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.

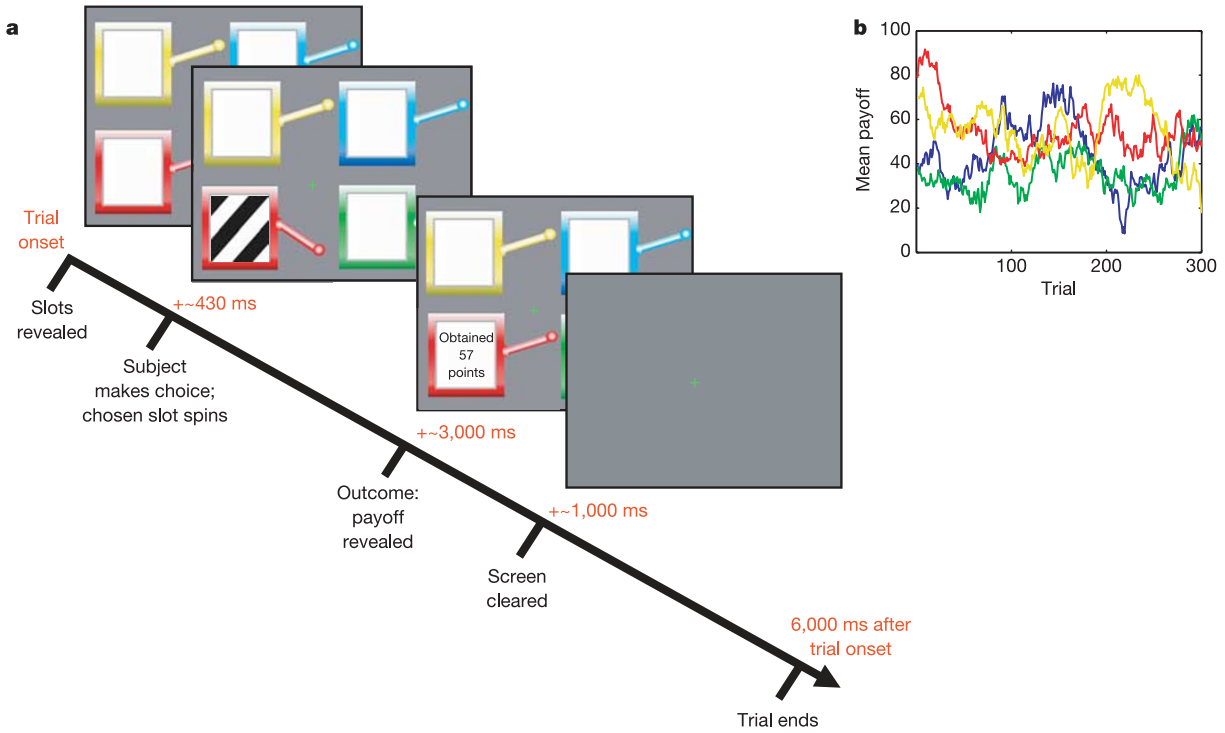

Figure 1 | **Task design.** **a**, Illustration of the timeline within a trial. Initially, four slots are presented. The subject chooses one, which then spins. Three seconds later the number of points won is revealed. After a further second the screen is cleared. The next trial is triggered after a fixed 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

