Submitted, Nature

Polar exploration: An anterior prefrontal substrate for exploratory decisions in humans

Nathaniel D. Daw[1,*], John P. O'Doherty[2,3,*], Peter Dayan[1], Ben Seymour[2], and Raymond J. Dolan[2]

1. Gatsby Computational Neuroscience Unit, Alexandra House, 17 Queen Square, London WC1N 3AR, UK

2. Wellcome Department of Imaging Neuroscience, UCL, 12 Queen Square, London WC1N 3BG, UK

3. Present affiliation: Division of Humanities and Social Sciences, California Institute of Technology, 1200 E California Blvd, Pasadena, CA 91125, USA.

* The first two authors contributed equally to this work, and ordering was determined arbitrarily.

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 must balance the desire to select what appears the richest option on the basis of accumulated experience, against the desire to choose a less familiar but potentially more advantageous option. Substantial neural, behavioral and computational evidence suggests a dopaminergic/striatal network mediates learning to exploit. However, the neural and computational substrates of exploration, a key and computationally challenging aspect of behavioral flexibility, are less clear. In a human gambling task, we show that subjects' choices can be characterized by an approximate solution to the dilemma that is well studied in the computational literature. Furthermore, by using this computational model to classify decisions as exploratory or exploitative, we show that an area of the frontopolar cortex is differentially active during exploratory decisions. In contrast, striatal and more medial-caudal prefrontal regions exhibit activity characteristic of an involvement in value-based decision making. The results suggest a model of action selection under uncertainty that involves switching between exploratory and exploitative behavioral modes, and indicate a computationally precise regulatory role for frontopolar cortex in exploration.

Much is known about exploitative aspects of appetitive choice, particularly in the mammalian brain. A wealth of studies suggest the involvement of two systems: the striatum[2–6] and particularly its dopaminergic innervation[7,8], which learns the utilities of actions, and expresses habits[9]; and various regions of prefrontal cortex, which are associated with more reflective planning[10]. Substantial reinforcement learning[11] (RL) theory, including the temporal difference algorithm (TD) underpins both the nature[7,8] and the interaction[12] of these systems.

However, this knowledge sheds little light on one of the most pervasive characteristics of adaptive behavior, namely a proclivity to explore. Far from representing idle curiosity, exploration is critical for animals to discover how best to harvest resources, such as food or water, in an unknown and dynamic environment. Although well studied both ecologically[13] and theoretically[1], the neural mechanisms that underpin exploratory action are far from clear. Exploration is an informationally and computationally refined capacity that demands careful regulation. On this basis we can conjecture an involvement of areas that influence and modulate the reflective planning system, notably regions in the anterior prefrontal cortex, including the frontal pole[14]. However, little is so far known about the specific computational functions of these areas.

We used a combined fMRI and behavioral paradigm to investigate the neural and computational substrates of exploratory behavior. We studied the patterns of brain activation in 14 healthy subjects in a four-armed bandit task in which participants made repeated choices between four slot machines (**Figure** 1; see **Supplementary Methods**). The slots paid off points (which the subjects were told would be exchanged for money) noisily around four different means. Unlike standard slots, the mean payoffs drifted randomly from trial to trial, with subjects finding information about the current worth of a slot only through actively sampling it. 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.

In RL[11], strategies for exploration come in three flavors, which differ in how exploratory actions are directed. The simplest methods, known as “ε-greedy,” are undirected: they choose the “greedy” option (that believed to be best) most of the time, but occasionally (with probability ε) substitute a random action. A more sophisticated approach is to guide exploration by value, as in the “softmax” rule. Here, the decision to explore and the choice of which suboptimal action to take are determined probabilistically based on the actions’ relative expected values. Finally, exploration can additionally be directed by uncertainty toward actions whose consequences are least known. The optimal strategy for a class of bandit tasks has this characteristic[1], as do standard heuristics[15] for exploration in more general RL tasks such as ours, for which the optimal solution is computationally intractable. Exploration of uncertain options can be worthwhile because the information received can allow for better future choices and payoffs.

