Cognitive, Affective, & Behavioral Neuroscience (2018) 18:117–126
https://doi.org/10.3758/s13415-017-0556-2

Pure correlates of exploration and exploitation in the human brain

Tommy C. Blanchard[1] · Samuel J. Gershman[1]

Published online: 7 December 2017

Abstract
Balancing exploration and exploitation is a fundamental problem in reinforcement learning. Previous neuroimaging studies of the exploration–exploitation dilemma could not completely disentangle these two processes, making it difficult to unambiguously identify their neural signatures. We overcome this problem using a task in which subjects can either observe (pure exploration) or bet (pure exploitation). Insula and dorsal anterior cingulate cortex showed significantly greater activity on observe trials compared to bet trials, suggesting that these regions play a role in driving exploration. A model-based analysis of task performance suggested that subjects chose to observe until a critical evidence threshold was reached. We observed a neural signature of this evidence accumulation process in the ventromedial prefrontal cortex. These findings support theories positing an important role for anterior cingulate cortex in exploration, while also providing a new perspective on the roles of insula and ventromedial prefrontal cortex.

Keywords reinforcement learning · fMRI · decision making

Many decision problems pose a fundamental dilemma between exploration and exploitation: An agent can exploit the option that has yielded the greatest reward in the past or explore other options that may yield greater reward, at the risk of foregoing some reward during exploration. The optimal solution to the exploration–exploitation dilemma is generally intractable, and hence resource-bounded agents must apply heuristic strategies (Cohen, McClure & Yu, 2007). The specific strategy used by humans is an open question.

Some evidence suggests that humans adopt exploration strategies that sample options with probability proportional to their estimated expected values (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006) or their posterior probability of having the maximum value (Speekenbrink & Konstantinidis, 2015). Other studies suggest that humans employ an uncertainty-driven exploration strategy based on an explicit exploration bonus (Badre, Doll, Long, & Frank, 2012; Frank, Doll, Oas-Terpstra, & Moreno, 2009). Humans also sometimes employ more sophisticated exploration strategies using model-based reasoning (Knox, Otto, Stone, & Love, 2012; Otto, Knox, Markman, & Love, 2014; Wilson, Geana, White, Ludvig, & Cohen, 2014; Gershman & Niv, 2015).

Neural data can potentially constrain the theories of exploration by identifying dissociable correlates of different strategies. For example, Daw et al. (2006) identified a region of frontopolar cortex that was significantly more active for putative exploratory choices compared to putative exploitative choices during a multiarmed bandit task (see also Boorman, Behrens, Woolrich, & Rushworth, 2009). Suppression of activity in this region, using transcranial direct current stimulation, reduces exploration, whereas amplifying activity increases exploration (Beharelle, Polania, Hare, & Ruff, 2015). These findings suggest that there may exist a dedicated neural mechanism for driving exploratory choice, analogous to regions in other species that have been found to inject stochasticity into songbird learning (Olveczky, Andalman, & Fee, 2005; Woolley, Rajan, Joshua, & Doupe, 2014) and rodent motor control (Santos, Oliveira, Jin, & Costa, 2015).

The main challenge in interpreting these studies is that exploratory and exploitative choices cannot be identified unambiguously in standard reinforcement learning tasks, such as multiarmed bandits. When participants fail to choose the value-maximizing option, it is impossible to know whether this choice is due to exploration or to random error (i.e., unexplained variance in choice behavior not captured by the model). The same ambiguity muddies

✉ Samuel J. Gershman
gershman@fas.harvard.edu

[1] Department of Psychology and Center for Brain Science, Harvard University, 52 Oxford St., Room 295.05, Cambridge, MA 02138, USA

the interpretation of individual differences in parameters governing exploration strategies (e.g., the temperature parameter in the softmax policy). Furthermore, exploitative choices yield information, whereas exploratory choices yield reward, obscuring the conceptual difference between these trial types. Finally, identifying deviations from value maximization depend on inferences about subjective value estimates, which in turn depend on assumptions about the exploration strategy. Thus, there is no theory-neutral way to contrast neural activity underlying exploration and exploitation in most reinforcement learning tasks.

We resolve this problem by using an "observe or bet" task that unambiguously separates exploratory and exploitative choice (Navarro, Newell, & Schulze, 2016; Tversky & Edwards, 1966). On each trial, the subject chooses either to observe the reward outcome of each option (without receiving any of the gains or losses) or to bet on one option, in which case she receives the gain or loss associated with the option at the end of the task. By comparing neural activity on observe and bet trials, we obtain pure correlates of exploration and exploitation, respectively. This also allows us to look at neural responses to the receipt of information without it being confounded with the receipt of reward. To gain further insight into the underlying mechanisms, we use the computational model recently developed by Navarro et al. (2016) to generate model-based regressors. In particular, we identify regions tracking the subject's change in belief about the hidden state of the world, which in turn governs the subject's exploration strategy.

It is important to clarify at the outset that the correlates we identify are "pure" only in the sense that exploratory observe trials do not involve value-based choice or reward receipt, while exploitative bet trials do not involve information acquisition. This is not, of course, a complete catalogue of cognitive processes involved in task performance, and both trial types surely involve a number of common processes (e.g., visual perception, memory retrieval, motor control). Our goal in this study is to isolate a subset of these processes that are central to theories of reinforcement learning.

Materials and method

Subjects

We recruited 18 members of the Harvard community, through the Harvard Psychology Study Pool, to participate in the study. Eleven of the 18 subjects were female. Ages ranged from 21 to 36 years, with a median age of 26 years. All subjects were right-handed, native English speakers, and had no history of neurological or psychiatric disease. Informed consent was obtained from all subjects.

