Behavioral/Cognitive

# Primate Orbitofrontal Cortex Codes Information Relevant for Managing Explore–Exploit Tradeoffs

**Vincent D. Costa**[1] **and Bruno B. Averbeck**[2]

[1]Department of Behavioral Neuroscience, Oregon Health and Science University, Portland, Oregon 97239-3098, and [2]Laboratory of Neuropsychology, National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland 20892-4415

**Reinforcement learning (RL) refers to the behavioral process of learning to obtain reward and avoid punishment. An important component of RL is managing explore– exploit tradeoffs, which refers to the problem of choosing between exploiting options with known values and exploring unfamiliar options. We examined correlates of this tradeoff, as well as other RL related variables, in orbitofrontal cortex (OFC) while three male monkeys performed a three-armed bandit learning task. During the task, novel choice options periodically replaced familiar options. The values of the novel options were unknown, and the monkeys had to explore them to see if they were better than other currently available options. The identity of the chosen stimulus and the reward outcome were strongly encoded in the responses of single OFC neurons. These two variables define the states and state transitions in our model that are relevant to decision-making. The chosen value of the option and the relative value of exploring that option were encoded at intermediate levels. We also found that OFC value coding was stimulus specific, as opposed to coding value independent of the identity of the option. The location of the option and the value of the current environment were encoded at low levels. Therefore, we found encoding of the variables relevant to learning and managing explore– exploit tradeoffs in OFC. These results are consistent with findings in the ventral striatum and amygdala and show that this monosynaptically connected network plays an important role in learning based on the immediate and future consequences of choices.**

*Key words:* **decision-making; explore– exploit; monkey; orbitofrontal cortex; reinforcement learning**

**Significance Statement**

**Orbitofrontal cortex (OFC) has been implicated in representing the expected values of choices. Here we extend these results and show that OFC also encodes information relevant to managing explore– exploit tradeoffs. Specifically, OFC encodes an exploration bonus, which characterizes the relative value of exploring novel choice options. OFC also strongly encodes the identity of the chosen stimulus, and reward outcomes, which are necessary for computing the value of novel and familiar options.**

## Introduction

How do humans and animals decide when to explore a new opportunity or stick with what we know? Decisions to forego immediate rewards to explore options whose value is uncertain is known as the explore–exploit dilemma (Sutton and Barto, 1998). Exploration allows biological and artificial agents to resolve uncertainty and potentially discover more profitable alternatives (Daw et al., 2006; Wittmann et al., 2008; Djamshidian et al., 2011; Averbeck et al., 2013; Costa et al., 2014, 2019). Exploitative strategies emphasize an agents' ability to predict the immediate outcome of its choices. Exploitative strategies are effective in stable learning environments, but excessive reliance on prior knowledge impedes performance when circumstances change. Exploration facilitates learning when agents must make decisions under uncertainty. Excessive exploration, however, can impede performance by preventing agents from exploiting what they have newly learned. Managing explore–exploit tradeoffs is therefore a fundamental component of behavioral flexibility.

To balance exploration and exploitation biological agents need to know when exploration is advantageous. An efficient strategy for managing explore–exploit tradeoffs is to predict the immediate and future outcomes of each available choice option (Wilson et al., 2014b; Averbeck, 2015). Predicting whether choices will be immediately rewarded or unrewarded is easily computed based on past experience. Predicting how often choices are rewarded or unrewarded in the future is a more difficult computation, because it relies on prospection. Yet these predictions can be

Received Oct. 1, 2019; revised Jan. 26, 2020; accepted Feb. 9, 2020.

Author contributions: V.C. and B.A. designed research; V.C. and B.A. performed research; V.C. and B.A. analyzed data; V.C. and B.A. wrote the paper.

This work was supported by the Intramural Research Program of the National Institute of Mental Health (ZIA MH002928).

The authors declare no competing financial interests.

Correspondence should be addressed to Vincent D. Costa at costav@ohsu.edu.

https://doi.org/10.1523/JNEUROSCI.2355-19.2020

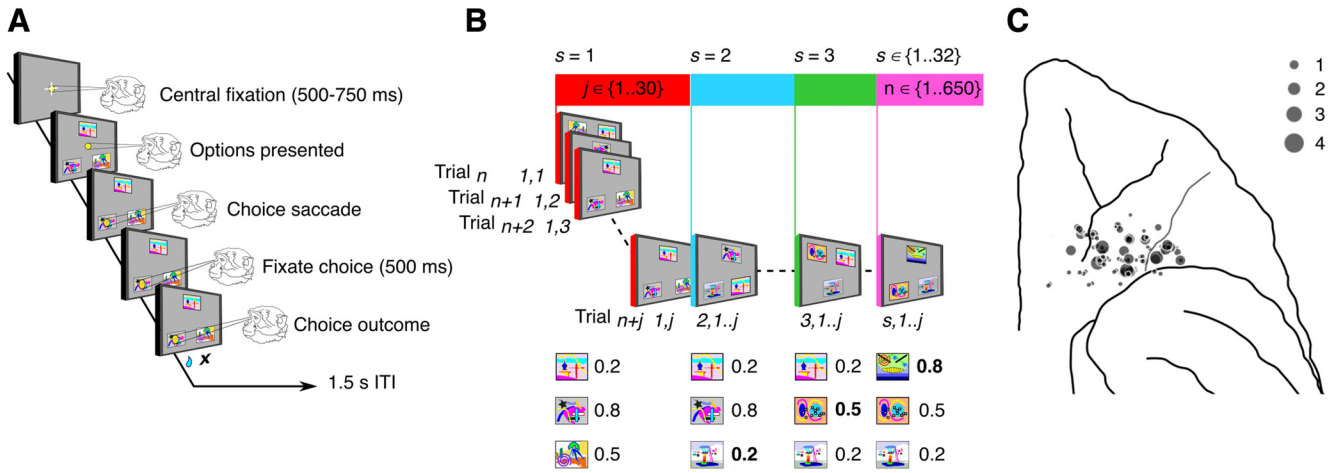


**Figure 1.** Task and recording locations. ***A***, Structure of an individual trial in the three-arm bandit task where monkeys indicated their choice by making a saccade to one of three options. Following each choice, the monkeys received either a fixed amount of juice reward with a probability conditioned on the stimulus or no reward. ***B***, Each block of 650 trials began with the presentation of three novel images. This set, *s*, of visual choice options was repeatedly presented to the monkey for a series of 10–30 trials. On a randomly selected trial between 10 and 30 one of the existing options was randomly selected and replaced with a novel image. This formed a new set of options that were presented for a series of 10–30 trials. Novel options were randomly assigned their own reward probabilities (0.2, 0.5, 0.8). Configurations in which all three options had the same reward probability were not allowed. This process of introducing a novel option to create a new set was repeated 32 times within a block. ***C***, MRI guided reconstruction of recording locations. Coronal T1-weighted MRI of electrodes lowered to specific depths for neuroimaging, were used to verify the trajectories and placement of the recording electrodes in each monkey. The number of cells recorded in the OFC at each site were projected on to template views from a standard macaque brain atlas.

integrated to decide when exploration is advantageous. During explore–exploit decision-making, such computations are encoded by neural activity in prefrontal cortical areas in humans (Daw et al., 2006; Zajkowski et al., 2017), as well as in the amygdala and ventral striatum in nonhuman primates (Costa et al., 2019).

