Supplementary Methods

Subjects and behavioral task

14 right-handed human subjects participated in the task. The subjects were pre-assessed to exclude those with a prior history of neurological or psychiatric illness. All gave informed consent, and the study was approved by the local ethics committee.

The task consisted of two sessions of 150 trials each, separated by a short break. On each trial, subjects were presented with pictures of four different colored slot machines (visible on a screen reflected in a head coil mirror), and selected one using a button box with their right hand (see **Fig.** 1a). Subjects had a maximum of 1.5 seconds in which to make their choice; if no choice was entered during that interval, a large red X was displayed for 4.2 seconds to signal an invalid missed trial (after which a new trial was triggered). Subjects usually responded well before the timeout, with a mean response time of ~430msecs Overall there were very few missed trials (typically 1 or 2 per subject). On valid trials, the chosen slot machine was animated and, three seconds later, the number of points earned was displayed. These points were displayed for 1 second and then the screen was cleared. The trial sequence ended 6 seconds after trial onset, followed by a jittered intertrial interval using a discrete approximation of a Poisson distribution with a mean of 2 seconds, before the next trial was triggered.

The payoff for choosing the *i*th slot machine on trial *t* was between 1 and 100 points, drawn from a Gaussian distribution (standard deviation $\sigma_o = 4$) around a mean $\mu_{i,t}$ and rounded to the nearest integer. At each timestep, the means diffused in a decaying Gaussian random walk, with $\mu_{i,t+1} = \lambda\mu_{i,t} + (1 - \lambda)\theta + v$ for each i. The decay parameter λ was 0.9836, the decay center θ was 50, and the diffusion noise v was zero-mean Gaussian (standard deviation $\sigma_d = 2.8$). Each subject was exposed to one of three instantiations of this process; one is illustrated in **Figure** 1B.

Subjects were instructed that they would be paid 'according to how many points you have won in total over the experiment,' and to expect average earnings of about 20 UK pounds. However, they were not advised of the actual exchange rate for points, nor of their cumulative point totals. At the completion of the task (due to behavioral protocol restrictions on differential treatment of subjects) each was paid 19 UK pounds.

Kalman filter model

The Kalman filter[1] is the Bayesian mean-tracking rule for the diffusion process described above. Assume the subject believes the process is governed by parameters $\hat{\sigma}_o$, $\hat{\sigma}_d$, $\hat{\lambda}$, and $\hat{\theta}$ (corresponding to σ_o , σ_d , λ, and θ above). Given, on trial *t*, a prior distribution over the true mean payoffs $\mu_{i,t}$ as independent Gaussians, $N(\hat{\mu}_{i,t}^{pre}, \hat{\sigma}_{i,t}^{2\,pre})$, then if option c_t is chosen and payoff r_t received, the posterior mean for that option is:

$$\hat{\mu}_{c_t,t}^{post} = \hat{\mu}_{c_t,t}^{pre} + \kappa_t \delta_t$$

with prediction error $\delta_t = r_t - \hat{\mu}_{c_t,t}^{pre}$ and learning rate ("gain") $\kappa_t = \hat{\sigma}_{c_t,t}^{2\,pre} / (\hat{\sigma}_{c_t,t}^{2\,pre} + \hat{\sigma}_o^2)$. The posterior variance for the chosen option is

$$\hat{\sigma}_{c_t,t}^{2\,post} = (1-\kappa_t)\hat{\sigma}_{c_t,t}^{2\,pre}$$

The posterior mean and variance for the unchosen options are unchanged by the observation. Taking into account the diffusion process, the prior distributions on the subsequent trial are given by $\hat{\mu}_{i,t+1}^{pre} = \hat{\lambda}\hat{\mu}_{i,t}^{post} + (1-\hat{\lambda})\hat{\theta}$ and $\hat{\sigma}_{i,t+1}^{2\,pre} = \hat{\lambda}^2\hat{\sigma}_{i,t}^{2\,post} + \hat{\sigma}_d^2$ for all *i*. The recursive process is initialized with prior distribution $N(\hat{\mu}_{i,0}^{pre}, \hat{\sigma}_{i,0}^{2\,pre})$.