Task procedure

Subjects performed the task in two sessions. In the first session, subjects were familiarized with the task and performed five blocks outside of the fMRI scanner. In the second session, subjects performed two blocks of the task out of the scanner, and an additional four to five (depending on time constraints) in the scanner. Subjects were paid $10 for the first session and $35 for the second. They also received a bonus in the form of an Amazon gift card, at an amount of $0.10 per point earned in the task.

Subjects performed a dynamic version of the "observe or bet" task (Tversky & Edwards, 1966; Navarro et al., 2016). In this task, subjects were asked to predict which of two lights (red or blue) will light up on a machine. On each trial, a single light is activated. The machine always has a bias—on a particular block, it either will tend to light up the blue or the red light. On each trial, subjects could take one of three actions: bet blue, bet red, or observe. If the subjects bet blue or red, they gained a point if they correctly predicted which light would light up but lost one if they were incorrect. Importantly, they were not told if they gained or lost a point, and they also did not see what light actually lit up. Instead, subjects could only see which light was activated by taking the observe action. Observing did not cost any points, but subjects relinquished their opportunity to place a bet on that trial. Thus, subjects were compelled to choose between gaining information about the current bias (by observing) or using the information they had gathered up to that point to obtain points (by betting).

Each block consisted of 50 trials. On each block, the machine was randomly set to have a blue or a red bias. The biased color caused the corresponding light to be active on 80% of the trials. There was also a 5% chance that the bias would change during the block. This change was not signaled to the subject in any way and could only be detected through taking "observe" actions.

Computational model

To understand performance in our task mechanistically, we fit a computational model to the choice behavior, created to qualitatively match the features of the optimal decision strategy and shown to best fit subject behavior out of four candidate process models (Navarro et al., 2016). Central to the model is an evidence tally that starts with a value of zero. Positive evidence reflects evidence that the bias is blue, and negative reflects evidence that the bias is red. Thus, low absolute numbers reflect a state of uncertainty about the bias. Each time an observation is made, the evidence value changes by +1 if blue is observed and −1 if red is observed.

The relevance of old observations diminishes over time, modeled using an evidence decay parameter, α. The evidence

decay parameter dictates what proportion of evidentiary value is lost on each trial. Thus, the evidence tally value is calculated as follows:

$$e_t = x_t + (1-\alpha)e_{t-1}, \quad (1)$$

where e is the evidence tally, t is the current trial, x_t is the observation on the current trial (zero if a bet action is taken), and α is the evidence decay parameter. This evidence accumulation process is an instance of the the linear operator learning rule that has a long history in theories of learning (Bush & Mosteller, 1951) and differs from typical error-driven reinforcement learning algorithms that have been used in most studies of reinforcement learning (e.g., Daw et al., 2006). A number of studies have suggested evidence decay over time, which can capture perseverative tendencies of human subjects (Erev & Roth, 1998; Worthy, Pang & Byrne, 2013).

The other main component of the model is a decision threshold. The threshold is a value at which the learner will switch from observing to betting. In the model used here (the best fitting model reported in Navarro et al., 2016), the decision threshold follows a piece-wise linear structure across trials: it remains constant until a specific trial, at which point it changes at a constant rate until the final trial. The initial threshold, the trial at which the threshold begins changing (the change point), and the terminal value of the threshold are all parameters fit to the data.

Finally, because decision-makers are noisy, we also include a response stochasticity parameter, σ. Assuming a normally distributed noise term for each trial, n_t, with a zero mean and a standard deviation of σ, the probability of betting blue is then:

$$p(bet\ blue) = P(e_t + n_t \geq d_t) = \Phi\left(\frac{e_t + d_t}{\sigma}\right), \quad (2)$$

where e_t, n_t, and d_t are the evidence tally, decision noise, and the decision boundary on trial t, respectively, and Φ is the cumulative distribution function for a standard normal distribution.

Following Navarro et al. (2016), we used hierarchical Bayesian methods to estimate individual model parameters from the blocks performed outside the scanner. For the i-th subject, we set the priors on our model's parameters as follows (these are the same priors used by Navarro et al., 2016). For the response stochasticity parameter:

$$\begin{aligned} \sigma_i &\sim Exp(\lambda) \\ \lambda &\sim Gamma(1,1). \end{aligned} \quad (3)$$

For the evidence decay parameter:

$$\begin{aligned} \alpha_i &\sim Beta(a_1+1, a_2+1) \\ a_j &\sim Gamma(1,1). \end{aligned} \quad (4)$$

For the initial value of the decision threshold, d_{0i}:

$$\begin{aligned} d_{0i} &\sim Gamma(g_{01}, g_{02}) \\ g_{0j} &\sim Exp(1). \end{aligned} \quad (5)$$

For the terminal value of the decision threshold, d_{1i}:

$$\begin{aligned} d_{1i} &\sim Gamma(g_{11}, g_{12}) \\ g_{1j} &\sim Exp(1). \end{aligned} \quad (6)$$

For the threshold changepoint, c_i:

$$\begin{aligned} c_i &\sim Beta(b_1+1, b_2+1) \\ b_j &\sim Gamma(1,1). \end{aligned} \quad (7)$$

We implemented the model in Stan (Stan Development Team, 2016) and used Markov chain Monte Carlo sampling to approximate the posterior distribution over parameters. For the fMRI analysis, we used the posterior median parameter values for each subject to generate model-based regressors.