The orbitofrontal cortex (OFC) is known to be important for encoding the subjective value of choices (Padoa-Schioppa and Assad, 2006; Schoenbaum et al., 2009; Padoa-Schioppa and Cai, 2011). It is also important for updating predictions about the outcomes of choices based on unobservable or partially observable information (Schuck et al., 2016). For instance, people and animals with OFC damage have difficulty learning unsignaled reversals in the contingencies between choices and outcomes (Iversen and Mishkin, 1970; Dias et al., 1996; Schoenbaum et al., 2003b; Groman et al., 2019), although in monkeys this is likely mediated by neighboring ventrolateral prefrontal cortex (Rudebeck et al., 2017a). OFC lesions also lead to deficits in realizing that a once valuable choice is no longer desirable, after the outcome it predicts is devalued outside the decision context (Rudebeck et al., 2013b, 2017a). OFC is hypothesized to be critical for integrating multisensory representations of choices and outcomes with information about past choices and outcomes in partially observable choice scenarios, to infer the current task state and make accurate predictions about the likely outcome of a particular choice (Wilson et al., 2014a). Although this theoretical view focuses on the role of OFC in making predictions about immediate outcomes it can be extended to its role in representing task states to make predictions about how current choices influence future outcomes. This suggests that OFC should encode computations relevant for managing explore–exploit tradeoffs. However, its role in exploratory decision-making has largely gone unexamined.

To determine the computational role of the OFC in explore–exploit decisions, we recorded neural activity from monkeys as they performed a task where explore–exploit tradeoffs were induced by introducing novel choice options. We used a model-based reinforcement learning algorithm, a partially observable Markov decision process (POMDP), to quantify the value of exploration and exploitation and related these values to the activity of individual neurons (Averbeck, 2015; Costa et al., 2019). We found that neuronal activity in OFC not only encodes the immediate value of choices based on what the monkeys had already learned, consistent with its known role in encoding economic value (Padoa-Schioppa and Cai, 2011), but also the potential future gains and losses associated with exploring novel choice options and the overall richness of the reward environment.

## Materials and Methods

*Subjects.* The experiments were performed on three adult male rhesus macaques (*Macca mulatta*), aged 6–8 years and weighing 7.2–9.3 kg. The monkeys were pair-housed when possible, and they had access to food 24 h/d. On days when recordings were performed, the monkeys earned their fluid through performance on the task. On non-testing days the monkeys were given *ad libitum* access to water. All procedures were reviewed and approved by the NIMH Animal Care and Use Committee.

*Experimental setup.* The monkeys were operantly trained to perform an oculomotor three-armed bandit task. The monkeys were seated in a primate chair facing a 19 inch LCD monitor (1024 × 768 resolution, 40 cm from the monkey's eyes), on which the visual stimuli were presented. Task control was performed using the Monkeylogic behavioral control system (Asaad and Eskandar, 2008). The monkey's eye movements were monitored using an Arrington Viewpoint eye tracking system (Arrington Research) and sampled at 1 kHz. Juice rewards (0.08–0.17 ml) were delivered using a pressurized juice delivery system (Mitz, 2005).

*Task design and stimuli.* The monkeys performed multiple blocks (650 trials each) of a three-armed bandit reinforcement learning task (Fig. 1) used previously (Costa et al., 2014, 2019) and based on a task originally used to study learning in human subjects (Wittmann et al., 2008; Djamshidian et al., 2011; Averbeck et al., 2013). During the task the monkey had to choose among three images. Each image had an associated probability of juice reward. In each trial, the monkey first acquired and held central fixation for a variable length of time (500–750 ms). After the fixation period, three peripheral choice targets were presented at the vertices of a triangle. To minimize spatial preferences, the main vertex of the triangle was randomly chosen to point up or down on each trial and the locations of the three stimuli were randomized from trial-to-trial. The monkeys were required to saccade to and maintain fixation on one of the peripheral targets for 500 ms. We excluded trials on which the

monkey made more than one saccade (<1% of all trials). After the response, a juice reward was delivered probabilistically. Within each block of 650 trials, 32 novel stimuli were introduced. A novel stimulus was introduced every 10–30 trials, with the interval chosen pseudorandomly. No single choice option could be available for >160 consecutive trials. When a novel stimulus was introduced it randomly replaced one of the existing choice options. At the start of a block, the three initial choice options were assigned reward probabilities of 0.2, 0.5, or 0.8. Novel options were pseudorandomly assigned one of these reward probabilities when they were introduced. The assigned reward probabilities were fixed for each stimulus. The only constraint was that there could not be three cues with the same probability of reward.

The visual stimuli were naturalistic scenes from the website Flickr (http://www.flickr.com). They were screened for image quality, discriminability, uniqueness, size, and color to obtain a final daily set of 210 images (35 images per block, up to 6 blocks). Images were never repeated across sessions. To avoid choices driven by perceptual pop out, choice options were spatial frequency and luminance normalized using functions adapted from the SHINE toolbox for MATLAB (Willenbockel et al., 2010).

*Neurophysiological recordings.* Monkeys were implanted with titanium headposts for head restraint. In a separate procedure, monkeys were fit with 28 × 36 mm recording chambers oriented to allow bilateral vertical grid access to the amygdala, ventral striatum and orbitofrontal cortex. The amygdala and ventral striatum data have been published separately (Costa et al., 2019). The OFC recordings were performed in partially overlapping recordings sessions. We recorded the activity of 146 single neurons from the OFC of 3 monkeys ($N = 12, 63$, and 71 single neurons across Monkeys H, F, and N). Chamber placements were planned and verified through MR T1-weighted scans of grid coverage with respect to the target structures.

