LETTERS

Cortical substrates for exploratory decisions in humans

Nathaniel D. Daw[1]*, John P. O'Doherty[2]*†, Peter Dayan[1], Ben Seymour[2] & Raymond J. Dolan[2]

Decision making in an uncertain environment poses a conflict between the opposing demands of gathering and exploiting information. In a classic illustration of this 'exploration–exploitation' dilemma[1], a gambler choosing between multiple slot machines balances the desire to select what seems, on the basis of accumulated experience, the richest option, against the desire to choose a less familiar option that might turn out more advantageous (and thereby provide information for improving future decisions). Far from representing idle curiosity, such exploration is often critical for organisms to discover how best to harvest resources such as food and water. In appetitive choice, substantial experimental evidence, underpinned by computational reinforcement learning[2] (RL) theory, indicates that a dopaminergic[3,4], striatal[5–9] and medial prefrontal network mediates learning to exploit. In contrast, although exploration has been well studied from both theoretical[1] and ethological[10] perspectives, its neural substrates are much less clear. Here we show, in a gambling task, that human subjects' choices can be characterized by a computationally well-regarded strategy for addressing the explore/exploit dilemma. Furthermore, using this characterization to classify decisions as exploratory or exploitative, we employ functional magnetic resonance imaging to show that the frontopolar cortex and intraparietal sulcus are preferentially active during exploratory decisions. In contrast, regions of striatum and ventromedial prefrontal cortex exhibit activity characteristic of an involvement in value-based exploitative decision making. The results suggest a model of action selection under uncertainty that involves switching between exploratory and exploitative behavioural modes, and provide a computationally precise characterization of the contribution of key decision-related brain systems to each of these functions.

Exploration is a computationally refined capacity, demanding careful regulation. Two possibilities for this regulation arise. On the one hand, we might expect the involvement of cognitive, prefrontal control systems[11] that can supervene[12] over simpler dopaminergic/striatal habitual mechanisms. On the other hand, theoretical work on optimal exploration[1,13] indicates a more unified architecture, according to which actions can be assessed with the use of a metric that integrates both primary reward and the informational value of exploration, even in simple, habitual decision systems.

We studied patterns of behaviour and brain activity in 14 healthy subjects while they performed a 'four-armed bandit' task involving repeated choices between four slot machines (Fig. 1; see Supplementary Methods). The slots paid off points (to be exchanged for money) noisily around four different means. Unlike standard slots, the mean payoffs changed randomly and independently from trial to trial, with subjects finding information about the current worth of a slot only through sampling it actively. This feature of the experimental design, together with a model-based analysis, allowed us to study exploratory and exploitative decisions under uniform conditions, in the context of a single task.

We asked subjects in post-task interviews to describe their choice strategies. The majority (11 of 14) reported occasionally trying the different slots to work out which currently had the highest payoffs (exploring) while at other times choosing the slot they thought had the highest payoffs (exploiting). To investigate this behaviour quantitatively, we considered RL (ref. 2) strategies for exploration. These strategies come in three flavours, differing in how exploratory actions are directed. The simplest method, known as 'ε-greedy', is undirected: it chooses the 'greedy' option (the one believed to be best) most of the time, but occasionally (with probability ε) substitutes a random action. A more sophisticated approach is to guide exploration by expected value, as in the 'softmax' rule. With softmax, the decision to explore and the choice of which suboptimal action to take are determined probabilistically on the basis of the actions' relative expected values. Last, exploration can additionally be directed by awarding bonuses in this latter decision towards actions whose consequences are uncertain: specifically, to those for which exploration will be most informative. The optimal strategy for a restricted class of simple bandit tasks has this characteristic[1], as do standard heuristics[14] for exploration in more complicated RL tasks such as ours, for which the optimal solution is computationally intractable.

We compared the fit of three distinct RL models, embodying the aforementioned strategies, to our subjects' behavioural choices. All the models learned the values of actions with the use of a Kalman filter (see Supplementary Methods), an error-driven prediction algorithm that generalizes the temporal-difference learning algorithm (used in most RL theories of dopamine) by also tracking uncertainty about the value of each action. The models differed only in their choice rules. We compared models by using the likelihood of the subjects' choices given their experience, optimized over free parameters. This comparison (Supplementary Tables 1 and 2) revealed strong evidence for value-sensitive (softmax) over undirected (ε-greedy) exploration. There was no evidence to justify the introduction of an extra parameter that allowed exploration to be directed towards uncertainty (softmax with an uncertainty bonus): at optimal fit, the bonus was negligible, making the model equivalent to the simpler softmax. We conducted additional model fits (see Supplementary Information) to verify that these findings were not an artefact of our assumptions about the yoking of free parameters between subjects.