When we asked subjects to describe their choice strategies in post-task interviews, the majority (11 out 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 behavior quantitatively, we compared the fit of three distinct RL models to our subjects' behavioral choices, embodying the aforementioned exploration strategies. All of the models learned the values of actions using a Kalman filter (see **Supplementary Methods**), an error-driven prediction algorithm that generalizes the TD algorithm by also tracking the uncertainty about the value of each action. The models differed only in their choice rules. We compared models using the likelihood of the subjects' choices given their experience, optimized over free parameters. This comparison (**Supplementary Table** 1) 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 toward uncertainty (softmax with an uncertainty bonus).

Having fit the softmax learning model to subjects' behavior, and having validated it against the other candidates, we subsequently used it to generate regressors containing value predictions, prediction errors, and choice probabilities for each subject on each trial. We used statistical parametric mapping (SPM) to identify brain regions in which neural activity correlated significantly with the model's signals. Consistent with previous studies[4–6], a prediction error correlated significantly with activity in both the ventral and dorsal striatum (see **Supplementary Table** 2). The results confirm a role for these structures in value learning and control. Other, cortical structures associated with this subcortical network[16] also showed significant value-related correlation. Specifically, we found activity in medial orbitofrontal cortex (OFC) correlated with the magnitude of the obtained payoff (**Figure** 2a), a finding consistent with prior evidence indicating that this region is involved in coding the relative value of different reward stimuli, including abstract rewards[17,18]. Furthermore, activity in medial and lateral OFC, extending into ventro-medial prefrontal cortex (PFC), correlated with the probability assigned by the model to the action actually chosen on a given trial (**Figure** 2b). This probability is a relative measure of the expected reward value of the chosen action, and the activity is thus consistent with a role for orbital and adjacent medial prefrontal cortex in encoding predictions of future reward[19,20].

We then sought to identify brain activity that selectively reflected whether actions were chosen for their exploratory or exploitative potential. We inferred that such a signal would reflect a form of control mechanism facilitating switching of behavioral strategies between different modes. We hypothesized that such a signal would be located in PFC, the principal cortical region known to mediate behavioral control[21]. 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 compared the pattern of brain activity associated with these exploratory and exploitative trials. This comparison revealed that a region of prefrontal cortex, incorporating anterior frontopolar cortex bilaterally (**Figure** 3a) was significantly more active during actions classified as exploratory. (Other areas outside prefrontal cortex showing activity related to exploration include medial and lateral premotor cortex, intraparietal sulcus and cerebellum; see **Supplementary Table 3** for details.) Activation in the right frontopolar cortex survived whole-brain correction (at $P<0.05$, corrected for multiple comparisons using false discovery rate). Average BOLD timecourses from this region (**Figure** 3b) demonstrate phasic increases and decreases in activity timelocked, respectively, to subjects' exploratory and exploitative decisions. No brain areas were found to exhibit higher activity during exploitation than exploration, that survived either whole-brain correction or the additional diagnostics discussed below.

Compared to exploitation, exploratory choices tend to favor less valuable, lower probability, and more uncertain targets. We investigated whether these potential confounds could explain our data in a post-hoc multiple linear regression analysis of frontopolar timecourses. We found that none of these variables could account for the differential frontopolar activity. The same analysis also ruled out other potential confounds such as stay versus switch, actual reward received, and reaction times. Interestingly, we did reveal that frontopolar activity following exploratory choices was correlated (negatively, $P<.01$) with the overall probability of exploration determined by the model, with the highest average response seen when exploration was chosen most against the odds. This observation, and a similar finding that activity in a region of dorsolateral PFC correlated negatively with the probability assigned by the model to the chosen action, is in keeping with the idea that additional cognitive control is needed to force exploration when exploitation seems most favorable.