In all three monkeys, we recorded from OFC using single tungsten microelectrodes (FHC or Alpha Omega; 0.8–1.5 MΩ at 1 kHz). The electrodes were advanced to their target location by an 8-channel micromanipulator (NAN Instruments) that was attached to the recording chamber. Additional MR T1-weighted scans with lowered electrodes were performed to verify recording trajectories (Fig. 1). Multichannel spike recordings were acquired with a 16-channel data acquisition system (Tucker Davis Technologies). Spike signals were amplified, filtered (0.3–8 kHz), and digitized at 24.4 kHz. Spikes were initially sorted online on all-channels using real-time window discrimination. Digitized spike waveforms and timestamps of stimulus events were saved for sorting off-line. Units were graded according to isolation quality and multiunit recordings were discarded. Neurons were isolated while the monkeys viewed a nature film and before they engaged in the bandit task. Other than isolation quality, there were no selection criteria for deciding whether to record a neuron.

*Data analysis.* The values of choice options in the task were modeled using an infinite horizon, discrete time, discounted, POMDP. Details of this model were published separately (Averbeck, 2015). The model estimates the utility, $u_t(s_t)$ associated with individual states, $s_t$:

$$u_t(s_t) = max_{a \in A_{s_t}} \left\{ r(s_t, a) + \gamma \sum_{j \in S} p(j \mid s_t, a) u_{t+1}(j) \right\}.$$

In the task, states are defined by the number of times each option has been chosen, $c_i$, and the number of times it was rewarded when chosen, $r_i$. From these values for each option the immediate reward value is given by, $r(s_t, a = i) = \frac{r_i + 1}{c_i + 2}$. The extra 1 in the numerator and 2 in the denominator comes from a $\beta$ prior. The future expected value (FEV) is calculated from the expected value of the next state. In this task, when an option is chosen, its value is incremented, $c_i \leftarrow c_i + 1$ and the reward is correspondingly incremented if it is delivered $r_i \leftarrow r_i + 1$, or not if it is not delivered $r_i \leftarrow r_i$. In this way choices and rewards drive state transitions. The model was fit using value iteration (Puterman, 1994). For additional details, see Averbeck (2015).

We use three quantities, *IEV*, *FEV*, and *Bonus* to characterize the values of choice options. The immediate expected value (IEV) used in the paper is given by the immediate reward value for action $a$, $IEV_a = r(s_t, a)$ which is the first term in the utility equation. The IEV is an estimate of the probability that choosing an option will lead to a reward. The future expected value, FEV is given by the second term in the utility equation, $FEV_a = \gamma \Sigma_{j \in s} p(j \mid s_t, a) u_{t+1}(j)$. The FEV is the discounted sum of the future expected rewards, when one chooses optimally, where the discount is given by $\gamma$. The FEV is higher if options with higher IEVs are available, because the animal can earn more rewards by picking these in the future. The FEV of each option also depends on how often the monkey has sampled the available options. On each trial, the difference in the FEV of each option relative to the average FEV of the three options quantifies the relative gain or loss in future rewards resulting from exploration. We refer to this quantity as the exploration Bonus.

$$Bonus_{a=i} = FEV_{a=i} - \langle FEV_j \rangle_j.$$

For example, each time a novel option is chosen and an outcome is observed the Bonus associated with that option decreases, whereas the Bonus values of the alternative options are already low, or even negative, because they have (usually) already been sampled. Behaviorally, the monkeys often explored novel options. Therefore, novel options were often chosen and the Bonus associated with novelty decreased over trials. At the same time the Bonus value for the alternative options became less negative because the monkeys frequently explored the novel options rather than exploiting the familiar options, in the first few trials. This dissociates sampling and habituation and therefore exploration value and simple perceptual novelty. An option must be chosen and not simply seen for its exploration Bonus to decrease. To predict choices the IEV and Bonus values from the POMDP were passed through a softmax function to generate choice probabilities, $p(choice = i) = \exp(b_1 iev_i + b_2 Bonus_i) / \Sigma_{j=1:3} \exp(b_1 iev_j + b_2 Bonus_j)$. The parameters $b_1$ and $b_2$ were optimized to predict choices using fminsearch in MATLAB, as done previously (Costa et al., 2019).

We quantified choice behavior by computing the fraction of times the monkey chose either the novel choice option, best alternative option, or worst alternative option. The best alternative option was defined as the option with the highest IEV, not including the most recently given option. In cases in which the remaining alternative options had equivalent IEVs, the best alternative was defined as the option with the higher action value as estimated by the POMDP model. We also computed the fraction of times the monkeys chose novel options based on the a priori assigned reward probabilities of those options.

For neural data analysis, all trials on which monkeys chose one of the three stimuli were analyzed. Trials in which the monkey broke fixation, failed to make a choice, or attempted to saccade to more than one option were excluded (<1% of all trials). On valid trials, the firing rate of each cell was computed in 200 ms bins, advanced in 50 ms increments, time locked to the monkey's initiation of a saccade to the chosen option, or in some cases delivery of the juice reward. We fit a fixed-effects ANOVA model to these windowed spike counts for each individual cell. The ANOVA included factors for IEV, Bonus, FEV, chosen stimulus, reward outcome of current trial, reward outcome of previous trial, saccade direction to select option, and trials since the novel option were introduced to control for perceptual familiarity. Two ANOVA models were fit. In the first, IEV was modeled as two factors. The a priori reward rate of the option (i.e., 0.8, 0.5 or 0.2, modeled as a fixed effect) and the difference between the current estimate of reward and this IEV value, which modeled the learning of value. In this model chosen stimulus was nested under a priori reward rate. This allowed us to analyze the separate effects of reward value and stimulus on neural activity. All IEV significant effects plotted include significance of either factor that was used to model IEV.

In the second model, IEV was used as a continuous factor, and interacted with stimulus identity. In both models, FEV, Bonus, and trial were continuous factors. All other factors were modeled with factor levels.

## Results

### Task and recordings

We recorded neural activity from three monkeys while they performed a three-armed bandit task (Fig. 1). In each trial, the mon-

