, C.R., Trepel, C., and Poldrack, R.A. (2007). The neural basis of loss aversion in decision-making under risk. Science *315*, 515–518.

Valentin, V.V., Dickinson, A., and O'Doherty, J.P. (2007). Determining the neural substrates of goal-directed learning in the human brain. J. Neurosci. *27*, 4019–4026.

Venkatraman, V., Payne, J.W., Bettman, J.R., Luce, M.F., and Huettel, S.A. (2009). Separate neural mechanisms underlie choices and strategic preferences in risky decision making. Neuron *62*, 593–602.

Wittmann, B.C., Daw, N.D., Seymour, B., and Dolan, R.J. (2008). Striatal activity underlies novelty-based choice in humans. Neuron *58*, 967–973.

Yin, H.H., Knowlton, B.J., and Balleine, B.W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. Eur. J. Neurosci. *19*, 181–189.

Yin, H.H., Ostlund, S.B., Knowlton, B.J., and Balleine, B.W. (2005). The role of the dorsomedial striatum in instrumental conditioning. Eur. J. Neurosci. *22*, 513–523.

Supplemental material for "Model-based influences on humans' choices and striatal prediction errors"

Nathaniel D. Daw, Samuel J. Gershman, Ben Seymour, Peter Dayan & Raymond J. Dolan

Supplemental experimental procedures

Computational model of behavior

The task consists of three states (first stage: s_A; second stage: s_B and s_C), each with two actions (a_A and a_B). The goal of both the model-based and model-free subcomponents of the algorithm is to learn a state-action value function $Q(s,a)$ mapping each state-action pair to its expected future value. On trial t, we denote the first-stage state (always s_A) by $s_{1,t}$, the second-stage state by $s_{2,t}$, the first- and second-stage actions by $a_{1,t}$ and $a_{2,t}$, and the first- and second-stage rewards as $r_{1,t}$ (always zero) and $r_{2,t}$.

The model free algorithm was SARSA(λ) temporal difference learning (Rummery and Niranjan, 1994). At each stage i of each trial t, the value for the visited state-action pair was updated according to:

$$Q_{TD}(s_{i,t}, a_{i,t}) = Q_{TD}(s_{i,t}, a_{i,t}) + \alpha_i \delta_{i,t}$$

where

$$\delta_{i,t} = r_{i,t} + Q_{TD}(s_{i+1,t}, a_{i+1,t}) - Q_{TD}(s_{i,t}, a_{i,t}) \quad [1]$$

and α_i are free learning-rate parameters. (We allow different learning rates α_1 and α_2 for the two task stages, to ensure our primary analyses of top-level effects are not affected by any potential difference in learning or behavior between the stages. Such effects might arise if there were differences in learning from state transitions vs rewards, and because any second-level state/action is sampled less frequently than the top-level ones.) Note that, for the first-stage choice, $r_{1,t} = 0$ and the RPE is instead driven by the second-stage value, $Q_{TD}(s_{2,t}, a_{2,t})$; conversely at the second stage, we define $Q_{TD}(s_{3,t}, a_{3,t}) = 0$, since there is no further value in the trial apart from the immediate reward $r_{2,t}$. Since this task has only two stages per trial, the only effect of the eligibility parameter λ (Sutton and Barto, 1998) is, at the end of each trial, to modulate an additional stage-skipping update of the first-stage action by the second-stage RPE, $Q_{TD}(s_{1,t}, a_{1,t}) = Q_{TD}(s_{1,t}, a_{1,t}) + \alpha_1 \lambda \delta_{2,t}$. Note that this model assumes that eligibility traces are cleared between episodes (i.e., that eligibility does not carry over from trial to trial), which appears to be a reasonable simplification since eligibility carryover would be inconsistent with the episodic structure of the task, about which subjects were instructed; though see Walton et al. (2010).

In general, a model-based RL algorithm works by learning a transition function (mapping state-action pairs to a probability distribution over the subsequent state), and immediate reward values for each state, then computing cumulative state-action values by iterative expectation over these. Specialized to

the structure of the current task, this amounts to, first, simply deciding which first-stage action maps to which second-stage state (since subjects were instructed that this was the structure of the transition contingencies), and second, learning immediate reward values for each of the second-stage actions (the immediate rewards at the first stage being always zero).

We characterized transition learning by assuming subjects simply chose between the two possibilities: $P(s_B|s_A,a_A)=0.7, P(s_C|s_A,a_B)=0.7$, or, vice versa $P(s_B|s_A,a_A)=0.3, P(s_C|s_A,a_B)=0.3$ (with $P(s_C|s_A,a_A)=1-P(s_B|s_A,a_A)$ and $P(s_B|s_B,a_A)=1-P(s_C|s_A,a_B)$, according to whether more transitions had so far occurred to s_B following a_A plus s_C following a_B, or, vice versa, to s_C following a_A plus s_B following a_B. (In analyses not reported here, we verified that this scheme, which settles on the true transition matrix after the first few trials and is consistent with subjects' instructions, fit their choices better than traditional incremental learning schemes for estimating transition matrices. The specific values 0.7/0.3 are chosen without loss of generality; if these are changed, other free parameters of the algorithm will rescale to give the same overall choice likelihood.)