Having characterized subjects' behaviour computationally, we used the best-fitting softmax model to generate regressors containing value predictions, prediction errors and choice probabilities for each subject on each trial. We used statistical parametric mapping to

[1]Gatsby Computational Neuroscience Unit, University College London (UCL), Alexandra House, 17 Queen Square, London WC1N 3AR, UK. [2]Wellcome Department of Imaging Neuroscience, UCL, 12 Queen Square, London WC1N 3BG, UK. †Present address: Division of Humanities and Social Sciences, California Institute of Technology, 1200 East California Boulevard, Pasadena, California 91125, USA.
*These authors contributed equally to this work.

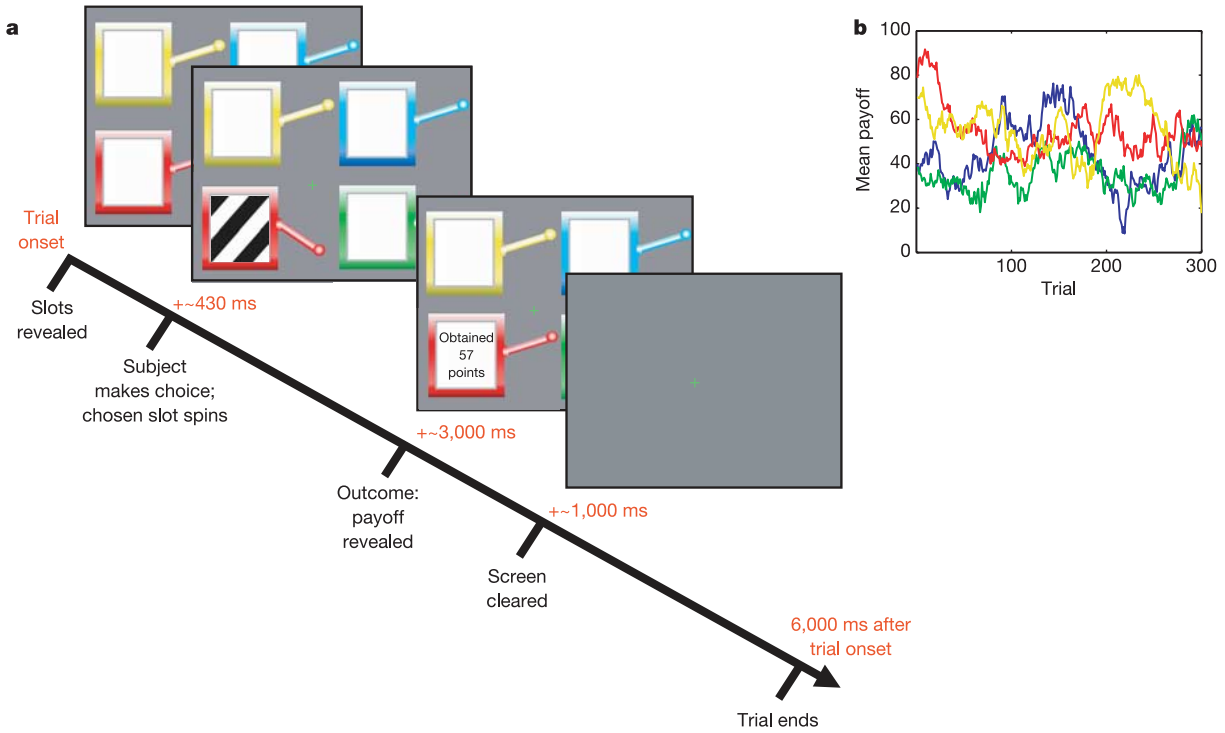

Figure 1 | **Task design.** **a**, Illustration of the timeline within a trial. Initially, four slots are presented. The subject chooses one, which then spins. Three seconds later the number of points won is revealed. After a further second the screen is cleared. The next trial is triggered after a fixed trial length of 6 s and an additional variable inter-trial interval (mean 2 s). **b**, Example of mean payoffs that would be received for choosing each slot machine (four coloured lines) on each trial, demonstrating their independent random diffusion. The payoff received for a particular choice is corrupted by gaussian noise around this mean.