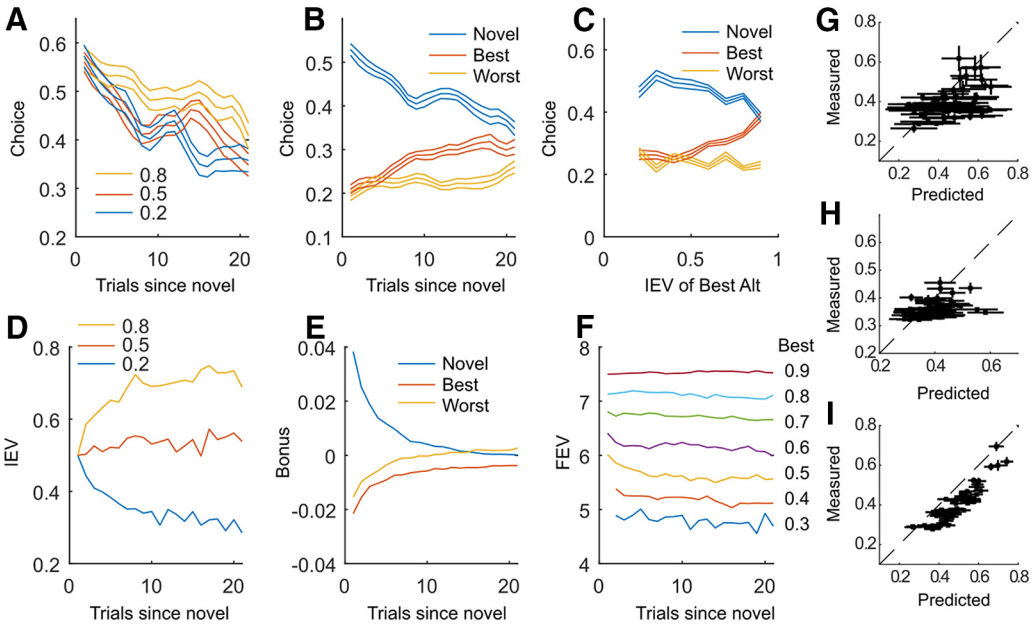


**Figure 2.** Behavior and model. ***A***, Fraction of times the monkeys chose the novel option as a function of the number of trials since the novel option was introduced. Data plotted separately for novel options assigned reward probabilities of 0.2, 0.5, and 0.8. ***B***, Fraction of times the monkeys chose the novel option as a function of trials since novel, relative to the best alternative and worst alternative familiar options. Best and worst were defined as the option with the best IEV and worst IEV, not including the novel option. ***C***, Fraction of times the novel option was selected, relative to the best and worst familiar options, as a function of the IEV of the best alternative option. ***D***, Average IEV of novel options estimated by POMDP model, as a function of trials since the introduction of a novel option. Data are plotted separately for options with IEV of 0.8, 0.5, and 0.2. Note that these are model value estimates, and not behavioral choices. The IEV should asymptotically approach the true value of the choice, which is 0.8, 0.5, or 0.2. ***E***, Average Bonus of novel option, relative to best and worst familiar option, as a function of trials since introduction of novel option. ***F***, Average FEV of the chosen option, as a function of the IEV of the best option currently available. ***G–I***, Predicted versus measured choice probabilities for individual Monkeys H, F, and N, respectively. These values are for all choices, not just for the novel options. Correlation between predicted and measured for ***G***: $r = 0.48 \pm 0.079$ ($N = 9$ sessions); ***H***: $r = 0.29 \pm 0.052$ ($N = 23$ sessions); ***I***: $r = 0.60 \pm 0.033$ ($N = 42$ sessions).

keys fixated a central location (Fig. 1*A*). After a hold period, three stimuli were presented around fixation. The monkeys made a saccade to one of the stimuli to indicate their choice. Reward was then stochastically delivered, depending on an a priori reward probability assigned to the stimulus. The task was run in 650 trial blocks (Fig. 1*B*). At the beginning of each block, we introduced three novel stimuli the monkey had not seen before. The initial stimuli were assigned reward probabilities of 0.2, 0.5, and 0.8. The monkey had to learn by exploring each of the stimuli, which was the best. Every 10–30 trials during the block, one of the stimuli was randomly selected and replaced by a novel stimulus with a randomly assigned reward probability of 0.2, 0.5, or 0.8. The only constraint was that all three stimuli could not have the same reward probability. There were 32 novel stimuli introduced in each block. When a novel stimulus was introduced, the monkeys had to decide whether to explore the novel option or continue to exploit one of the familiar options. While the monkeys performed the task, neural activity was recorded in OFC (Fig. 1*C*).

### Analysis of behavioral data

The task was challenging, because there were three options, and options were frequently replaced. However, the monkeys were able to learn to differentially choose the options depending on their a priori defined reward values (Fig. 2*A*; Value: $F_{(2,5)} = 9.0$, $p = 0.024$). They also showed a novelty preference (Fig. 2*B*; Novel/Best/Worst: $F_{(2,3)} = 218.1$, $p < 0.001$) and this evolved across trials (Trial × Novel/Best/Worst: $F_{(40,4319)} = 3.2$, $p < 0.001$). Although they had a preference for the novel option, their choices did depend on the value of the best alternative (Fig. 2*C*; $F_{(6,774)} = 4.7$, $p = 0.007$), such that they chose the best alternative more when it had higher value.

We used a model-based reinforcement learning algorithm (POMDP; see Materials and Methods) to quantify the behavior. The algorithm generates three values for each chosen option. The IEV (Fig. 2*D*; note these are the IEV of the chosen option, these values do not reflect the monkey's choices), which is the estimated reward probability associated with the object (i.e., the number of times the option has been rewarded divided by the number of times it has been chosen). An exploration Bonus associated with that option (Fig. 2*E*, Bonus), which is the relative additional value that can be obtained in the future, if an option is explored in the current trial (i.e., the FEV of the option relative to the other 2 options). The Bonus comes about because the model assumes that any option that has not been explored might have a high reward probability assigned to it. The distribution over reward probabilities of novel options, before they have been chosen, is given by a broad prior, which allows a probability for reward rates >0.8. It is also possible that an unexplored option has a reward rate <0.2 because the prior is symmetric. But a low reward rate option would not be selected by an optimal agent once its reward rate was known. Therefore, if an option is explored and found to be better than other available options, it can be selected until it is replaced. However, if an option is explored and found to be worse than the other available options, one can switch back to selecting the best alternative. Therefore, an unexplored option might be better than any other options currently available. If it is it will be selected. And if it is not better than the other options it will not be selected. This tradeoff is reflected in the exploration Bonus. After an option has been explored, a better estimate of its true reward probability can be formed, and this leads to a decrease in the exploration Bonus.