fMRI acquisition

Neuroimaging data were collected using a 3 Tesla Siemens Magnetom Prisma MRI scanner (Siemens Healthcare, Erlangen, Germany) with the vendor's 32-channel head coil. Anatomical images were collected with a T1-weighted magnetization-prepared rapid gradient multiecho sequence (MEMPRAGE, 176 sagittal slices, TR = 2530 ms, TEs = 1.64, 3.50, 5.36, and 7.22ms, flip angle = 7°, 1 mm 3 voxels, FOV = 256 mm). All blood-oxygen-level-dependent (BOLD) data were collected via a T2*-weighted echo-planar imaging (EPI) pulse sequence that employed multiband RF pulses and Simultaneous Multi-Slice (SMS) acquisition (Feinberg et al., 2010; Moeller et al., 2010; Xu et al., 2013). For the task runs, the EPI parameters were 69 interleaved axial-oblique slices (25 degrees toward coronal from ACPC alignment, TR = 2000 ms, TE = 35 ms, flip angle = 80°, 2.2 mm^3 voxels, FOV = 207 mm, SMS = 3). The SMS-EPI acquisitions used the CMRR-MB pulse sequence from the University of Minnesota.

fMRI preprocessing and analysis

Data preprocessing and statistical analyses were performed using SPM12 (Wellcome Department of Imaging Neuroscience, London, UK). Functional (EPI) image volumes were realigned to correct for small movements occurring between scans. This process generated an aligned set of images

and a mean image per subject. Each participant's T1-weighted structural MRI was then coregistered to the mean of the realigned images and segmented to separate out the gray matter, which was normalized to the gray matter in a template image based on the Montreal Neurological Institute (MNI) reference brain. Using the parameters from this normalization process, the functional images were normalized to the MNI template (resampled voxel Size 2-mm isotropic) and smoothed with an 8-mm full-width at half-maximum Gaussian kernel. A high-pass filter of 1/128 Hz was used to remove low-frequency noise, and an AR(1) (Autoregressive 1) model was used to correct for temporal autocorrelations.

We designed a general linear model model to analyze BOLD responses. This model included an event for observe decisions and another for bet decisions, time locked to the beginning of the decision period. We also included an event for the onset of feedback (either the observation of which light turned on, or just a visual of the machine with the bet that was made). For the onset of feedback, we included a parametric modulator that was the change in the absolute value of the evidence tally resulting from the observed outcome. Thus, this value would be negative and due entirely to evidence decay on a bet trial, and could be positive or negative on an observation trial depending on whether the observation provided more evidence in favor of betting or observing. Events were modeled with a 1-s duration.

Regions of interest

Regions of interest (ROIs) were constructed by combining structural ROIs with previously defined functional ROIs. Specifically, to define anatomically constrained value-based ROIs, we found the overlap between the structural ROIs from Tzourio-Mazoyer et al. (2002) and the value-sensitive functional ROIs from Bartra, McGuire, and Kable (2013). We also took the specific vmPFC and striatum ROIs from Bartra et al. (2013). For frontopolar cortex, we constructed a spherical ROI with a radius of 10 voxels, centered at the peak of activation reported by Daw et al. (2006). Similarly, for rostrolateral prefrontal cortex, the spherical ROI (10-voxel radius) was constructed using the coordinates given in Badre et al. (2012).

Code and data availability

Code and behavioral data are available on GitHub (https://github.com/TommyBlanchard/ObserveBet). The brain imaging data are available upon request.

Results

Behavioral results

Eighteen subjects performed a dynamic version of the "observe or bet" task (Fig. 1; see Materials and Method section for details). On each trial, subjects chose to either observe an outcome (without gaining or losing points) or bet on the outcome (without observing the outcome but redeeming points at the end of the experiment). The outcome probability had a small probability of changing during the course of each block of 50 trials.

Normative behavior on this task predicts several distinctive behavioral patterns (Navarro et al., 2016). On the first trial that subjects bet following a series of observe actions, they should bet on the color seen last. The intuition is that observing a color should either make your belief about the outcome probability stronger or weaker, and subjects should always bet on the outcome with the higher probability. If the subject observed on the previous trial, they were not certain enough to place a bet based on their current belief. Observing a surprising outcome (i.e., the outcome that is less strongly predicted by the subject's current belief) should push the belief towards the opposite decision threshold and therefore make the subject more likely to either observe or bet on the last-observed outcome. Indeed, subjects did strongly tend to bet on the last observed outcome on the first trial following an observe action, on average doing this 95.1% of the time (see Fig. 2a).

Subjects should also gradually reduce the probability of observing over the course of a block. This is because they start with no information about the outcome probability and thus must start by accumulating some information, but this tendency to explore will eventually yield to betting (exploitation) when the evidence becomes sufficiently strong. Again, subjects follow this pattern, observing 85.3% of the time on the first trial in a block and betting 98.4% on the final trial (see Fig. 2b).

Next, we implemented a previously developed computational model and fit it to subjects' choice data (Navarro et al., 2016). This model consists of an evidence tally that tracks how much evidence the learner currently has about the outcome probability, and a decision threshold that captures when the subject switches between observe and bet behaviors (see Fig. 2c). We fit this model to each subject's behavior from the prescanning blocks and used the fitted model to construct regressors for our fMRI analysis (see Method section). Behavior was stable across prescanning and scanning blocks (see Fig. 2d–e).