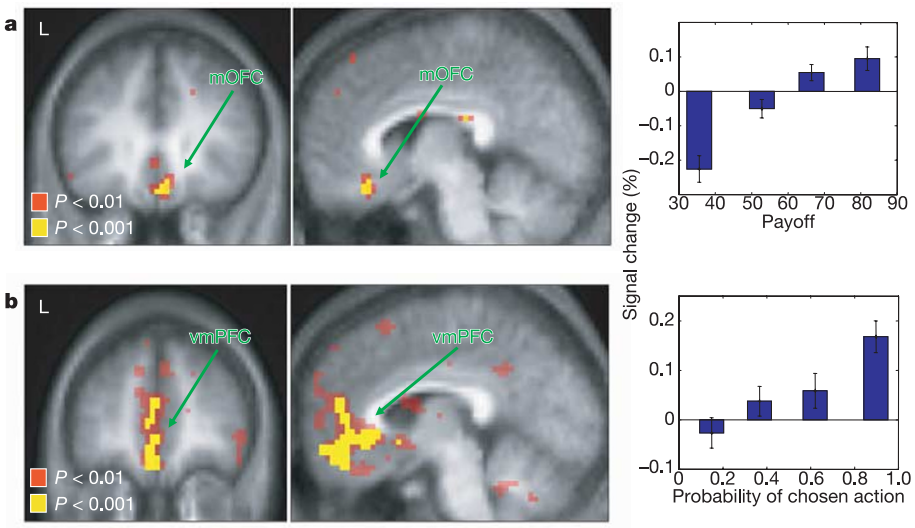

Figure 2 | **Reward-related activations.** Activation maps (yellow, $P < 0.001$; red, $P < 0.01$ to illustrate the full extent of the activations) are superimposed on a subject-averaged structural scan. **a**, Region of medial orbitofrontal cortex (mOFC) correlating significantly with the number of points 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.).

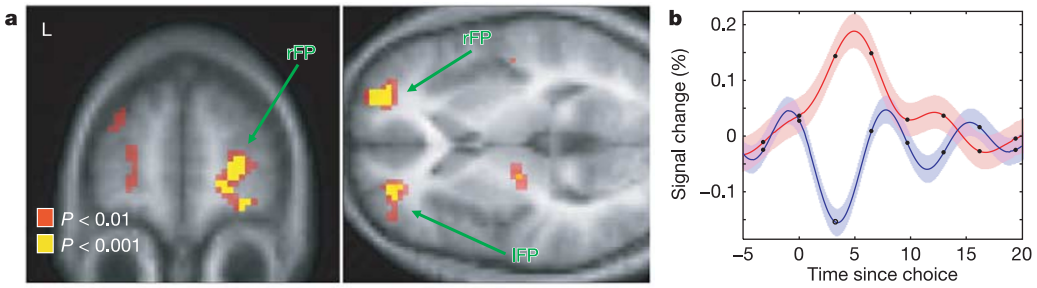

Figure 3 | Exploration-related activity in frontopolar cortex. a, Regions of left and right frontopolar cortex (lFP, rFP) showing significantly increased activation on exploratory compared with exploitative trials. Activation maps (yellow, $P < 0.001$; red, $P < 0.01$) are superimposed on a subject-averaged 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).

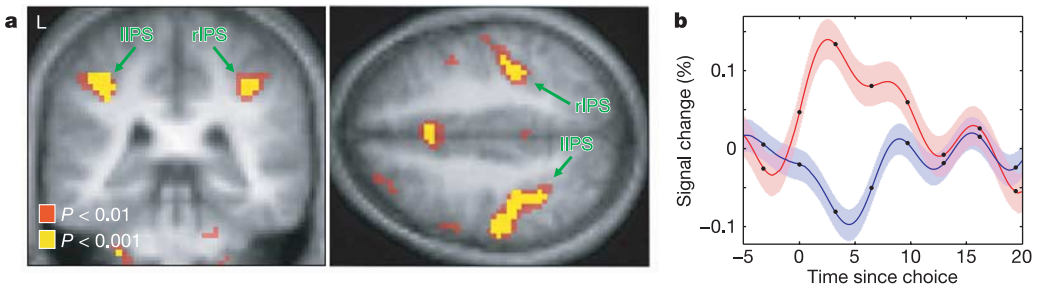

Figure 4 | Exploration-related activity in intraparietal sulcus. a, Regions of left and right intraparietal sulcus (lIPS and rIPS) showing significantly increased activation on exploratory compared with exploitative trials. Activation maps (yellow, $P < 0.001$; red, $P < 0.01$) are superimposed on a 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).