These results have important implications for both computational and neural accounts of action selection. The finding that a brain region is discretely implicated in exploration (particularly a prefrontal, high-level control structure[14]) suggests a theory involving discrete switching between exploratory and exploitative modes, rather than more integrated accounts such as uncertainty bonus schemes[1,15]. These latter schemes tightly entangle exploration and exploitation, because they work by choosing actions with respect to a single value metric that prizes both information gathering and primary reward. The fact that we find neither behavioral nor neural evidence for such an account is relevant to theories of dopamine function, since the value of information in a unified exploratory framework has been suggested to explain the responses of dopamine neurons to novel stimuli[22]. Another neuromodulator, norepinephrine, has previously been suggested to regulate the overall tendency to explore[23,24], by controlling the gain in exactly the sort of softmax-style exploration rule implicated by our results.

Given a mode-switching exploration strategy, frontopolar cortex seems well suited for directing it. The specific role of this most rostral of prefrontal regions is still under vigorous study[14], but on the basis of previous imaging results, it is broadly associated with high-level control. Indeed this region has been hypothesized to sit atop a hierarchy of nested prefrontal controllers[25] implicated in evaluating one's internal thought processes[26] and in mediating between different goals[27], subgoals[28], or cognitive processes[14]. However, no previous work has offered a computationally specific account of the type advanced here. In keeping with our observations, frontopolar lesions are associated with problems in task switching[29]; it is also possible that exploratory problems contribute to perseverative deficits associated more generally with frontal dysfunction[30].

A striking feature of our data is that they suggest that even healthy subjects may be poor explorers. When exploratory decisions are guided by value at the cost of neglecting a proper accounting of uncertainty, as suggested here, choices can be pathological due either to excessive or insufficient exploration. Notably, as perhaps in many gambling situations, options well known to be only moderately worse than their alternatives appear to be persistently attractive.

Acknowledgments: We thank Jian Li, Samuel McClure, Brooks King-Casas, and P. Read Montague for generously sharing their unpublished data on exploration. Funding was from a Royal Society USA Research Fellowship (ND), the Gatsby Foundation (ND, PD), the EU BIBA project (ND, PD), and a Wellcome Trust Programme Grant (JOD, RD).

Figure Captions

Figure 1

(a) Illustration of timeline within a trial. At the beginning of each trial the four slots are presented. Subject's task is to choose one of the four slots. The chosen slot then spins. Three seconds later the outcome is revealed (number of points won). After a further second the screen is cleared. The next trial is triggered after a fixed trial length of 6 seconds and an additional variable inter-trial interval as described in the methods (mean 2 seconds).

(b) Example of mean payoffs that would be received for choosing each slot machine (four colored lines) on each trial, demonstrating their independent random drift. The payoff received on a particular trial is corrupted by further Gaussian noise around this mean.

Figure 2

(a) Regions of ventromedial prefrontal cortex (including medial and lateral OFC and adjacent medial prefrontal cortex) found to show significant correlations with the probability assigned by the computational model to the subject's choice of slot. Activation maps are shown superimposed on a subject-averaged structural scan. A statistical threshold of $p<0.001$ is shown in yellow, and at $p<0.01$ in red (to illustrate the full extent of the activations). Co-ordinates of the activated areas are: medial OFC [-3, 45, -18, peak z=5.62], lateral OFC (not shown in figure): [45, 36, -15, peak z=4.6], medial PFC [-3, 33, -6, peak z=4.62]. The correlation for the medial PFC region is also illustrated as a bar plot, showing average BOLD response to decision as a function of choice probability.

(b) Region of medial OFC correlating significantly with the magnitude of points received at the time that the outcome is revealed. Activation maps are shown superimposed on a subject-averaged structural scan. A statistical threshold of $p<0.001$ is shown in yellow, and at $p<0.01$ in red (to illustrate the full extent of the activations). Co-ordinates of the activated area is: [3,30,-21, peak z=3.87]. The correlation is also illustrated as a bar plot, showing average BOLD response to outcome as a function of payoff amount.

Figure 3

(a) Regions of left and right frontopolar cortex (lFP and rFP respectively) showing significantly increased activation on exploratory compared to exploitative trials. Activation maps are shown superimposed on a subject averaged structural scan. A statistical threshold of p<0.001 is shown in yellow, and at p<0.01 in red (to illustrate the full extent of the activations). Co-ordinates of activated areas are: lFP [-27 48 4, peak z = 3.49] and rFP [27 57 6, peak z = 4.13].