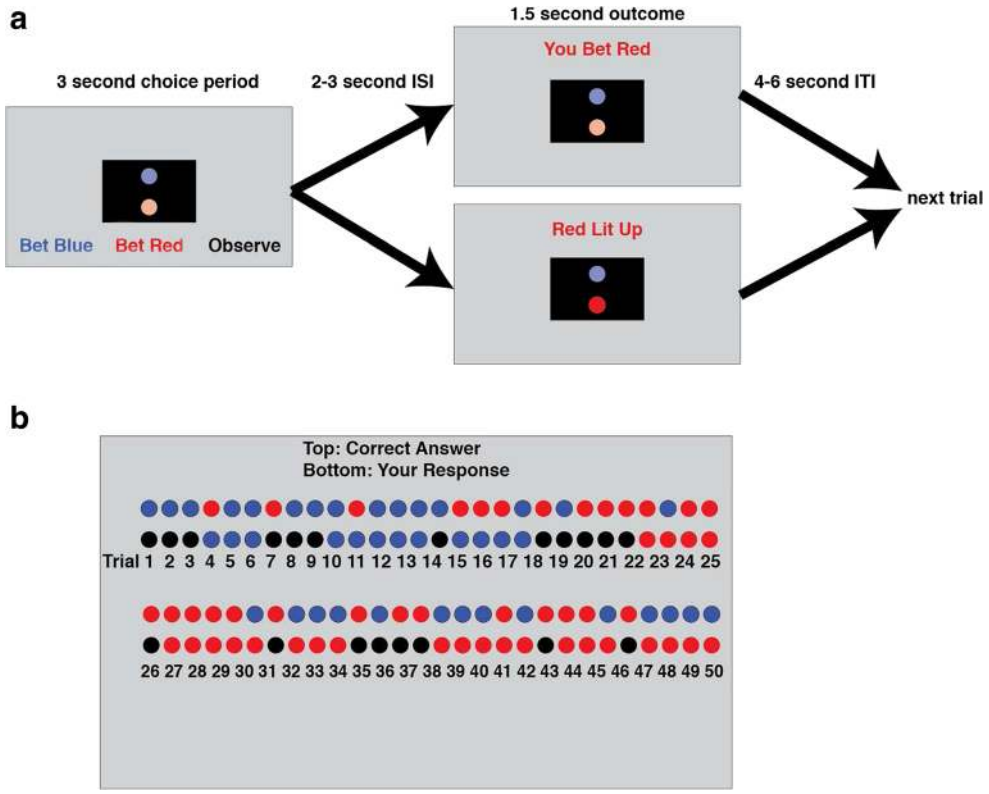

Fig. 1 **a** "Observe or bet" task. Subjects first made a choice between betting blue, betting red, or observing. They then waited through a variable-length interstimulus interval (during which nothing was on the screen). Then, for 1.5 s, subjects were shown the outcome of their action—if they bet, they were simply told which color they bet; if they observed, they were told which color lit up. This was followed by a variable length intertrial interval. **b** End of block score screen. At the end of each block of the task, subjects were shown what had happened on each trial. They saw one row of colored circles indicating what lit up on each trial, and a second row showing what their action had been on that trial (red or blue for betting, black for observing). They were also told their score for that block (for more details on the task, see the Materials and Method section). (Color figure online)

fMRI results

In a follow-up session, our 18 subjects returned and performed the "observe or bet" task in an fMRI scanner. Our model contained regressors for the appearance of stimuli, when a subject observed, when a subject bet, and the change in the absolute value of the evidence tally (see Materials and Method section).

We first attempted to identify regions associated with the decision to explore versus exploit (i.e., observe vs. bet). We chose to specifically investigate brain regions previously associated with value-based decision-making or exploration. Specifically, we examined the frontal pole and rostrolateral prefrontal cortex, which have both previously been implicated in balancing exploration and exploitation (Badre et al., 2012; Boorman et al., 2009; Daw et al., 2006; Donoso, Collins, & Koechlin, 2014). We also investigated the striatum, ventromedial prefrontal cortex (vmPFC), insula, and dorsal anterior cingulate cortex (dACC), all of which play a role in value-based decision-making (Bartra et al., 2013). We analyzed the signal in each of these ROIs, averaged across voxels (see Materials and Method section for details of ROI construction).

In each of our predefined ROIs, we calculated an observe–bet contrast for each subject and evaluated statistical significance using a one-sample t test. We found a significant positive effect (observe > bet) in insula and dACC ($t = 4.20$, $p < .001$ and $t = 2.80$, $p = .006$, respectively; see Table 1; Fig. 3a). The peaks of these effects were at 32, 22, −8 for the right insula; −30, 16, −8 for left insula; and 8, 16, 46 for dACC. The effects in all other ROIs did not pass the error-corrected threshold of $p < .008$ (Bonferroni correction with six comparisons and $\alpha = 0.05$). We then performed a whole-brain analysis with cluster family-wise error correction using the bspmview package (Spunt, 2016). We found a bilateral effect in thalamus that passed the error-corrected threshold of $p < .05$ (see Fig. 3b; peak at 8, −14, 2).

One potential concern with this analysis is that if people tend to switch from observing to betting more frequently than vice versa, any contrast between observe and bet trials would be confounded with task switching effects. Indeed, subjects