identify brain regions in which neural activity was significantly correlated with the model's internal signals. Consistent with previous studies[7–9] was our observation that a prediction error was correlated significantly with activity in both the ventral and dorsal striatum (see Supplementary Table 3). Other, cortical, structures linked to this subcortical network[15] also showed significant value-related correlations. Specifically, we found activity in medial orbitofrontal cortex to be correlated with the magnitude of the obtained payoff (Fig. 2a), a

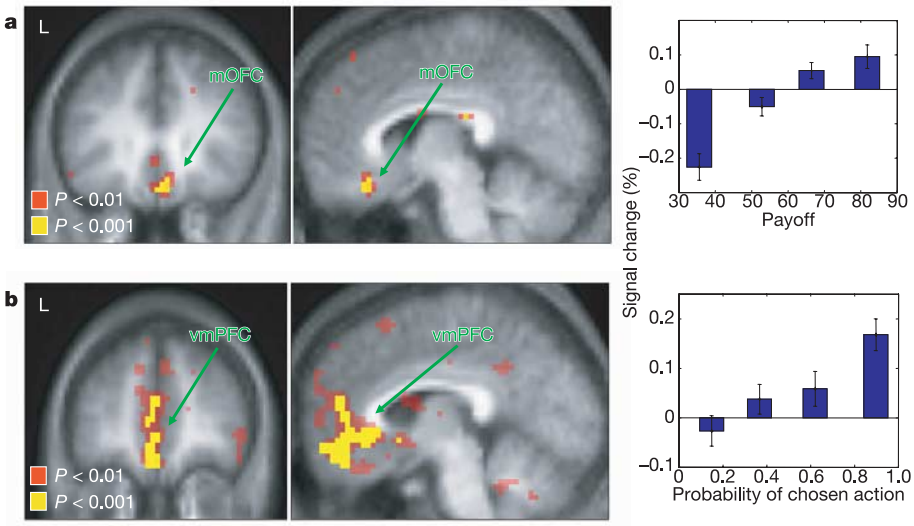

Figure 2 | **Reward-related activations.** Activation maps (yellow, $P < 0.001$; red, $P < 0.01$ to illustrate the full extent of the activations) are superimposed on a subject-averaged structural scan. **a**, Region of medial orbitofrontal cortex (mOFC) correlating significantly with the number of points received. The coordinates of the activated area are [3,30,−21, peak $z = 3.87$]. The bar plot shows the average BOLD response to outcome, binned by amount won (error bars represent s.e.m.). **b**, Regions of ventromedial prefrontal cortex (vmPFC; including medial and lateral orbitofrontal cortex and adjacent medial prefrontal cortex) correlating significantly with the probability assigned by the computational model to the subject's choice of slot. The coordinates of the activated areas are as follows: medial orbitofrontal, [−3,45,−18, peak $z = 5.62$]; lateral orbitofrontal (not illustrated), [45,36,−15, peak $z = 4.6$]; medial prefrontal, [−3,33,−6, peak $z = 4.62$]. The bar plot shows the average medial prefrontal BOLD response to decision, binned by choice probability (error bars represent s.e.m.).