At the second-stage (the only one where immediate rewards were offered), the problem of learning immediate rewards is equivalent to that for TD above, since $Q_{TD}(s_{2,t},a_{2,t})$ is just an estimate of the immediate reward $r_{2,t}$; with no further stages to anticipate, the SARSA learning rule reduces to a delta-rule for predicting the immediate reward. Thus the two approaches coincide at the second stage, and we define $Q_{MB}=Q_{TD}$ at those states.

Next, using Bellman's equation, we define the model-based values of the first level actions as

$$Q_{MB}(s_A,a_j)=P(s_B|s_A,a_j)\max_{a\in\{a_A,a_B\}}Q_{TD}(s_B,a)+P(s_C|s_A,a_j)\max_{a\in\{a_A,a_B\}}Q_{TD}(s_C,a)$$

and assume these are recomputed at each trial from the current estimates of the transition probabilities and rewards.

Finally, to connect the values to choices, we define net action values at the first stage as the weighted sum of model-based and model-free values $Q_{net}(s_A,a_j)=wQ_{MB}(s_A,a_j)+(1-w)Q_{TD}(s_A,a_j)$ where *w* is a weighting parameter. At the second stage, $Q_{net}=Q_{MB}=Q_{TD}$. We then assume the probability of a choice is softmax in Q_{net}:

$$P(a_{i,t}=a|s_{i,t})=\frac{\exp(\beta_i[Q_{net}(s_{i,t},a)+p\cdot\mathrm{rep}(a)])}{\sum_{a'}\exp(\beta_i[Q_{net}(s_{i,t},a')+p\cdot\mathrm{rep}(a')])} \quad [2]$$

Here, the free inverse temperature parameters β_i control how deterministic are the choices, and we allow β_1 and β_2 to differ between the stages. (This captures any differences in choice reliability between the stages; note that this also renders redundant a time-discount parameter.) The indicator function $\mathrm{rep}(a)$ is defined as 1 if *a* is a top-stage action and is the same one as was chosen on the previous trial, zero otherwise. Together with the free parameter *p*, this captures first-order perseveration ($p>0$) or switching ($p<0$) in the first-stage choices (Lau and Glimcher, 2005; also visible in Figure 2c). We do not include such autocorrelation for the second-stage choices, simply because (since different second-level

states are visited from trial to trial) choice repetition at the second stage is less likely to play a large role, and it is also less clear how best to define it.

In total, the algorithm contains 7 free parameters (β_1, β_2, α_1, α_2, λ, p, w), and nests pure model-based (w = 1, with arbitrary α_1 and λ) and model-free (w = 0) learning as special cases.

For neural analysis, we defined a generalized version of Equation 1, measuring RPEs with respect to net model-based/model-free values Q_{net}:

$$\delta_{net,i,t} = r_{i,t} + Q_{net}(s_{i+1,t}, a_{i+1,t}) - Q_{net}(s_{i,t}, a_{i,t}) \quad [3]$$

and took its partial derivative with respect to the parameter w mixing Q_{TD} and Q_{MB} into Q_{net}, which we refer to as the "difference regresssor" since it is the difference between δ_{net} computed for w=1 and w=0.

fMRI procedures

Functional imaging was conducted using a 1.5T Siemens Sonata MRI scanner to acquire gradient echo T2*-weighted echo-planar images (EPI) with blood oxygenation level dependent (BOLD) contrast. We employed a special pulse sequence designed to optimize functional sensitivity in OFC (Deichmann et al., 2003). This consisted of tilted acquisition in an oblique orientation at 30 degrees to the AC-PC line, as well as application of a preparation pulse with a duration of 1ms and amplitude of -2mT/m in the selection direction. The sequence enabled 36 axial slices of 3mm thickness and 3mm in-plane resolution to be acquired with a repetition time (TR) of 3.24s. Coverage was obtained from the base of the orbitofrontal cortex and medial temporal lobes to the superior border of the dorsal anterior cingulate cortex. Participants were placed in a light head restraint within the scanner to limit head movement during acquisition. A field map was also recorded for distortion correction of the acquired EPI images, using a double echo FLASH sequence (64 oblique transverse slices, slice thickness = 2 mm, gap between slices = 1 mm, TR = 1170 ms, $\alpha = 90°$, short TE = 10 ms, long TE = 14.76 ms, BW = 260 Hz/pixel, PE direction anterior–posterior, FOV = 192×192 mm^2, matrix size 64 × 64, flow compensation). A T1-weighted structural image was also acquired for each subject.