Note that the heart of this procedure is an error-driven learning rule of the same form as TD or other delta-rule methods — the difference is the additional tracking of uncertainties $\hat{\sigma}_{i,t}^2$, which determine

the trial-specific learning rates κ_t. In general, uncertainties decrease for sampled options and increase for unsampled ones.

Together with this tracking rule, we examined three choice rules, each of which determined the probability $P_{i,t}$ of choosing option i on trial t as a function of the estimated payoffs. The ε-greedy rule is:

$$P_{i,t} = \begin{cases} 1-3\varepsilon & i = \arg\max(\hat{\mu}_{i,t}^{pre}) \\ \varepsilon & \text{otherwise} \end{cases}$$

with exploration parameter ε. (If there is a tie for the winning action, they are made equally probable.) The softmax rule is:

$$P_{i,t} = \frac{\exp(\beta\hat{\mu}_{i,t}^{pre})}{\sum_j \exp(\beta\hat{\mu}_{j,t}^{pre})}$$

with exploration parameter β. Finally, we tested a rule in which an exploration bonus[2] of φ standard deviations was added to the expected mean payoff, and choices were softmax in this adjusted value:

$$P_{i,t} = \frac{\exp(\beta[\hat{\mu}_{i,t}^{pre} + \varphi\hat{\sigma}_{i,t}^{pre}])}{\sum_j \exp(\beta[\hat{\mu}_{j,t}^{pre} + \varphi\hat{\sigma}_{j,t}^{pre}])}$$

Note that this model nests uncertainty bonuses within a softmax scheme: it reduces to the simple softmax model for $\varphi = 0$ (as was nearly the case in our behavioral fits) and to classic deterministic uncertainty-bonus exploration as β approaches infinity with φ positive. Between these regimes, the model spans hybrids combining contributions of both approaches differentially according to the parameters.

Behavioral analysis

We evaluated the three models using Bayesian model comparison techniques[3]. We took the parameters $\hat{\sigma}_d$, $\hat{\lambda}$, $\hat{\theta}$, $\hat{\mu}_{i,0}^{pre}$, $\hat{\sigma}_{i,0}^{pre}$, ε or β, and φ to be free (holding σ_o constant due to model degeneracy). For each model, we fit these to the subjects' choice data by maximizing the likelihood of the observed choices

$$\prod_s \prod_t P_{c_{s,t},t}$$

compounded over subjects s and trials t. Here, $c_{s,t}$ denotes the choice made by subject s on trial t, and the underlying value estimates $\hat{\mu}_{i,t}^{pre}$ and uncertainties $\hat{\sigma}_{i,t}^{pre}$ were computed using the actual sequence of choices and outcomes through trial t - 1. (Fewer than 1% of trials, in which a response was not entered, were omitted.)

A combination of nonlinear optimization algorithms (Matlab optimization toolbox) was used to optimize the parameter fits, together with a search of different starting locations. We report negative log likelihoods (smaller values indicate better fit), both pure and penalized for model complexity (Bayesian information criterion; BIC[4]). We also report a pseudo-r^2 statistic[5], defined as $(r - l)/r$ where l and r are, respectively, the log likelihoods of the data under the model and under purely random choices ($P_{c_{s,t},t} = .25$ for all t).

The ε-greedy choice rule resists optimization since its likelihood is undifferentiable. We therefore optimized parameters in two steps, first using a differentiable approximation in which the "max" operation was replaced with a very sharp softmax, $P_{i,t} = \varepsilon + (1 - 4\varepsilon) \cdot \exp(\beta_t \hat{\mu}_{i,t}^{pre}) / \sum_j \exp(\beta_t \hat{\mu}_{j,t}^{pre})$ (with the softmax sharpness β_t taken to be 100 divided by the L2 norm of the vector of mean-adjusted value estimates, $\hat{\mu}_{i,t}^{pre} - \sum_j \hat{\mu}_{j,t}^{pre}$, to keep the softmax sharp at the scale of the values).

Locally optimal parameters for the approximate rule were then tuned for the exact rule using a non-

gradient search. The approximation was found to be tight (typically within 10 log likelihood points), suggesting that this is an effective way to optimize the original function.