(b) Phasic increases and decreases in frontopolar BOLD signal, timelocked to exploratory and exploitative choices. Percent signal changes from peak frontopolar voxel, were upsampled and aligned to choices, and averaged over 1,516 exploratory and 2,647 exploitative trials. Colored fringes denote error bars (SEM); black dots indicate the sampling frequency (3.24 secs-1) of the original signals (though sample alignment varied from trial to trial).

References

1. Gittins, J.C. & Jones, D. A dynamic allocation index for the sequential design of experiments. in *Progress in Statistics* (ed. Gani, J.) 241–266 (North Holland, Amsterdam, 1974).

2. 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).

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

4. 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).

5. 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).

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

7. 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).

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

9. Packard, M.G. & Knowlton, B.J. Learning and memory functions of the basal ganglia. *Annu. Rev. Neurosci.* **25**, 563-593 (2002).

10. Owen, A.M. Cognitive planning in humans: Neuropsychological, neuroanatomical and neuropharmacological perspectives. *Prog. Neurobiol.* **53**, 431-450 (1997).

11. Sutton, R.S. & Barto, A.G. *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998).

12. Daw, N.D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* (2005). (in press).

13. Charnov, E.L. Optimal foraging: The marginal value theorem. *Theor. Popul. Biol.* **9**, 129-136 (1976).

14. Ramnani, N. & Owen, A.M. Anterior prefrontal cortex: Insights into function from anatomy and neuroimaging. *Nature Reviews Neuroscience* **5**, 184–194 (2004).

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

16. 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).

17. O'Doherty, J., Kringelbach, M.L., Rolls, E.T., Hornak, J. & Andrews, C. Abstract reward and punishment representations in the human orbitofrontal cortex, *Nat. Neurosci.* **4**, 95-102 (2001).

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

19. Gottfried, J.A., O'Doherty, J. & Dolan, R.J. Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* **301**, 1104-1107 (2003).

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

21. Miller, E.K. and Cohen, J.D., An integrative theory of prefrontal cortex function. *Annu. Rev. Neurosci.* **24**, 167-202 (2001).

22. Kakade, S. & Dayan, P. Dopamine: Generalization and bonuses. *Neural Netw.* **15**, 549–559 (2002).

23. 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).

24. Doya, K. Metalearning and neuromodulation. *Neural Netw.* **15**, 495-506 (2002).

25. Koechlin, E., Ody, C. & Kouneiher, F.A. The architecture of cognitive control in the human prefrontal cortex. *Science* **302**, 1181–1185 (2003).

26. Simons, J.S., Owen, A.M., Fletcher, P.C. & Burgess, P.W. Anterior prefrontal cortex and the recollection of contextual information. *Neuropsychologia* **43**, 1774–1783 (2005).

27. Koechlin, E., Basso, G., Pietrini, P., Panzer, S. & Grafman, J. The role of the anterior prefrontal cortex in human cognition. *Nature* **399**, 148–151 (1999).

28. Braver, T.S. & Bongiolatti, S.R. The role of frontopolar cortex in subgoal processing during working memory. *Neuroimage* **15**, 523-536 (2002).

29. Burgess, P.W., Veitch, E., de Lacy Costello, A. & Shallice, T. The cognitive and neuroanatomical correlates of multitasking. *Neuropsychologia* **38**, 848–863 (2000).

30. Stuss, D.T. *et al.* Wisconsin Card Sorting Test performance in patients with focal frontal and posterior brain damage: Effects of lesion location and test structure on separable cognitive processes. *Neuropsychologia* **38**, 388–402 (2000).