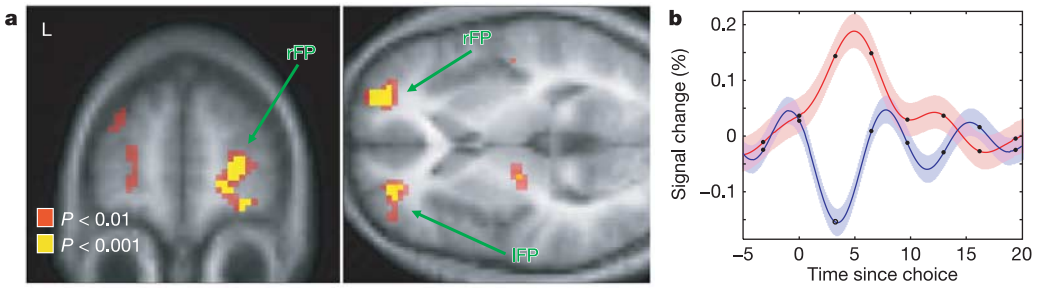

Figure 3 | Exploration-related activity in frontopolar cortex. a, Regions of left and right frontopolar cortex (lFP, rFP) showing significantly increased activation on exploratory compared with exploitative trials. Activation maps (yellow, $P < 0.001$; red, $P < 0.01$) are superimposed on a subject-averaged structural scan. The coordinates of activated areas are [−27,48,4, peak $z = 3.49$] for lFP and [27,57,6, peak $z = 4.13$] for rFP. **b**, rFP BOLD time courses averaged over 1,515 exploratory (red line) and 2,646 exploitative (blue line) decisions. Black dots indicate the sampling frequency (although, because sample alignment varied from trial to trial, time courses were upsampled). Coloured fringes show error bars (representing s.e.m.).

finding consistent with previous evidence indicating that this region is involved in coding the relative value of different reward stimuli, including abstract rewards[16,17]. Furthermore, activity in medial and lateral orbitofrontal cortex, extending into ventro-medial prefrontal cortex, was correlated with the probability assigned by the model to the action actually chosen on a given trial (Fig. 2b). In the softmax model, this probability is a relative measure of the expected reward value of the chosen action, and the observed profile of activity is thus consistent with a role for orbital and adjacent medial prefrontal cortex in encoding predictions of future reward[18,19]. The same quantity was negatively correlated with activity in a small area of dorsolateral prefrontal cortex (left: −39,36,42, peak $z = 3.38$; right: 36,33,33, peak $z = 3.27$); that is, higher activity was seen there for lower-probability choices.

We next sought to identify brain activity that selectively reflected whether actions were chosen for their exploratory or exploitative potential. To test for such a signature, we classified trials according to whether the actual choice was the one predicted by the model to be the dominant slot machine with the highest expected value (exploitative) or a dominated machine with a lower expected value (exploratory). We then directly compared the pattern of brain activity associated with these exploratory and exploitative trials. We found no area that exhibited significantly higher activity for exploitative than exploratory decisions (employing whole-brain correction for multiple comparisons). However, the opposite contrast revealed several activations. First, right anterior frontopolar cortex (Fig. 3a) was significantly more active during decisions classified as exploratory ($P < 0.05$, corrected whole-brain for multiple comparisons with false discovery rate; activation was noted bilaterally at $P < 0.001$ uncorrected but did not survive whole-brain correction on the left). Average blood-oxygenation-level-dependent (BOLD) signal time courses from the region (Fig. 3b) demonstrated phasic increases and decreases in activity that were time-locked to subjects' exploratory and exploitative decisions, respectively.

Because the prefrontal cortex is the principal cortical region implicated in behavioural control[20], the signal we observed in anterior frontopolar cortex could reflect a control mechanism facilitating the switching of behavioural strategies between exploratory and exploitative modes. This most rostral of prefrontal regions is known to be associated with high-level control[21]. This region sits atop a proposed hierarchy of nested prefrontal controllers[22] and is implicated in mediating between different goals, subgoals[23] or cognitive processes[21].

Differential activation during exploratory trials was also observed bilaterally in anterior intraparietal sulcus (whole-brain corrected at $P < 0.05$; Fig. 4), bordering on the postcentral gyrus. The sulcus has repeatedly been implicated in decision making in both humans[15,19] and primates[24–26], with different subregions being associated with different output modalities. In lateral intraparietal area LIP, associated with saccades, neurons also carry information about decision variables such as the reward expected for a saccade[24–26]; the area perhaps serves as an interface between frontal areas (where such information may be calculated) and motor output. The anterior border of the sulcus, close to our exploration-related activation, is associated with grasping and manual manipulation[27], raising the possibility that such information (here, that associated with exploration) might also reach parietal regions involved in the button-press actions in our task.

Last, we used a multiple regression analysis to verify that differential activity in frontopolar and intraparietal regions during exploratory trials was not better explained by any of several potentially confounding factors such as switching between options or reaction times (see Supplementary Information and Supplementary Tables 4 and 5).