As is standard in similar behavioral analyses[5-7] with a limited number of trials per subject, for each model, we fit the behavior of all subjects using a single instance of most of the model parameters ($\hat{\sigma}_d$, $\hat{\lambda}$, $\hat{\theta}$, $\hat{\mu}_{i,0}^{pre}$, $\hat{\sigma}_{i,0}^{pre}$ and φ). However, to capture some effects of inter-subject variability, we fit the parameter controlling the "noisiness" of choices (β or ε) individually for each subject and model.

To investigate whether our conclusions might be influenced by sharing of parameters between subjects, we also conducted an alternative analysis fitting all parameters individually for each subject.

Imaging procedure

The functional imaging was conducted using a 1.5 Tesla Siemens Sonata MRI scanner to acquire gradient echo T2* weighted echo-planar images (EPI) images with BOLD (blood oxygenation level dependent) contrast. We employed a special sequence designed to optimize functional sensitivity in OFC and medial temporal lobes[8]. This consisted of tilted acquistion in an oblique orientation at 30* to the AC-PC line, as well as application of a preparation pulse with a duration of 1 msec. and amplitude of –2 mT/m in the slice selection direction. The sequence enabled 36 axial slices of 3 mm thickness and 3 mm in-plane resolution to be acquired with a repetition time (TR) of 3.24 seconds. Coverage was obtained from the base of the orbitofrontal cortex and medial temporal lobes to the superior border of the dorsal anterior cingulate cortex. Subjects were placed in a light head restraint within the scanner to limit head movement during acquisition. Functional imaging data were acquired in two separate 385-volume runs. A T1-weighted structural image was also acquired for each subject.

Imaging analysis

Image analysis was performed using SPM2 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, U.K.). To correct for subject motion, the images were realigned to the first volume, spatially normalized to a standard T2* template with a resampled voxel size of $3mm^3$, and spatial smoothing was applied using a Gaussian kernel with a full width at half maximum (FWHM) of 8mm. Intensity normalization and high pass temporal filtering (using a filter width of 128 secs) were also applied to the data.

For the statistical analysis, each trial was modeled as having 2 time points: the time of the decision (arbitrarily set to be midway between the time of presentation of the bandits and the time of the recorded key press indicating choice of a specific bandit - on average 210 msecs after trial onset), and the time of the presentation of the outcome (3 seconds after recorded key press). We constructed regressors containing trial-by-trial outputs from the softmax model: classification of choices as greedy or non, prediction errors δ_t and choice probabilities $P_{c_{s,t},t}$. For the prediction error regressor, we simulated a TD signal using an impulse for the prediction error δ at the time of outcome, and an additional impulse at the time of decision (of size $\hat{\mu}^{pre}_{c_{s,t},t} - \hat{\mu}^{pre}_{avg,t}$ for an average-obtained value $\hat{\mu}^{pre}_{avg,t}$ tracked the same as the other means but regardless of subject choice). An alternative analysis, in which the prediction error impulses at decision and outcome were modeled using separate regressors and then studied in conjunction, produced nearly identical results. The other regressors (greedy vs non greedy and choice probability) were modeled at the time of the decision alone. We also entered the number of points won on each trial as an additional parametric modulator set at the time of outcome. 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 regressors in the SPM analysis to account for residual effects of scan to scan motion. To enable inference at the group level, the regression fits of each computational signal from each individual subject were taken to allow second level, random effects group statistics to be computed.

Results are reported in areas of interest at $p<0.001$ uncorrected. To show the full spatial extent of activations we also show effects significant at $p<0.01$ uncorrected.

The structural T1 images were co-registered to the mean functional EPI images for each subject and normalized using the parameters derived from the EPI images. Anatomical localization was carried out by overlaying the t-maps on a normalized structural image averaged across subjects, and with reference to an anatomical atlas[9].