Figure 1

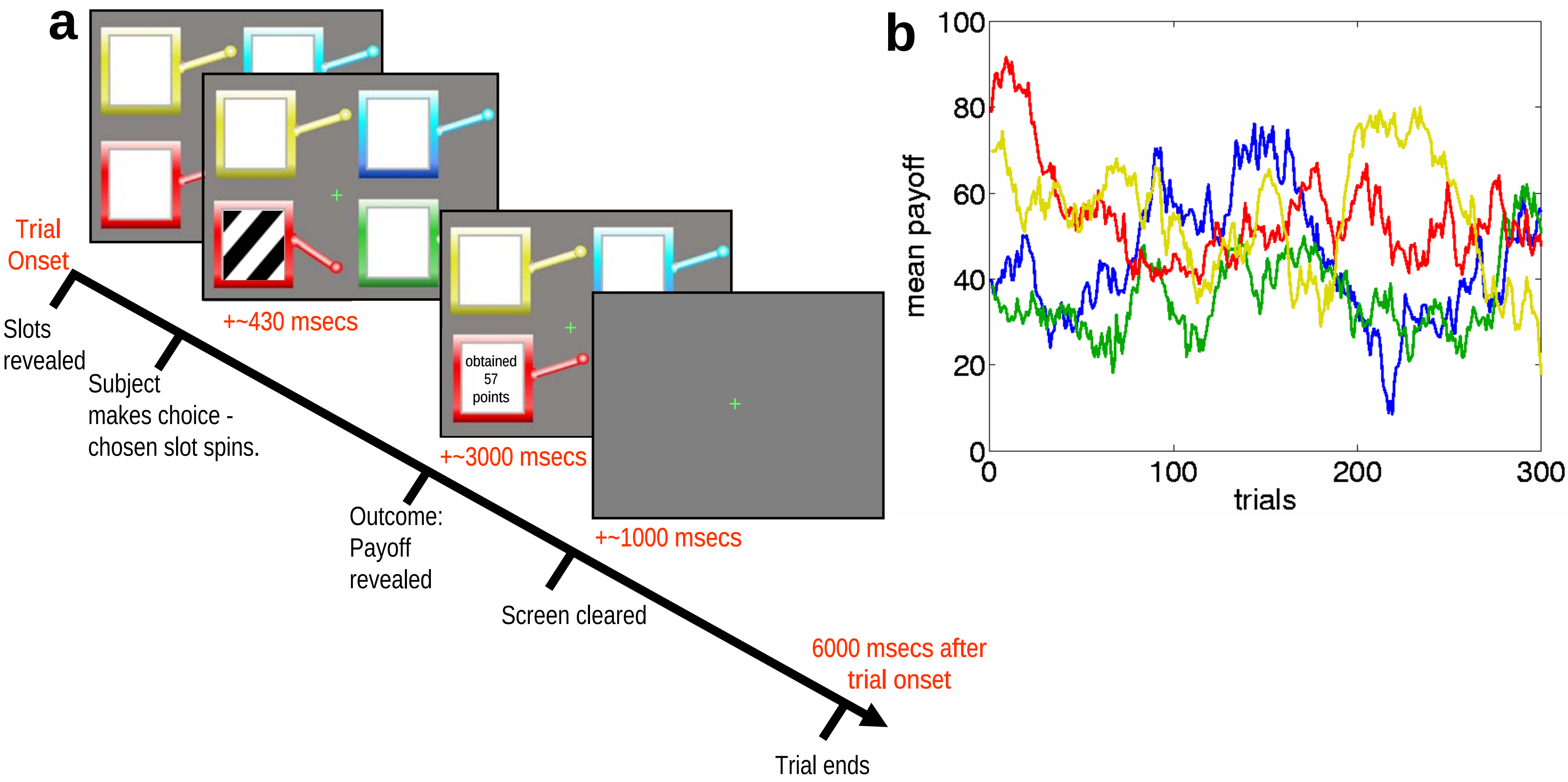

a
Trial
Onset
Slots
revealed
+~430 msecs
Subject
makes choice -
chosen slot spins.
obtained
57
points
+~3000 msecs
Outcome:
Payoff
revealed
+~1000 msecs
Screen cleared
6000 msecs after
trial onset
Trial ends
b
mean payoff
100
80
60
40
20
0
0
100
200
300
trials