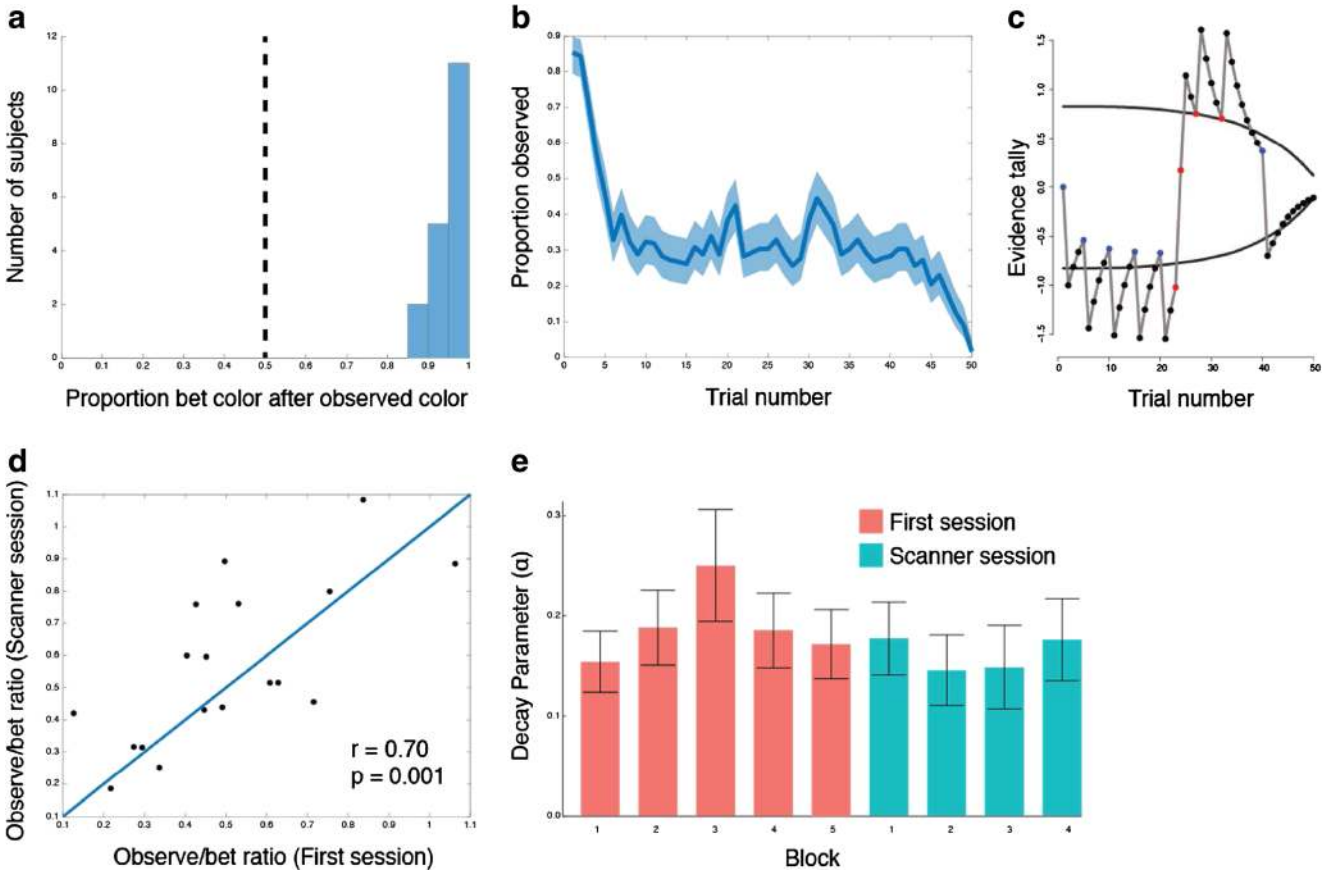

Fig. 2 Behavior on the "observe or bet" task. **a** Proportion of time each subject bet on the same color they observed on the previous trial. Vertical dashed line indicates random choice. **b** Proportion of trials subjects observed by trial number on each block (averaged across all subjects). Shaded region indicated the 95% confidence interval. **c** Visual representation of the model for one block. Circles indicate the action that was taken on that trial (black for bet, red for observed red, blue for observed blue). Gray line indicates the evidence tally on each trial. Black lines indicate the betting threshold (see Materials and Method section for model details). **d** Observe-to-bet ratio for each subject for the initial behavioral session and the scanner session. Line indicates the point of equality for the two sessions. **e** Average evidence decay parameter across all subjects for each block. (Color figure online)

were significantly more likely to switch following an observe trial ($p < .001$, signed rank test). If this lead to differential switch costs, then we would expect that responses should be slower on bet trials than on observe trials, consistent with the empirical data, $t(17) = 2.2$, $p < .05$ (mean difference: 48 ms). Thus, our data do not allow us to completely rule out a task switching confound.

Next, we investigated whether the BOLD signal in any regions covaried with changes in the absolute value of the evidence tally (a variable we termed the "update" regressor). In other words, we wanted to know which areas might be involved in using outcome information to update the decision policy. We again investigated the same six ROIs (see Table 2), finding a significant negative relationship between the update regressor and the BOLD signal in vmPFC ($t = -2.82$, $p = .006$; peak of cluster at −4, 36, −16). The negative effect means that vmPFC is more active when predictions are confirmed (i.e., updated less). No effects in any of our other ROIs passed Bonferroni correction. After examining these specific areas, we performed a whole-brain

Table 1 Table of values for the ROI analyses for the group-level observe–bet contrast

Brain region	t value	*p* value	Peaks	Cluster size (voxels)
Insula	4.20	<.001	32, 22, −8 (Right)	388 (Right)
			−30, 16, -8 (Left)	282 (Left)
Dorsal Anterior Cingulate	2.80	.006	8, 16, 46	307

Note. Degrees of freedom = 17. Bonferroni-corrected *p* value threshold with $\alpha = 0.05$ is 0.008. Significant effects were found in insula and dorsal anterior cingulate

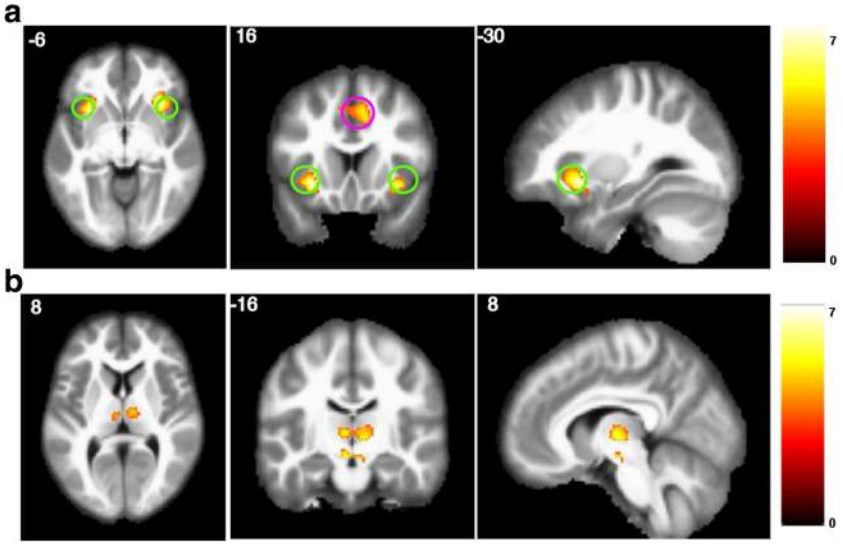

Fig. 3 Observe–bet contrast. **a** Clusters within the significant ROIs, with threshold set at $p < .001$, uncorrected. The ROI for insula is circled in green, the ROI for ACC is circled in magenta. **b** Whole-brain analysis with cluster family-wise error shows an effect in thalamus, peak activity at 8, −14, 2. (Color figure online)

analysis (see Fig. 4). No additional areas reached significance when performing whole-brain correction.