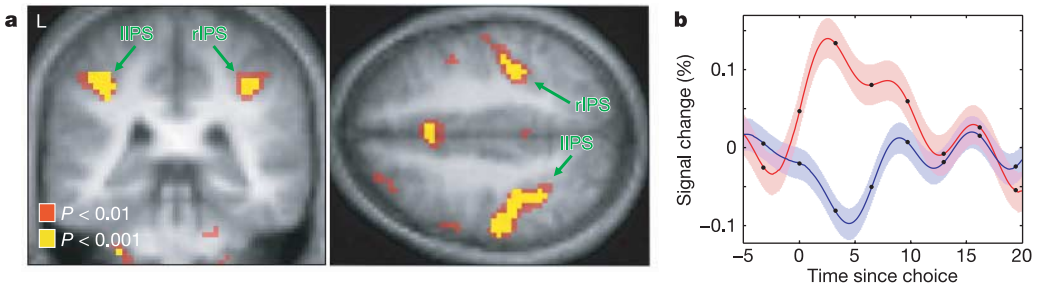

Figure 4 | Exploration-related activity in intraparietal sulcus. a, Regions of left and right intraparietal sulcus (lIPS and rIPS) showing significantly increased activation on exploratory compared with exploitative trials. Activation maps (yellow, $P < 0.001$; red, $P < 0.01$) are superimposed on a subject-averaged structural scan. The coordinates of the activated areas are [−29,−33,45, peak $z = 4.39$] for lIPS and [39,−36,42, peak $z = 4.16$] for rIPS. **b**, lIPS BOLD time courses averaged over 1,515 exploratory (red line) and 2,646 exploitative (blue line) decisions. Black dots indicate the sampling frequency (although, because sample alignment varied from trial to trial, time courses were upsampled). Coloured fringes show error bars (representing s.e.m.).

These results have important implications for both computational and neural accounts of action selection. The finding of brain regions discretely implicated in exploration (and particularly that one of them is a prefrontal, high-level control structure[21]) is consistent with a theory in which exploration is accomplished by overriding an exploitative tendency, but troubling for accounts such as uncertainty bonus schemes[1,14], which more tightly entangle exploration and exploitation. Such anatomical separation would be unlikely under these latter schemes, because they work by choosing actions with respect to a unified value metric that simultaneously prizes both information gathering and primary reward. Just such an exploration-encouraging value metric has previously been suggested to explain why dopamine neurons respond to novel, neutral stimuli[13]; such anomalous responses in an otherwise typically appetitive signal remain puzzling in view of our failure here to find either behavioural or neural evidence for such an account.

Exploration has a central role in the acquisition of adaptive behaviour in environments that change. Characteristic expressions of frontal pathology[28] include impairments in task switching as well as behavioural perseveration, which might relate, at least in part, to a core deficit in exploration. As one might expect for such a critical function, subcortical systems are also implicated in the control of exploration, with noradrenaline being suggested as regulating a global propensity to explore[29,30], a factor captured in our model in terms of the parameter regulating competition in the softmax rule. Last, self-directed exploration of the form studied here is an example of a refined cognitive function that is ubiquitous but hard to pin down in regular designs (because exploratory and exploitative responses are apparently seamlessly mixed). We were able to capture it only through a tight coupling of computational modelling, behavioural analysis and functional neuroimaging.

METHODS

Fourteen right-handed healthy human subjects participated in an fMRI scan (using a 1.5 T Siemens Sonata scanner) while repeatedly choosing between animated slot machines. One of three candidate reinforcement learning models for their behaviour was selected, and its parameters estimated, by maximizing the cumulative likelihood of the subjects' choices given the model and parameters. Trials were classified according to the model as exploratory or exploitative, and trial-by-trial estimates of subjects' predictions about slot machine payoffs (and the error or mismatch between those predictions and received payoffs) were generated by running the model progressively on the subjects' actual choices and winnings. A general linear model implemented in SPM2 (Wellcome Department of Imaging Neuroscience, Institute of Neurology, UCL) was used to locate brain voxels where the measured BOLD signal was significantly correlated with these model-generated signals. Regions identified as significantly correlated with exploration were subjected to a subsequent multiple regression analysis to investigate whether other, confounding factors might better account for the observed activity. For a detailed description of the experimental and analytical techniques, see Supplementary Methods.

Received 7 February; accepted 30 March 2006.