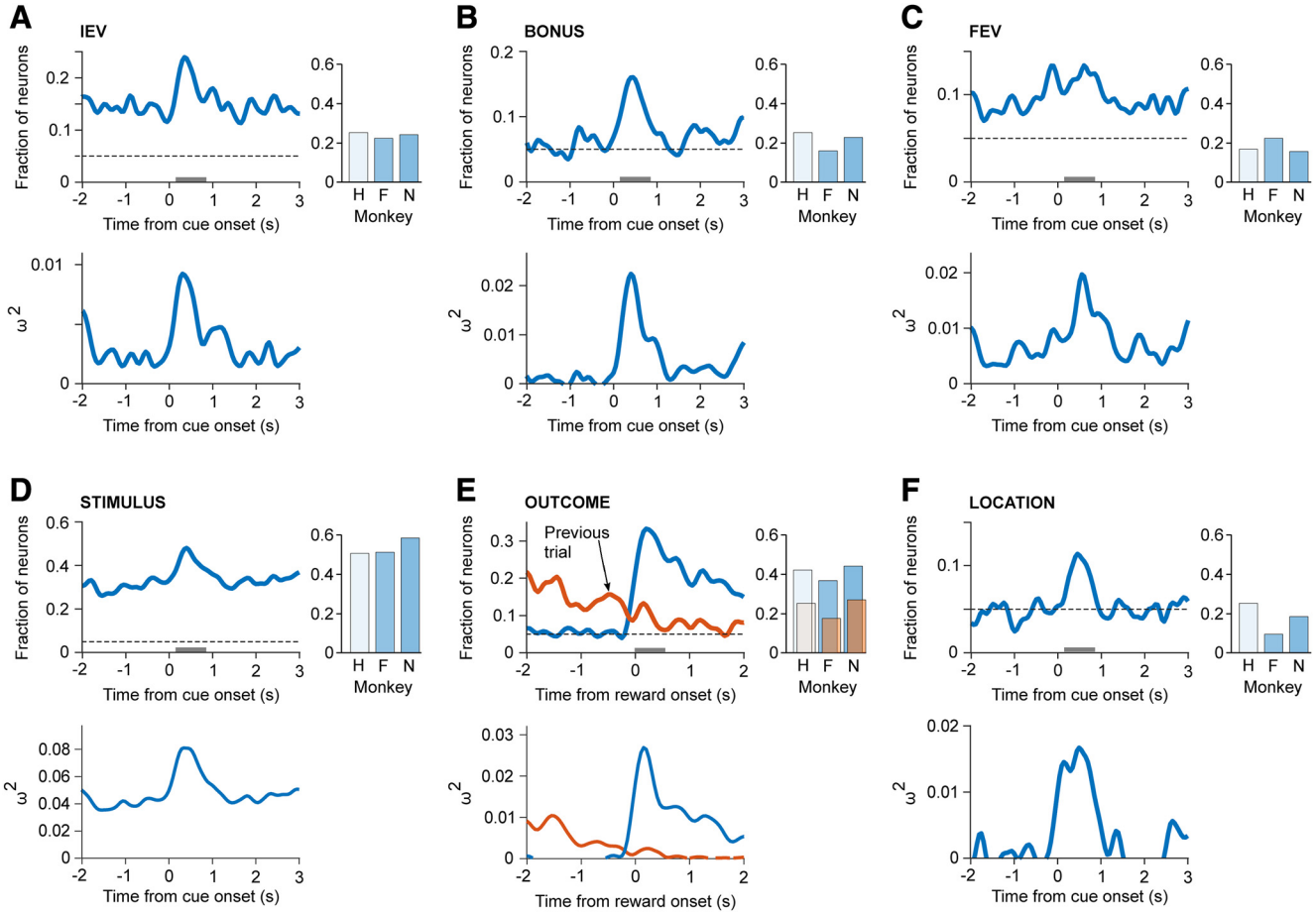


**Figure 3.** Task relevant neural encoding. The fraction of the population of neurons that significantly ($p < 0.05$) encoded task relevant variables (top) and their associated effect size (bottom, $\omega^2$; Olejnik and Algina, 2000). A sliding-window ANOVA was performed on spikes counted in 200 ms bins, advanced in 50 ms increments. Effect size estimates were calculated for all neurons showing a significant effect in at least 5 windows from 0 to 1200 ms after cue or reward onset. Bar plots indicate the fraction of significant neurons encoding each task factor for each of the three monkeys. The shading of each bar reflects the number of neurons recorded in each monkey (Monkey H = 12; Monkey F = 63; $N = 71$ neurons). ***A***, IEV of the chosen option. ***B***, Exploration bonus associated with chosen option. ***C***, FEV of the chosen option. ***D***, Stimulus-specific identity of the chosen option. ***E***, Outcome of the current (blue) and previous choice (orange) trial. ***F***, Screen location of the chosen option.

And, finally, the model also estimates the FEV of an option, which is the sum of the future discounted expected rewards associated with choosing an option in the current trial, and then making optimal decisions in all future trials. The FEV primarily reflects the IEV of the best option in the current set (Fig. 2*F*), because one expects to obtain more rewards over the relevant future time horizon if there is a good option (i.e., high IEV) available, relative to when there is not a good option available (i.e., the best option has a low IEV). The algorithm was able to predict the choices of the individual monkeys, reasonably well (Fig. 2*G–I*).

### Analysis of neural data

Next, we performed a moving-window ANOVA to assess the fraction of neurons that encoded the task-relevant variables. The ANOVA was performed on spike counts in 200 ms bins, with 50 ms steps between bins. We found that all task relevant variables were encoded in the population (Fig. 3*A–F*). We also examined effect size and found that it was similar to the fraction of neurons encoding each variable (Fig. 3, bottom), and comparable to results in the ventral striatum and amygdala (Costa et al., 2019). Approximately 15% of neurons encoded the IEV of the upcoming choice during the ITI and baseline hold periods (Fig. 3*A*). This increased to 24% of the neurons at the time of choice. The exploration bonus was encoded at chance levels during the ITI and hold periods (Fig. 3*B*) and this increased to 17% of the neurons at the time of choice. The FEV, which characterizes the overall-reward environment, was represented by ~10% of the neurons during the ITI and baseline hold periods (Fig. 3*C*). The encoding of FEV then increased to ~13% of the neurons at the time of choice. FEV was therefore encoded, but not strongly.

We also found that the identity of the chosen stimulus was encoded during the ITI and baseline hold period (Fig. 3*D*). Stimulus encoding was the strongest factor represented in our population, consistent with previous results (Wallis and Miller, 2003). Approximately 30% of the neurons encoded the identity of the chosen stimulus during the ITI and baseline hold periods, and this increased to ~50% of the neurons at the time of choice. This could be seen clearly in single example neurons, which showed considerable variability in response to different high value stimuli (Fig. 4).