Discussion

Using a reinforcement learning task that cleanly decouples exploration and exploitation, our study provides the first pure neural correlates of these processes. Insula and dorsal anterior cingulate cortex showed greater activation for observe (exploration) trials compared to bet (exploitation) trials. Ventromedial prefrontal cortex showed greater activation for bet compared to observe trials, although this result did not survive correction for multiple comparisons across the regions of interest that we examined. We also found behavioral evidence favoring a heuristic approximation of the Bayes-optimal exploration strategy (Navarro et al., 2016): The probability of exploration changed dynamically as evidence was accumulated. These dynamics were accompanied by a neural correlate in the vmPFC that negatively correlated with the size of the belief update, suggesting that this region may encode the degree to which outcomes match prior expectations.

Table 2 Values for the ROI analyses for the "update" contrast

Brain region	*t* value	*p* value	Peaks	Cluster size (voxels)
vmPFC	**−2.82**	**.006**	−4, 36, −16	203

Note. Degrees of freedom = 17. Bonferroni-corrected *p* value threshold with $\alpha = 0.05$ is 0.008, a threshold that only the effect in vmPFC (highlighted in bold) passes.

The anterior cingulate cortex has figured prominently in past research on the exploration–exploitation dilemma, though its computational role is still unclear. Consistent with our findings, the anterior cingulate shows increased activity during exploration in multi-armed bandit (Amiez, Sallet, Procyk, & Petrides, 2012; Daw et al., 2006; Karlsson, Tervo, & Karpova, 2012; Quilodran, Rothe, & Procyk, 2008), foraging (Hayden, Pearson, & Platt, 2011; Kolling, Behrens, Mars, & Rushworth, 2012) and sequential problem-solving tasks (Procyk, Tanaka, & Joseph, 2000). Some evidence suggests that the anterior cingulate reports the value of alternative options (Blanchard & Hayden, 2014; Boorman, Rushworth, & Behrens, 2013; Hayden et al., 2011; Kolling et al., 2012); when this value exceeds the value of the current option, the optimal policy is to explore. Shenhav, Botvinick, and Cohen (2013) have argued that exploration is a control-demanding behavior, requiring an override of the currently dominant behavior in order to pursue long-term greater long-term rewards. In this framework, anterior cingulate reports the expected long-term value of invoking cognitive control.

The insula has also been implicated in several studies of the exploration–exploitation dilemma. Li, McClure, King-Casas, and Montague (2006) found insula activation in response to changes in reward structure during a dynamic economic game. These changes were accompanied by rapid alterations in the behavioral strategy. In a study of adolescents, Kayser, Op de Macks, Dahl, & Frank (2016) found that resting-state connectivity between rostrolateral prefrontal cortex and insula distinguished "explorers" from "nonexplorers" on a temporal decision-making task. Finally, using positron emission tomography while subjects performed a bandit task, Ohira et al. (2013) reported that insula activity was correlated both with

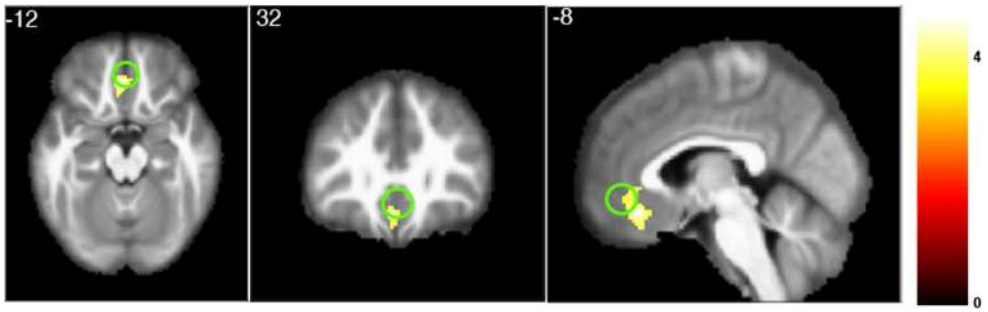

Fig. 4 Update contrast. Cluster within the significant ROI. Green circle shows the ROI for vmPFC. Threshold set at $p < .001$, uncorrected. (Colori figure online)

peripheral catecholamine concentration and response stochasticity. These results are consistent with our finding that insula was positively associated with exploration, though they do not provide insight into the region's specific contribution.

Surprisingly, we did not find a statistically significant effects of exploration in either frontopolar cortex or rostrolateral prefrontal cortex. Several influential studies have identified these regions as playing an important role in regulating exploration and exploitation (Badre et al., 2012; Beharelle et al., 2015; Boorman et al., 2009; Daw et al., 2006). It is not clear why we did not find effects in these regions; it is possible that our ROI selection procedure failed to identify the relevant voxels, or that these regions are primarily involved in other kinds of tasks (e.g., standard bandit or temporal decision-making tasks). One approach to this issue would be to define subject-specific functional ROIs using these other tasks and then interrogate regional responses using the observe or bet task. Another possibility is that substantive differences in task design and analysis account for the lack of activation. For example, Daw et al. (2006) defined exploratory versus exploitative trials based on whether subjects chose the option with highest expected value, whereas in our study subjects might choose options with either high or low expected value on exploratory trials.