1. Gittins, J. C. & Jones, D. in *Progress in Statistics* (ed. Gani, J.) 241–266 (North-Holland, Amsterdam, 1974).
2. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, Massachusetts, 1998).
3. Montague, P. R., Dayan, P. & Sejnowski, T. J. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16,** 1936–1947 (1996).
4. Bayer, H. M. & Glimcher, P. W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* **47,** 129–141 (2005).
5. Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C. & Fiez, J. A. Tracking the hemodynamic responses to reward and punishment in the striatum. *J. Neurophysiol.* **84,** 3072–3077 (2000).
6. Knutson, B., Westdorp, A., Kaiser, E. & Hommer, D. fMRI visualization of brain activity during a monetary incentive delay task. *Neuroimage* **12,** 20–27 (2000).
7. McClure, S. M., Berns, G. S. & Montague, P. R. Temporal prediction errors in a passive learning task activate human striatum. *Neuron* **38,** 339–346 (2003).
8. O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H. & Dolan, R. J. Temporal difference models and reward-related learning in the human brain. *Neuron* **38,** 329–337 (2003).
9. O'Doherty, J. P. *et al.* Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304,** 452–454 (2004).
10. Charnov, E. L. Optimal foraging: The marginal value theorem. *Theor. Popul. Biol.* **9,** 129–136 (1976).
11. Owen, A. M. Cognitive planning in humans: Neuropsychological, neuroanatomical and neuropharmacological perspectives. *Prog. Neurobiol.* **53,** 431–450 (1997).
12. Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioural control. *Nature Neurosci.* **8,** 1704–1711 (2005).
13. Kakade, S. & Dayan, P. Dopamine: Generalization and bonuses. *Neural Netw.* **15,** 549–559 (2002).
14. Kaelbling, L. P. *Learning in Embedded Systems* (MIT Press, Cambridge, Massachusetts, 1993).
15. McClure, S. M., Laibson, D. I., Loewenstein, G. & Cohen, J. D. Separate neural systems value immediate and delayed monetary rewards. *Science* **306,** 503–507 (2004).
16. O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J. & Andrews, C. Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neurosci.* **4,** 95–102 (2001).
17. O'Doherty, J. Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Curr. Opin. Neurobiol.* **14,** 769–776 (2004).
18. Gottfried, J. A., O'Doherty, J. & Dolan, R. J. Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* **301,** 1104–1107 (2003).
19. Tanaka, S. C. *et al.* Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neurosci.* **7,** 887–893 (2004).
20. Miller, E. K. & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24,** 167–202 (2001).
21. Ramnani, N. & Owen, A. M. Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. *Nature Rev. Neurosci.* **5,** 184–194 (2004).
22. Koechlin, E., Ody, C. & Kouneiher, F. A. The architecture of cognitive control in the human prefrontal cortex. *Science* **302,** 1181–1185 (2003).
23. Braver, T. S. & Bongiolatti, S. R. The role of frontopolar cortex in subgoal processing during working memory. *Neuroimage* **15,** 523–536 (2002).
24. Platt, M. L. & Glimcher, P. W. Neural correlates of decision variables in parietal cortex. *Nature* **400,** 233–238 (1999).
25. Sugrue, L. P., Corrado, G. S. & Newsome, W. T. Matching behaviour and the representation of value in the parietal cortex. *Science* **304,** 1782–1787 (2004).
26. Dorris, M. C. & Glimcher, P. W. Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron* **44,** 365–378 (2004).
27. Grefkes, C. & Fink, G. R. The functional organization of the intraparietal sulcus in humans and monkeys. *J. Anat.* **207,** 3–17 (2005).
28. Burgess, P. W., Veitch, E., de Lacy Costello, A. & Shallice, T. The cognitive and neuroanatomical correlates of multitasking. *Neuropsychologia* **38,** 848–863 (2000).
29. Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J. & Aston-Jones, G. The role of locus coeruleus in the regulation of cognitive performance. *Science* **283,** 549–554 (1999).
30. Doya, K. Metalearning and neuromodulation. *Neural Netw.* **15,** 495–506 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Li, S. McClure, B. King-Casas and P. R. Montague for sharing their unpublished data on exploration, and Y. Niv, Z. Gharamani and C. Camerer for discussions. Funding was from a Royal Society USA Research Fellowship (N.D.), the Gatsby Foundation (N.D., P.D.), the EU BIBA project (N.D., P.D.), and a Wellcome Trust Programme Grant (J.O.D., R.D.).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to N.D. (daw@gatsby.ucl.ac.uk) or J.O.D. (jdoherty@hss.caltech.edu).

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**.