For the analysis and visualization of timecourse data from regions identified in the SPM analysis, raw signal timecourses were extracted from each region using the peak voxel from each individual subject from within a 10mm sphere centered on the group peak co-ordinate, after adjusting the data for the effects of motion (and mean correcting the signal). For alignment, these timecourses were upsampled to 10 Hz using a Fourier transform, averaged over trials and plotted. The upsampled OFC and medial PFC timecourses were modeled using a hemodynamic impulse at each outcome or decision time (respectively); least-squares response coefficients were grouped in evenly spaced bins and averaged over trials to produce the bar plots in Figure 2.

For each region showing differential activity between exploratory and exploitative trials, a multiple regression analysis was conducted to investigate whether the differential BOLD responses could be explained by any potentially confounding factors. The dependent variable was a per-trial estimate of the BOLD response (extracted by modeling the peak timecourses using impulses for each decision convolved with the canonical hemodynamic response, sampled at image acquisition times, and minimizing squared error); independent variables were the explore/exploit labeling and 10 other factors. These were the value, choice probability, and uncertainty (prior variance) accorded by the model to the chosen option ("val chosen", "prob chosen", "unc chosen" in **Supplementary Table** 4); the modeled value and probability of the highest-valued option ("val max" and "prob max"); the reaction time; the obtained reward; a binary variable signaling whether the choice was the same as the previous one ("switch"); the length in trials of any preceding uninterrupted run on the chosen option ("runlength chosen"); and the fraction of time the chosen option had also been chosen in the recent past (using an exponentially windowed running average with decay constant 0.9 per trial; "propensity chosen").

Supplementary Discussion

Behavioral analysis: Subject heterogeneity

Our conclusions are based on analyses in which all subjects' behavior was modeled as being produced by a single, shared, instance of most of the free parameters, with any heterogeneity captured through subject-specific fits of the parameters controlling choice noisiness (β or ε). We also investigated fully individualized fits with separate parameters for each subject. There were a number of indications that these fits were less reliable than the ones on which we focus: many parameters attained extreme values; the examination of estimated Hessians of the likelihood at the

optima suggested parameters were more poorly identified; and some of the modeled signals correlated less strongly with fMRI measurements, suggesting the many additional parameters had been overfit to behavior. Nonetheless, the results support the same general conclusions. Notably, there was little evidence that uncertainty bonuses could account for the exploration that the subjects exhibited.

To probe the effects of the uncertainty bonus over individuals and the population, we investigated these individual fits in a number of ways. First, an asymptotic approximation of the variance of a parameter estimate can be obtained from the inverse Hessian of the likelihood function at the optimum; according to this measure, the bonus coefficient φ was insignificantly different from zero (i.e., by less than two standard deviations) in thirteen of the fourteen subjects. Alternatively, the likelihood of choice data for models with and without the bonus, penalized for model complexity, may be compared for each subject individually; here, the bonus was modestly but significantly helpful for about half the subjects (7/14 according to BIC, and 8/14 according to the Aikake information criterion and the likelihood ratio test at $P<.05$). But, in fact, the best-fitting bonus coefficient was as often negative – i.e., *discouraging* exploration – as positive. (A negative coefficient was found in 8/14 subjects including 4 of the 8 for whom the bonus significantly improved the data likelihood.) This suggests that this model feature was generically capturing autocorrelation among the choices, but not specifically an exploratory tendency. Finally, since in the model, the uncertainty bonus is nested within a softmax choice rule, we compared the contribution of each strategy to producing exploration. We found that the majority of decisions classed as exploratory when the model was fit without the bonus (i.e., actions chosen despite not having the highest predicted value) were not explained by the inclusion of bonuses (i.e., the sum of the predicted value plus the bonus was still smaller for the chosen option than for some alternative, so softmax was still required to produce the decision). This was true for 89.9% of exploratory trials over all subjects (individuals ranged between 78.2% and 100%). Thus, the predominant mode of exploration even with bonuses included appeared to be softmax. In short, although including this

model feature improved fit for some subjects, it does not appear to have captured the exploratory strategy that they were adopting.

Behavioral analysis: Fit parameters

Supplementary Table 2 lists the best fitting parameters for each of the three behavioral models. These appear plausibly identified and broadly similar between models (except for the large initial uncertainty, $\hat{\sigma}_{i,0}^{2\,pre}$, in the ε-greedy model, a feature that impacts only the first few trials). Parameters are similar to those actually used to generate the payoffs, except that subjects' behavior is best explained by assuming that they overestimate the speed of diffusion in the payoffs, $\hat{\sigma}_d$, an effect particularly apparent in the softmax fits. Since large values of this parameter induce high learning rates, this is an indication that subjects are more sensitive to the most recent experience with a bandit than they optimally should be.