Our model-based analysis posits that an important computation governing exploration is the updating of the belief state. We found a *negative* effect of updating in the vmPFC, indicating that this region was more active when expectations were confirmed. One way to interpret this finding is that the ventromedial prefrontal cortex signals a match between outcomes and expectations (i.e., a kind of "confirmation" or "match" signal). An analogous match signal has been observed in a visual same/different judgment task (Summerfield & Koechlin, 2008). In a related vein, Stern, Gonzalez, Welsh, and Taylor (2010) reported that signals in vmPFC correlated with "underconfidence" (the degree to which self-reported posterior probabilities underestimate objective posterior probabilities), consistent with the hypothesis that reduced updating will elicit greater vmPFC activity.

In the context of reinforcement learning and decision-making tasks, the ventromedial prefrontal cortex has more commonly been associated with reward expectation (Bartra et al., 2013) rather than outcome-expectation comparisons. Nonetheless, a number of studies have reported evidence accumulation correlates in this region or nearby regions (d'Acremont, Fornari, & Bossaerts, 2013; Chan, Niv, & Norman, 2016). More research is needed to pinpoint the relationship between these findings and exploration during reinforcement learning.

One limitation of our approach is that exploration is confounded with time: subjects are less likely to observe on later trials. A promising approach to dealing with this issue would be to use a yoked control condition in which subjects see the same sequence of trials without the trial types being contingent on their own actions (cf. Wang & Voss, 2014). However, this yoked control is imperfect insofar as it essentially eliminates the exploration–exploitation trade-off.

Another limitation of our approach is that we only considered a single model in detail, one developed specifically to approximate the Bayes-optimal strategy on the observe-or-bet task (Navarro et al., 2016). Navarro and colleagues compared this model to several variants, which differed in terms of their assumptions about evidence decay and decision thresholds. They concluded, on the basis of qualitative and quantitative measures of model fit, that both decaying evidence and declining thresholds were necessary to account for the choice data. Although this is still a fairly restricted space of models, it is worth pointing out that most conventional reinforcement learning models cannot address the task at all: Because the observe action does not accrue any points, it will always be assigned a value of zero by model-free algorithms like Q-learning. Nonetheless, the model developed by Navarro and colleagues invokes cognitive mechanisms that are shared across many other models, such as incremental adjustment of expectations (as in Q-learning) and decisions based on a stochastic threshold-crossing (as in sequential sampling models). The interface of these mechanisms has recently become an important focus of research in reinforcement learning (Frank et al., 2015; Pedersen, Frank, & Biele, 2017).

Finally, we must keep in mind that while the observe-or-bet task provides "pure" correlates by decoupling information acquisition and action selection, there are many other cognitive processes involved in exploration and exploitation, which

may be shared across observe and bet trials. Thus, we cannot decisively conclude that this contrast has perfectly isolated the critical computations underlying exploration and exploitation. It is unlikely that any single task will be able to achieve complete purity in this sense, so our findings should be understood as complementing, rather than superseding, previous studies of exploration and exploitation, all of which have their strengths and weaknesses.

In summary, the main contribution of our study is the isolation of neural correlates specific to exploration. The major open question is computational: What exactly do the insula and anterior cingulate contribute to exploration? As discussed in the preceding paragraphs, the literature is well-supplied with hypotheses, but our study was not designed to discriminate between them. Thus, an important task for future research will be to use tasks like "observe or bet" in combination with experimental manipulations (e.g., volatility or the distribution of rewards) that are diagnostic of underlying mechanisms.

Acknowledgements We are grateful to Joel Voss for helpful comments on an earlier draft. This research was carried out at the Harvard Center for Brain Science with the support of the Pershing Square Fund for Research on the Foundations of Human Behavior. This work involved the use of instrumentation supported by the NIH Shared Instrumentation Grant Program, Grant No. S10OD020039. We acknowledge the University of Minnesota Center for Magnetic Resonance Research for use of the multiband-EPI pulse sequences.

Compliance with ethical standards

Conflict of interest The authors declare no competing financial interests.

References

Amiez, C., Sallet, J., Procyk, E., & Petrides, M. (2012). Modulation of feedback related activity in the rostral anterior cingulate cortex during trial and error exploration. *NeuroImage, 63*, 1078–1090.

Badre, D., Doll, B. B., Long, N. M., & Frank, M. J. (2012). Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron, 73*, 595–607.

Bartra, O., McGuire, J. T., Kable, J. W. (2013). The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *NeuroImage, 76,* 412–27.

Beharelle, A. R., Polania, R., Hare, T. A., & Ruff, C. C. (2015). Transcranial stimulation over frontopolar cortex elucidates the choice attributes and neural mechanisms used to resolve exploration-exploitation trade-offs. *Journal of Neuroscience, 35*(43), 14544–14556.

Blanchard, T. C., & Hayden, B. Y. (2014). Neurons in dorsal anterior cingulate cortex signal postdecisional variables in a foraging task. *Journal of Neuroscience, 34,* 646–655.

Boorman, E. D., Behrens, T. E., Woolrich, M. W., & Rushworth, M. F. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron, 62*, 733–743.

Boorman, E. D., Rushworth, M. F., & Behrens, T. E. (2013). Ventromedial prefrontal and anterior cingulate cortex adopt choice and default reference frames during sequential multi-alternative choice. *The Journal of Neuroscience, 33*, 2242–2253.

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review, 58,* 313–323.

Chan, S. C. Y., Niv, Y., & Norman, K. A. (2016). A probability distribution over latent causes in the orbitofrontal cortex. *Journal of Neuroscience, 36,* 7817–7828.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362, 933–942.

d'Acremont, M., Fornari, E., & Bossaerts, P. (2013). Activity in inferior parietal and medial prefrontal cortex signals the accumulation of evidence in a probability learning task. *PLOS ONE,* 9, e1002895.