Figure 2

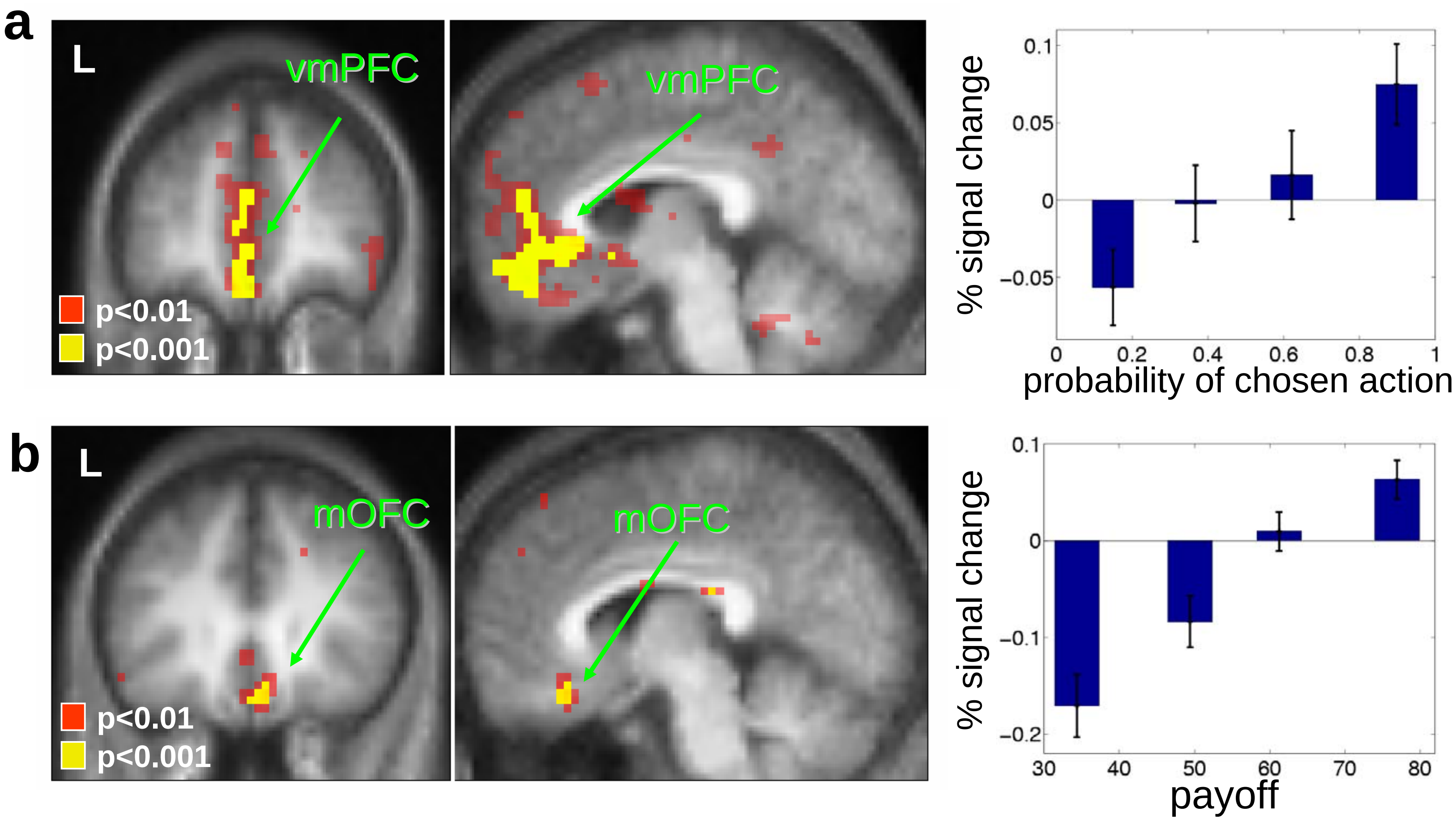