Imaging analysis: Multiple regression

Compared with exploitation, exploratory choices tend to favor less valuable, lower probability, and more uncertain targets. We therefore subjected all of the regions showing differential activity during exploration and exploitation to a further, post-hoc multiple linear regression analysis (**Supplementary Table** 4), to investigate whether such potential confounds could account for the differences in activity. Additional explanatory factors in the regression included reaction time, actual reward received, stay versus switch (intended to control for processes such as attentional disengagement[10], thought to involve parietal cortex), and two measures of the degree of recent preference for the chosen option (intended to control for the strength of habitual responding). None of these variables could explain the differential responding during exploratory trials in right frontopolar or bilateral IPS areas (which each still correlated with exploration at $P<.001$ uncorrected). However, the original SPM analysis identified a number of additional areas as differentially active during exploration (**Supplementary Table** 5). As can be seen in

Supplementary Table 4, with confounds taken into account, activity each of these areas was less strongly and significantly correlated with exploration than was activity in the frontopolar and IPS regions. None of these regions was significantly correlated with exploration at P<.001, and in some cases activity was better explained by several confounding factors. These correlations (notably right supplementary motor area with a measure of uncertainty) merit future investigation, since the present study concentrated its statistical power on the balance between exploration and exploitation.

Another noteworthy trend from this analysis (though not reaching significance at the high threshold discussed here) was that frontopolar decision activity was additionally correlated (positively, P=.002) with the probability of the apparently optimal action – that is, the probability of exploitation. The highest net responses would therefore be seen when exploration is chosen most against the odds. This observation (and also the finding, discussed in the main article, of inverse correlation between activation in a dorsolateral PFC region and modeled choice probability) is in keeping with the idea that additional cognitive control is needed to enforce exploration when exploitation seems most favorable.

Supplementary References

1. Anderson, B.D.O. & Moore, J.B. *Optimal Filtering* (Prentice-Hall, Englewood Cliffs, NJ, 1979).

2. Kaelbling, L.P. *Learning in Embedded Systems* (MIT Press, Cambridge, Mass., 1993).

3. Kass, R.E. & Raftery, A.E. Bayes factors. *Journal of the American Statistical Association* **90**, 7730–795 (1995).

4. Schwarz, G. Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464 (1978).

5. Camerer, C. & Ho T.-H., Experience-weighted attraction learning in normal form games. *Econometrica* **67**, 827-874 (1999).

6. O'Doherty, J. *et al.* Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454 (2004).

7. Ho, T-H., Camerer, C., & Chong, J-K. The economics of learning models: A self-tuning theory of learning in games. Working paper, University of California, Berkeley (2004).

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

9. Duvernoy, H.M. *The Human Brain* (Vienna, Springer-Verlag, 1999).

10. Posner MI, Walker JA, Friedrich FH & Rafal RD. Effects of parietal injury on covert orienting of attention. *J. Neurosci.* **4**:1863-1874 (1984).

	ε-greedy	softmax	uncertainty
-LL	4190.6	3972.1	3972.1
pseudo-r^2	0.27353	0.31141	0.31141
# parameters	19	19	20
BIC	4269.8	4051.3	4055.4

Supplementary Table 1: Quality of behavioral fits to 4,161 choices from 14 subjects, for three models. -LL: Negative log likelihood. BIC: Bayesian information criterion.