The current and previous trial reward outcomes were strongly encoded in the population (Fig. 3*E*). The current trial reward (i.e., a comparison of firing rate in rewarded and non-rewarded trials) reached ~35% after reward outcome. The fraction of neurons representing the reward remained elevated through the ITI and baseline hold periods, with ~10–15% of the neurons representing the previous trial outcome at the time of the current trial outcome. Whether

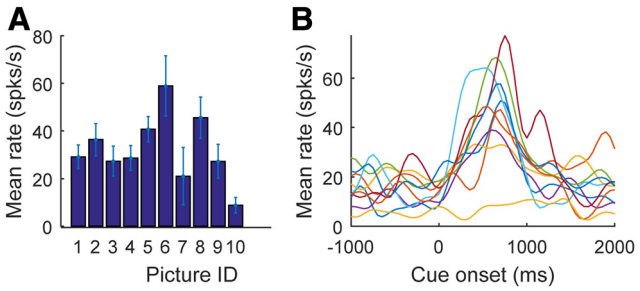


**Figure 4.** Single neuron example of variability in stimulus encoding. ***A***, Average response to chosen stimulus in a window from 0 to 400 ms after cue onset. ***B***, Spike density functions showing average response to the 10 different high-valued stimuli. Each line represents the mean for a different stimulus.

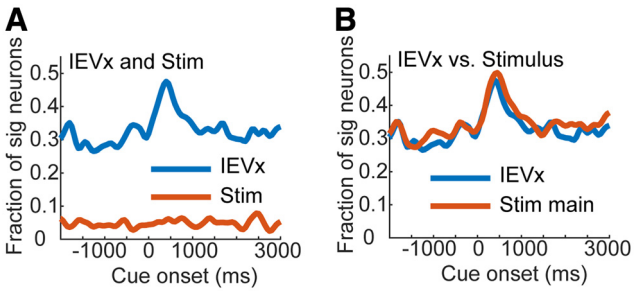


**Figure 5.** Alternative encoding model. All variables assessed at $p < 0.05$. ***A***, Encoding of IEV as an interaction (IEVx) with chosen stimulus identity and stimulus as an non-nested main effect (Stim). ***B***, Overlay of non-nested interaction from second ANOVA model and nested stimulus identity encoding from first ANOVA model.

this is a neural effect, or more simply reflects binding of molecules to peripheral taste receptors on the tongue is not clear, as it likely takes the monkeys some time to swallow the juice.

Finally, we also examined encoding of the screen location of the chosen option (Fig. 3*F*). Because the locations of the objects were randomized from trial-to-trial, encoding was at chance during the ITI and baseline hold. However, this increased to 11% of the neurons at the time of choice.

In summary, both the value and identity of the to-be-chosen option were represented during the ITI and baseline hold periods, reflecting the fact that monkeys learned the identities of the good options and their associated values. There was a substantial encoding of the reward outcome, and this continued through the ITI and baseline hold, such that the previous trial outcome was represented at the time of the current trial choice and outcome. Finally, the strongest factor encoded by the neural population was the perceptual identity of the chosen option. The IEV, FEV, and Bonus values were encoded at intermediate levels, and the location of the chosen option was encoded by only a small fraction of the population.

Next, we used the results of the ANOVA model to characterize several features of value encoding. First, we characterized the tuning properties of the cells that encoded value. We found that 18 of the 27 neurons that encoded value did so with monotonic (increasing or decreasing) tuning functions. This fraction differed from chance (binomial test, $p = 0.026$). There was an approximately equal representation of positive (8/18) and negative (10/18) tuning for value (binomial test, $p = 0.593$). We also asked whether single neurons explicitly encoded reward prediction errors (RPEs) because RPEs are important for learning in basic Rescorla–Wagner or Bayesian models. We did this be examining the fraction of neurons that encoded reward and expected reward with opposite signs at the time of reward delivery. However, we found that there was no enrichment for specific RPE encoding (11/18; binomial test, $p = 0.119$). We also did not find that single neurons tended to encode both value and reward. The presence of each of these in the single population was equivalent to the product of finding them individually in the population (Fisher test, $p = 0.431$).

When we examined encoding of stimulus and IEV in the first ANOVA model, we did so by nesting stimulus under the a priori value of the option (see Materials and Methods). This model shows us that the response to a high value stimulus depends on the stimulus identity (Fig. 4). However, this model does not allow us to test whether the response to IEV is stimulus dependent. To examine this in more detail, we used a second ANOVA model to examine neural activity. In this model, we modeled the interaction of IEV, coded as a single continuous variable, and stimulus identity. Stimulus identity was also entered as a main effect. This model has the same degrees of freedom as the original model, so we can directly compare prediction accuracy between the models. We found that the models were similar in their ability to predict neural activity. The average difference in $R^2$ was 0.1% across single neurons. This is not surprising as nesting in an ANOVA model is an interaction (Kutner et al., 2005), although it is different from the interaction in the second model. However, when we compared the residual variance, we found that the first model better predicted activity in 56% of the single neurons, and across the population the first model significantly outperformed the second model ($t$ test, $t_{(145)} = 2.64$, $p = 0.009$, mean difference in residual variance = 1.4, SEM = 0.54). Therefore, there was a slight preference for the nested model.

When we examined cue encoding using the second model, we found that the interaction between value and stimulus identity accounted for the stimulus encoding (Fig. 5*A*). Stimulus encoding only occurred at chance levels when the interaction of IEV and stimulus identity was also in the model. However, when we compared the encoding of this interaction using the second model, with nested stimulus encoding using the first model (Fig. 5*B*), we found that they were nearly identical. Therefore, using either approach we find that value coding is heavily dependent on the stimulus to which the value is associated, in single OFC neurons.

## Discussion

We performed an experiment in which we recorded neural activity from OFC while monkeys took part in a three-armed bandit reinforcement learning task. In the task, we periodically introduced new options, which allowed us to examine the explore–exploit tradeoff. This allowed us to examine encoding of the immediate and future expected values of options, as well as the exploration bonus. We found that OFC encoded task variables important for learning in the task, similar to what has been seen in the amygdala and ventral striatum (Costa et al., 2019). The strength of the encoding, however, differed across variables. The chosen stimulus had the largest effect on the neural population, followed by the reward outcome. The expected values of the chosen options were coded at intermediate levels, and the location and future expected values were encoded at low levels. We also found that chosen value encoding was stimulus specific. Therefore, responses depended not just on value, but also the visual stimulus associated with the value, similar to what we found in the amygdala and ventral striatum (Costa et al., 2019).

We found substantial stimulus identity encoding, consistent with a previous study (Wallis and Miller, 2003). This stimulus encoding was independent of the value of the option, because our ANOVA model examined differences in response for stimuli within each value category (i.e., 0.8, 0.5, and 0.2 reward probability). For example, the response to different cues that all coded a reward probability of 0.8 differed. The ANOVA model also controlled for additional confounding factors that could account for these differences, including Bonus value, the number of times the stimulus was seen, etc. This is perhaps not surprising, given how important it is for the monkeys to encode the identity of the chosen options. In the task, we introduced a new option every 10–30 trials and the monkeys had to quickly learn the values of the chosen options to maximize reward. Therefore, identifying the best image and picking it consistently across trials maximized rewards. The learning effects were also consistent with the elevated coding of the to-be-chosen stimulus during the ITI and baseline hold periods. The monkeys likely were remembering, across trials, the identities of the preferred options. These visual stimulus coding effects are also consistent with the organization of the anatomical inputs to OFC. There are strong visual and multisensory inputs (Morecraft et al., 1992; Carmichael and Price, 1995) as well as inputs from other visual recipient areas including the amygdala (Ghashghaei and Barbas, 2002).