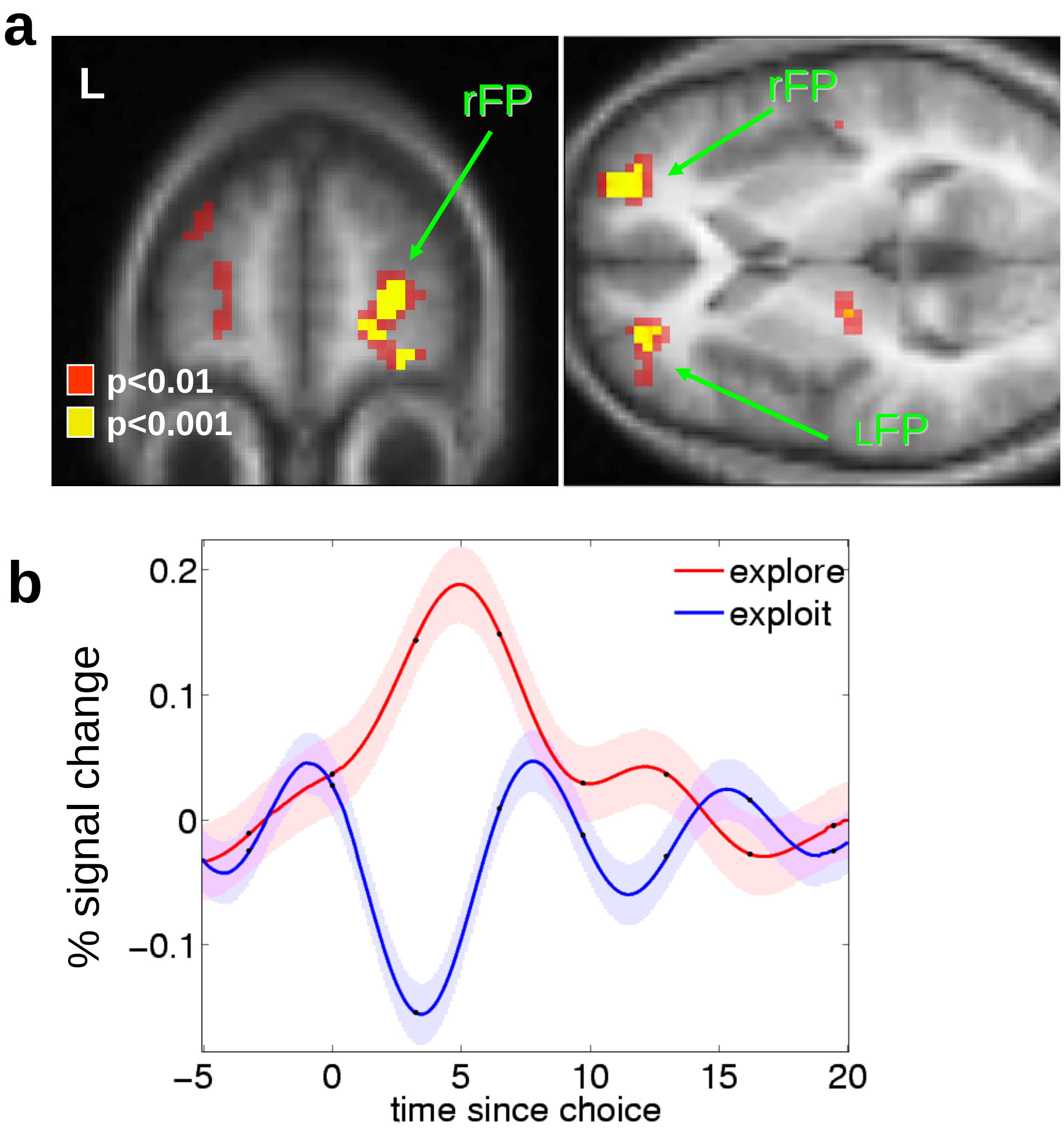