	ε-greedy	softmax	uncertainty		generative
ε or β	0.121 ± 0.0499	0.112 ± 0.0547	0.112 ± 0.0547		
φ	-	-	7.61e-6		
$\hat{\lambda}$	0.974	0.924	0.924	λ	0.9836
$\hat{\theta}$	49.2	50.5	50.4	θ	50.0
$\hat{\sigma}_d$	9.53	51.3	50.9	σ_d	2.80
$\hat{\sigma}_o$ (fixed)	(4.00)	(4.00)	(4.00)	σ_o	4.00
$\hat{\mu}_{i,0}^{pre}$	87.1	85.7	85.7		
$\hat{\sigma}_{i,0}^{2pre}$	3.36e+5	4.61	4.61		

Supplementary Table 2: Parameter fits to 4,161 choices from 14 subjects, for three models (ε-greedy, softmax, and uncertainty bonus). Parameters ε and β shown as mean ± 1 SD, over individual fits to each subject; other parameters were yoked between subjects. For comparison, the parameters used to generate the payoffs are also shown.

Prediction error		**MNI co-ordinates**			
	Side	**X**	**Y**	**Z**	**Z-score**
Ventral striatum (nucleus accumbens)	R	9	12	-9	3.35
Dorsal striatum (caudate nucleus)	R	9	0	18	3.19

Supplementary Table 3: Co-ordinates of ventral and dorsal striatum activity showing significant correlation with the prediction error signal from the computational model.

	left fpole	left ips	right ips	left pm	right sma	cereb1	cereb2
explore	0.49 (8.6E-5)	0.37 (1.4E-4)	0.39 (2.1E-4)	0.33 (0.003)	0.31 (0.015)	0.30 (0.005)	0.29 (0.013)
val chosen x 0.01	1.49 (0.088)	0.81 (0.231)	1.19 (0.104)	1.30 (0.088)	3.02 (0.001)	1.43 (0.052)	2.80 (0.001)
prob chosen	-1.07 (0.007)	-0.47 (0.120)	-0.60 (0.071)	-0.78 (0.023)	-1.22 (0.002)	-0.49 (0.135)	-1.23 (0.001)
unc chosen	-0.13 (0.231)	0.09 (0.259)	0.10 (0.247)	0.15 (0.103)	0.45 (4.5E-5)	0.08 (0.365)	0.24 (0.015)
val max x 0.01	-1.79 (0.020)	-1.43 (0.016)	-1.81 (0.005)	-1.83 (0.007)	-1.80 (0.022)	-1.04 (0.110)	-2.44 (0.001)
prob max	1.08 (0.002)	0.51 (0.059)	0.58 (0.048)	0.66 (0.030)	0.49 (0.173)	0.23 (0.431)	1.12 (0.001)
reward x 0.1	0.05 (0.890)	0.07 (0.803)	0.37 (0.238)	0.42 (0.196)	0.10 (0.793)	0.10 (0.739)	0.28 (0.417)
runlength chosen x 0.1	0.08 (0.178)	0.02 (0.668)	0.06 (0.250)	0.04 (0.420)	-0.01 (0.912)	0.09 (0.072)	-0.02 (0.752)
propensity chosen	-0.16 (0.508)	0.12 (0.496)	0.22 (0.270)	0.26 (0.197)	0.87 (3.1E-4)	-0.03 (0.884)	0.10 (0.650)
switch	0.09 (0.433)	-0.03 (0.759)	0.09 (0.322)	0.07 (0.471)	0.16 (0.138)	0.22 (0.016)	0.01 (0.923)
rt	-0.09 (0.604)	0.25 (0.064)	-0.05 (0.748)	0.30 (0.052)	0.58 (0.001)	0.13 (0.371)	0.04 (0.791)

Supplementary Table 4: Coefficients from multiple linear regression for 11 explanatory variables with significance (against the null hypothesis that the coefficient equals zero) in parentheses. The dependent variable is the per-trial BOLD signal change estimate at the time of decision. Coefficients significant at $P < .001$ are highlighted.

Explore > Exploit		**MNI co-ordinates**			
	Side	**X**	**Y**	**Z**	**Z-score**
Lateral premotor cortex	L	-57	3	36	4.92
Supplementary Motor Area	R	3	9	51	4.36
Cerebellum	R	21	-54	-30	5.42
	R	18	-57	-51	4.28

Supplementary Table 5: Additional regions showing significantly greater activity on exploratory compared to exploitative trials. We report only those areas surviving whole brain correction with false discovery rate (FDR) at $p<0.05$. None of these activations survived the additional multiple regression test against confounds described in **Supplementary Methods**.