We also found strong encoding of reward outcomes on the current and previous trial, also consistent with OFC recordings in other reward learning tasks (Simmons and Richmond, 2008). These results are not consistent, however, with previous reports using functional imaging, which did not find a representation of previous trial outcomes in OFC (Chau et al., 2015). However, this may be because of a lack of sensitivity in the fMRI experiments or a difference in the OFC location investigated. We find encoding of previous trial outcomes across the OFC, amygdala, and ventral striatum network (Costa et al., 2019). Whether the previous trial reward coding is generated centrally or reflects ongoing binding to peripheral taste receptors is not clear. However, it is clear that the signal for the reward in the current and previous trial is well encoded in OFC.

Previous studies of learning related signals in OFC have found that neurons reflect the ongoing chosen values of options (Wallis and Miller, 2003; Rudebeck et al., 2013a, 2017b). These learning related value representations are consistent with findings in studies in which choice-outcome associations are highly overlearned (Padoa-Schioppa and Assad, 2006; Kennerley and Wallis, 2009; Kennerley et al., 2009; Rudebeck et al., 2013a; Rudebeck and Murray, 2014). Therefore, chosen value signals in OFC are not specific to learning tasks. Studies have found that this encoding depends partially on amygdala inputs, as lesions of the amygdala substantially reduce value encoding in OFC in learning tasks (Schoenbaum et al., 2003a; Rudebeck et al., 2013a, 2017b). This is consistent with findings that the amygdala and the ventral striatum code similar value signals in our task (Costa et al., 2019). Other studies have found similar value coding, but have found different time courses for positive versus negative value updates (Morrison et al., 2011).

The amygdala, ventral striatum, medial dorsal thalamus, and caudal OFC are part of a monosynaptically connected network (Amaral et al., 1992; Neftci and Averbeck, 2019), which plays an important role in RL (Averbeck and Costa, 2017). Comparison of the neurophysiology data from this study and our previous study in which we reported results from the amygdala and ventral striatum (Costa et al., 2019) suggests that these structures code similar variables, and at similar levels. All the task relevant variables were encoded across structures, although at different levels. One salient feature is that the ventral striatum has almost no saccade direction coding (Costa et al., 2019), whereas both OFC and the amygdala have low levels of saccade direction encoding. Other work on closely related tasks suggests that dorsolateral prefrontal cortex also codes a range of value-related variables (Bartolo et al., 2019) with strongly enhanced encoding of saccade direction. Given the strong interconnectivity among these areas, it is not surprising that they code the same variables, as this is often found across monosynaptically connected structures (Chafee and Goldman-Rakic, 1998; Averbeck et al., 2009; Seo et al., 2012). Each area in this network does differ, however, in its connections outside the network. OFC has strong connections with other prefrontal areas (Barbas, 1993; Carmichael and Price, 1996). The amygdala receives strong visual inputs and projects back to visual areas (Amaral et al., 1992), and the ventral striatum receives perhaps the strongest dopamine input of all areas (Haber et al., 2000), and does not project directly back to cortex, but instead projects through the GPi/SNr, and then on to the thalamus (Alexander et al., 1986). These areas also have different microcircuit organization and local neurochemistry, which may reflect differences in computational and plasticity mechanisms. One hypothesis is that the primary role of cortex is setting up complex representations (Rigotti et al., 2013), over which the amygdala and ventral striatum then learn (Averbeck and Costa, 2017) with the amygdala updating values quickly using an activity-dependent mechanism, and the striatum updating values more slowly using a dopamine-dependent mechanism (Averbeck, 2017). Neurophysiology data may not be able to distinguish representations used for learning, from learning itself, and therefore cortical and subcortical structures may appear similar. Approaches which block synaptic plasticity or protein synthesis may be able to test these hypotheses.

In reinforcement learning models, states define all information relevant to making choices, or equivalently to predicting the outcomes of choices. It has been suggested that OFC plays an important role in state representation (Wilson et al., 2014a; Schuck et al., 2016). We have used a POMDP algorithm to model behavior in our task (Averbeck, 2015). This model allows us to calculate a principled exploration bonus value that we can use to examine neural activity. This is because in true model-based RL, the relative immediate and future values of novel options can be calculated directly. This differs from some models of the explore–exploit dilemma, where the values of novel options are floated as free parameters (Wittmann et al., 2008; Djamshidian et al., 2011; Costa et al., 2014). However, model-based RL requires an accurate representation of state, because values are tied directly to states. In the POMDP algorithm, states are not directly observable and have to be inferred. State inference and state transitions in our task, however, are governed by chosen objects and whether they are rewarded. These are the two strongest signals in the OFC data. Therefore, OFC does represent the information necessary to derive choice values, in our POMDP model-based framework, as well as the values themselves. Other recent work has also shown that medial prefrontal cortex in rodents may play a role in hidden state inference, when state is defined by the passage of time (Starkweather et al., 2018). Therefore, OFC and additional prefrontal areas appear to play a role in state representation or state inference.

There are numerous theories of OFC function, beyond state representation (Schoenbaum et al., 2009; Padoa-Schioppa and Cai, 2011; Rudebeck and Murray, 2014; Stalnaker et al., 2015). One of the early theories suggested that OFC was important for

reversal learning, or inhibitory control (Butter, 1969; Iversen and Mishkin, 1970). However, recent work has shown that these deficits were due to destruction of fibers of passage, and not cell bodies in OFC (Rudebeck et al., 2013b), at least in monkeys. One consistent finding, is that OFC plays an important role in devaluation tasks (Rudebeck et al., 2013b, 2017a; Murray and Rudebeck, 2018). In these tasks, monkeys first learn to make choices to obtain a preferred outcome (e.g., peanuts vs candy). One of the outcomes is then devalued by feeding it to satiation. After the outcome is devalued, the monkeys normally switch to preferring the non-devalued outcome. Amygdala lesions in moneys also lead to deficits in variations on these tasks (Rhodes and Murray, 2013), and these behaviors are mediated by circuits connecting medial-prefrontal cortex to the dorsal-medial-striatum in rodents (Hart et al., 2018). Thus, OFC plays an important role in reward devaluation in monkeys, but as part of a broader circuit, similar to what we have observed here with respect to its role in exploratory decision making. Reward devaluation and explore-exploit decision making are both instances of model-based RL but how they are related computationally is not obvious. Further development of theories of OFC function should attempt to account for both functions.