a
L
vmPFC
vmPFC
🟥 p<0.01
🟨 p<0.001
% signal change
0.1
0.05
0
−0.05
0
0.2
0.4
0.6
0.8
1
probability of chosen action
b
L
mOFC
mOFC
🟥 p<0.01
🟨 p<0.001
% signal change
0.1
0
−0.1
−0.2
30
40
50
60
70
80
payoff

Figure 3


a
L
rFP
rFP
LFP
🟥 p<0.01
🟨 p<0.001
b
explore
exploit
0.2
0.1
0
−0.1
% signal change
−5
0
5
10
15
20
time since choice

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 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 differently 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 σ_o = 4) around a mean $\mu_{i,t}$ and rounded to the nearest integer. At each timestep, the means drifted 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 drift noise v was zero-mean Gaussian (standard deviation σ_d = 2.8). Each subject was exposed to one of three instantiations of this process; one is illustrated in **Figure** 1B.

Kalman filter model

The Kalman filter[1] is the Bayesian mean-tracking rule for the drift process described 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 drift process, the prior distributions on the subsequent trial are given by $\hat{\mu}_{i,t+1}^{pre} = \lambda\hat{\mu}_{i,t}^{post} + (1-\lambda)\theta$ and $\hat{\sigma}_{i,t+1}^{2\,pre} = \lambda^2\hat{\sigma}_{i,t}^{2\,post} + \sigma_d^2$ 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 $\hat{\sigma}^2$, 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:

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

Behavioral analysis

We evaluated the three models using Bayesian model comparison techniques[3]. We took the inference parameters λ, σ_d, $\hat{\mu}_{i,0}^{pre}$, $\hat{\sigma}_{i,0}^{pre}$ and θ to be free, together with the choice parameters ε or β, and φ. For each model, we fit these (holding σ_o constant due to model degeneracy) 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.)

Due to the relatively small number of trials per subject, we fit the behavior of all subjects using a single instance of most of the model parameters (λ, σ_d, θ, $\hat{\mu}_{i,0}^{pre}$, $\hat{\sigma}_{i,0}^{pre}$ and φ), but to account partly for subject heterogeneity, we fit the parameter controlling the “noisiness” of choices (β or ε) individually for each subject. A combination of nonlinear optimization algorithms (Matlab optimization toolbox) was used 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[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 report results for an approximation in which the “max” operation was replaced with a very sharp softmax, $P_{i,t} = \varepsilon + (1-4\varepsilon)\cdot \exp(\beta\hat{\mu}_{i,t}^{pre}) / \sum_j \exp(\beta\hat{\mu}_{j,t}^{pre})$ (for a fixed β = 1,000). Investigation of the original rule using grid search and around local maxima of the approximate rule indicated that the likelihood reported is generous, i.e., that pure ε-greedy fits even more poorly.

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[6]. 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). 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[7].

For the post-hoc analysis of timecourse data from the frontopolar region, raw signal timecourses were extracted from this 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 and averaged and plotted. For the post-hoc multiple regression analysis, upsampled timecourses were projected onto a 32-sec. canonical hemodynamic response function aligned to each choice. The resultant per-trial response estimates were then used as targets for the multiple regression procedure. The same procedure was used to extract timecourses and per-trial response estimates for the medial PFC and OFC regions, which were grouped in evenly spaced bins and averaged to produce the bar plots in Figure 2.

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

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

	ε-greedy	softmax	uncertainty
-LL	4130.4	3972.1	3972.1
pseudo-r^2	0.28396	0.31141	0.31141
# parameters	19	19	20
BIC	4209.6	4051.3	4055.4

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

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

Explore > Exploit		**MNI co-ordinates**			
	Side	**X**	**Y**	**Z**	**Z-score**
Premotor cortex - Area 6	L	-57	3	36	4.92
Premotor cortex - Area 6	L	3	9	51	4.36
Intra-parietal sulcus	L	-29	-33	45	4.39
	R	39	-36	42	4.16
Cerebellum	R	21	-54	-30	5.42
	R	18	-57	-51	4.28

Supplementary Table 3: Regions outside prefrontal cortex 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.