Preprocessing and data analysis were performed using Statistical Parametric Mapping software implemented in Matlab (SPM5; Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK; and SPM8 for final results visualization and multiple comparison correction). Using the FieldMap toolbox (Hutton et al., 2002; Hutton et al., 2004), field maps were estimated from the phase difference between the images acquired at the short and long TEs in the FLASH sequence. The EPI images were corrected for subject motion by realigning them to the first volume, and simultaneously corrected for geometric distortion based on the field map and the interaction of distortion with motion (SPM5 "realign and unwarp"; Andersson et al., 2001; Hutton et al., 2002; Hutton et al., 2004). EPI images were then spatially normalized to the Montreal Neurological Institute template by warping the subject's anatomical image to an SPM segmentation template (SPM5 "segment and normalize") and applying these parameters to the functional images, resampled into 2x2x2 mm sized voxels, and smoothed using an 8 mm Gaussian kernel.

For statistical analysis, the data were scaled voxel-by-voxel onto their global mean and high-pass-filtered using a filter width of 128 secs.

fMRI analysis

The fMRI analysis was based around the timeseries of RPEs as generated from the simulation of the model over each subject's experiences. Note that as defined by Equation 1 above, this error is nonzero at two timepoints: the onset of the second stage, when $\delta_{1,t}$ is realized, and at the reward receipt, when $\delta_{2,t}$ is realized. These two RPEs ostensibly train the values that drive the two choices in the task. In general, RPE can also be defined at a third point, the start of the trial (the onset of stage 1; see Daw et al., 2006; Schonberg et al., 2007; Schonberg et al., 2010), but this is more difficult to define accurately enough to analyze parametrically since it depends on the value expectation prior to trial onset, a quantity which is not assessed behaviorally. Therefore, we do not include this timepoint in our parametric analysis of RPE effects (instead defining nuisance regressors to control variance there), but we separately subject activity at this timepoint to a complementary, factorial analysis as a relatively independent test of our conclusions (see ROI analyses in main text Experimental Procedures).

We included the RPE as a parametric regressor modulating impulse events at the second-stage onset and reward receipt. The regressor, from Equation 1, corresponds to the generalized model-based/model-free RPE (Equation 3) computed for the mixing parameter *w* = 0. We included an additional parametric regressor, defined at the same timepoints, containing the partial derivative of this timeseries with respect to *w*. Intuitively, the partial derivative captures how the RPE would change if it were computed according to a different value of *w* (Friston et al., 1998); in this case, it is just the difference between the RPEs computed with respect to model-based and model-free action values. Since this difference is zero at outcome time, but nonzero at the second-stage onset, to exclude the possibility that the difference effect would be confounded by a simple difference in average striatal activity between these two events, we mean-corrected the difference regressor's values at the choicepoint to zero mean within-subject, and also included an additional nuisance onset at the time of outcome reveal so as to capture any difference in mean activity between the choice and outcome events. We included another nuisance onset at thefirst-stage trial onset, modulated by two additional parametric regressors, also treated as nuisance effects: $P\left(a_{1,t}|s_A\right)$ (from Equation 2), as a normalized measure of the first-stage action value (Daw et al., 2006), and its partial derivative with respect to *w*.

These regressors were then convolved with the canonical hemodynamic response function, and entered into a regression analysis against each subject's fMRI data using SPM. The 6 scan-to-scan motion parameters produced during realignment were included as additional nuisance regressors in the SPM analysis to account for residual effects of scan to scan motion. To enable inference at the group level, the coefficient estimates for the RPE and difference regressors from each individual subject were taken to the second-level to allow random effects group statistics to be computed. To test the correspondence between behavioral and neural estimates of the model-based effect, we also included the per-subject estimate of the model-based effect (*w*, above) from the behavioral fits as a second-level covariate for the difference regressor.

References

Andersson, J., Hutton, C., Ashburner, J., Turner, R., and Friston, K. (2001). Modeling geometric deformations in EPI time series. Neuroimage *13*, 903-919.

Daw, N.D., O'Doherty, J.P., Dayan, P., Seymour, B., and Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. Nature *441*, 876-879.

Deichmann, R., Gottfried, J., Hutton, C., and Turner, R. (2003). Optimized EPI for fMRI studies of the orbitofrontal cortex. Neuroimage *19*, 430-441.

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

Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., and Turner, R. (2002). Image distortion correction in fMRI: A quantitative evaluation. Neuroimage *16*, 217-240.

Hutton, C., Deichmann, R., Turner, R., and Andersson, J.L.R. (2004). Combined correction for geometric distortion and its interaction with head motion in fMRI. In ISMRM 12 (Kyoto, Japan)).

Lau, B., and Glimcher, P.W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. J Exp Anal Behav *84*, 555-579.

Rummery, G., and Niranjan, M. (1994). On-line Q-learning using connectionist systems.

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

Schonberg, T., O'Doherty, J., Joel, D., Inzelberg, R., Segev, Y., and Daw, N. (2010). 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 *49*, 772-781.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement Learning: An Introduction (MIT Press).

Walton, M.E., Behrens, T.E., Buckley, M.J., Rudebeck, P.H., and Rushworth, M.F. (2010). Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. Neuron *65*, 927-939.