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

Donoso, M., Collins, A. G., & Koechlin, E. (2014). Human cognition: Foundations of human reasoning in the prefrontal cortex. *Science, 344*, 1481–1486.

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review, 88,* 848–881.

Feinberg, D. A., Moeller, S., Smith, S. M., Auerbach, E., Ramanna, S., Glasser, M. F., … Yacoub, E. (2010). Multiplexed echo planar imaging for sub-second whole brain fMRI and fast diffusion imaging. *PLOS ONE, 5,* e15710.

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, *12*, 1062–1068.

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T.V., Cavanagh, J. F., & Badre, D. (2015). fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *Journal of Neuroscience, 35,* 484–494.

Gershman, S. J., & Niv, Y. (2015). Novelty and inductive generalization in human reinforcement learning. *Topics in Cognitive Science, 7,* 391–415.

Hayden, B. Y., Pearson, J. M., & Platt, M. L. (2011). Neuronal basis of sequential foraging decisions in a patchy environment. *Nature Neuroscience*, *14*, 933–939.

Karlsson, M. P., Tervo, D. G. R., & Karpova, A. Y. (2012). Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. *Science, 338*, 135–139.

Kayser, A. S., Op de Macks, Z., Dahl, R. E., & Frank, M. J. (2016). A neural correlate of strategic exploration at the onset of adolescence. *Journal of Cognitive Neuroscience, 28,* 199–209.

Knox, W. B., Otto, A. R., Stone, P., & Love, B. C. (2012). The nature of belief-directed exploratory choice in human decision-making. *Frontiers in Psychology, 2,* 398.

Kolling, N., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2012). Neural mechanisms of foraging. *Science, 336*, 95–98.

Li, J., McClure, S. M., King-Casas, B., & Montague, P. R. (2006). Policy adjustment in a dynamic economic game. *PLOS ONE, 1,* e103.

Moeller, S., Yacoub, E., Olman, C. A., Auerbach, E., Strupp, J., Harel, N., & Uğurbil, K. (2010). Multiband multislice GE-EPI at 7 Tesla with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magnetic Resonance in Medicine, 63,* 1144–1153.

Navarro, D. J., Newell, B., & Schulze, C. (2016). Learning and choosing in an uncertain world: An investigation of the explore-exploit dilemma in static and dynamic environments. *Cognitive Psychology, 85,* 43–77.

Ohira, H., Matsunaga, M., Murakami, H., Osumi, T., Fukuyama, S., Shinoda, J., & Yamada J. (2013). Neural mechanisms mediating

association of sympathetic activity and exploration in decision-making. *Neuroscience, 246*, 362–374.

Olveczky, B. P., Andalman, A. S., & Fee, M. S. (2005). Vocal experimentation in the juvenile songbird requires a basal ganglia circuit. *PLOS ONE: Biology, 3,* 153.

Otto, A. R., Knox, W. B., Markman, A. B., & Love, B. C. (2014). Physiological and behavioral signatures of reflective exploratory choice. *Cognitive, Affective, & Behavioral Neuroscience, 14,* 1167–1183.

Pedersen, M. L., Frank, M. J., & Biele, G. (2017). The drift diffusion model as the choice rule in reinforcement learning. *Psychonomic Bulletin & Review,* 24(4), 1234–1251.

Procyk, E., Tanaka, Y. L., & Joseph, J. P. (2000). Anterior cingulate activity during routine and non-routine sequential behaviors in macaques. *Nature Neuroscience, 3*, 502–508.

Quilodran, R., Rothe, M., & Procyk, E. (2008). Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron, 57*, 314–325.

Santos, F. J., Oliveira, R. F., Jin, X., & Costa, R. M. (2015). Corticostriatal dynamics encode the refinement of specific behavioral variability during skill learning. *eLife, 4,* e09423.

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron, 79,* 217–240.

Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science, 7*, 351–367.

Spunt, B. (2016). spunt/bspmview: BSPMVIEW v.20161108. *Zenodo.* Retrieved from https://zenodo.org/record/168074

Stan Development Team (2016). RStan: The R interface to Stan (R Package Version 2.14.1) [Computer software]. Retrieved from http://mc-stan.org

Stern, E. R., Gonzalez, R., Welsh, R. C., & Taylor, S. F. (2010). Updating beliefs for a decision: Neural correlates of uncertainty and underconfidence. *Journal of Neuroscience, 30,* 8032–8041.

Summerfield, C. S., & Koechlin, E. (2008). A neural representation of prior information during perceptual inference. *Neuron, 59,* 336–347.

Tversky, A., & Edwards, W. (1966). Information versus reward in binary choices. *Journal of Experimental Psychology, 71*(5), 680–683.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., … Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage, 15,* 273–289.

Wang, J. X., & Voss, J. L. (2014). Brain networks for exploration decisions utilizing distinct modeled information types during contextual learning. *Neuron, 82,* 1171–1182.

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of Experimental Psychology: General, 143,* 2074–2081.

Woolley S. C., Rajan, R., Joshua, M., & Doupe, A. J. (2014). Emergence of context dependent variability across a basal ganglia network. *Neuron, 82,* 208–223.

Worthy, D. A., Pang, B., & Byrne, K. A. (2013). Decomposing the roles of perseveration and expected value representation in models of the Iowa gambling task. *Frontiers in Psychology*, *4*, 640.

Xu, J., Moeller, S., Auerbach, E. J., Strupp, J., Smith, S. M., Feinberg, D. A., … Ugurbil, K. (2013). Evaluation of slice accelerations using multiband echo planar imaging at 3 T. *NeuroImage, 83,* 991–1001.