### Conclusions

We found substantial representations of chosen stimulus identity and reward in OFC, consistent with previous studies (Wallis and Miller, 2003; Simmons and Richmond, 2008). Chosen value, which is often considered an important function of OFC, was represented but not as frequently. Other factors including the location of the chosen target, and the future expected value of options were encoded, but only rarely. These signals reflect the information necessary to learn to select the best options and balance the explore–exploit tradeoff in our task (Averbeck, 2015; Costa et al., 2019). We have found similar signals across the network of monosynaptically connected areas that includes the amygdala and ventral striatum (Costa et al., 2019). However, the extent to which these signals are computed locally in OFC, or inherited from other structures is not currently clear. Future work that combines multisite recording with inactivation of OFC circuitry will help disentangle the specific role of each area in these tasks.

## References


Alexander GE, DeLong MR, Strick PL (1986) Parallel organization of functionally segregated circuits linking basal ganglia and cortex. Annu Rev Neurosci 9:357–381.

Amaral DG, Price JL, Pitkanen A, Carmichael ST (1992) Anatomical organization of the primate amygdaloid complex. In: The amygdala: neurobiological aspects of emotion, memory, and mental dysfunction (Aggleton JP, ed), pp. 1–66. New York: Wiley.

Asaad WF, Eskandar EN (2008) A flexible software tool for temporally-precise behavioral control in Matlab. J Neurosci Methods 174:245–258.

Averbeck BB (2015) Theory of choice in bandit, information sampling and foraging tasks. PLoS Comput Biol 11:e1004164.

Averbeck BB (2017) Amygdala and ventral striatum population codes implement multiple learning rates for reinforcement learning. 2017 IEEE Symposium Series on Computational Intelligence (SSCI), Honolulu, HI, 2017, pp. 1–5.

Averbeck BB, Costa VD (2017) Motivational neural circuits underlying reinforcement learning. Nat Neurosci 20:505–512.

Averbeck BB, Crowe DA, Chafee MV, Georgopoulos AP (2009) Differential contribution of superior parietal and dorsal-lateral prefrontal cortices in copying. Cortex 45:432–441.

Averbeck BB, Djamshidian A, O'Sullivan SS, Housden CR, Roiser JP, Lees AJ (2013) Uncertainty about mapping future actions into rewards may underlie performance on multiple measures of impulsivity in behavioral addiction: evidence from Parkinson's disease. Behav Neurosci 127:245–255.

Barbas H (1993) Organization of cortical afferent input to orbitofrontal areas in the rhesus monkey. Neuroscience 56:841–864.

Bartolo R, Saunders RC, Mitz A, Averbeck BB (2019) Dimensionality, information and learning in prefrontal cortex. bioRxiv 823377.

Butter CM (1969) Perseveration in exinction and in discrimination reversal tasks following selective frontal ablations in *Macaca mulatta*. Physiol Behav 4:163–171.

Carmichael ST, Price JL (1995) Sensory and premotor connections of the orbital and medial prefrontal cortex of macaque monkeys. J Comp Neurol 363:642–664.

Carmichael ST, Price JL (1996) Connectional networks within the orbital and medial prefrontal cortex of macaque monkeys. J Comp Neurol 371:179–207.

Chafee MV, Goldman-Rakic PS (1998) Matching patterns of activity in primate prefrontal area 8a and parietal area 7ip neurons during a spatial working memory task. J Neurophysiol 79:2919–2940.

Chau BK, Sallet J, Papageorgiou GK, Noonan MP, Bell AH, Walton ME, Rushworth MF (2015) Contrasting roles for orbitofrontal cortex and amygdala in credit assignment and learning in macaques. Neuron 87:1106–1118.

Costa VD, Tran VL, Turchi J, Averbeck BB (2014) Dopamine modulates novelty seeking behavior during decision making. Behav Neurosci 128:556–566.

Costa VD, Mitz AR, Averbeck BB (2019) Subcortical substrates of explore–exploit decisions in primates. Neuron 103:533–545.e5.

Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. Nature 441:876–879.

Dias R, Robbins TW, Roberts AC (1996) Dissociation in prefrontal cortex of affective and attentional shifts. Nature 380:69–72.

Djamshidian A, O'Sullivan SS, Wittmann BC, Lees AJ, Averbeck BB (2011) Novelty seeking behaviour in Parkinson's disease. Neuropsychologia 49:2483–2488.

Ghashghaei HT, Barbas H (2002) Pathways for emotion: interactions of prefrontal and anterior temporal pathways in the amygdala of the rhesus monkey. Neuroscience 115:1261–1279.

Groman SM, Keistler C, Keip AJ, Hammarlund E, DiLeone RJ, Pittenger C, Lee D, Taylor JR (2019) Orbitofrontal circuits control multiple reinforcement-learning processes. Neuron 103:734–746.e3.

Haber SN, Fudge JL, McFarland NR (2000) Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. J Neurosci 20:2369–2382.

Hart G, Bradfield LA, Fok SY, Chieng B, Balleine BW (2018) The bilateral prefronto-striatal pathway is necessary for learning new goal-directed actions. Curr Biol 28:2218–2229.e7.

Iversen SD, Mishkin M (1970) Perseverative interference in monkeys following selective lesions of the inferior prefrontal convexity. Exp Brain Res 11:376–386.

Kennerley SW, Wallis JD (2009) Evaluating choices by single neurons in the frontal lobe: outcome value encoded across multiple decision variables. Eur J Neurosci 29:2061–2073.

Kennerley SW, Dahmubed AF, Lara AH, Wallis JD (2009) Neurons in the frontal lobe encode the value of multiple decision variables. J Cogn Neurosci 21:1162–1178.

Kutner MH, Nachtsheim CJ, Neter J, Li W (2005) Applied linear statistical models. New York: McGraw-Hill.

Mitz AR (2005) A liquid-delivery device that provides precise reward control for neurophysiological and behavioral experiments. J Neurosci Methods 148:19–25.

Morecraft RJ, Geula C, Mesulam MM (1992) Cytoarchitecture and neural afferents of orbitofrontal cortex in the brain of the monkey. J Comp Neurol 323:341–358.

Morrison SE, Saez A, Lau B, Salzman CD (2011) Different time courses for learning-related changes in amygdala and orbitofrontal cortex. Neuron 71:1127–1140.

Murray EA, Rudebeck PH (2018) Specializations for reward-guided decision-making in the primate ventral prefrontal cortex. Nat Rev Neurosci 19:404–417.

Neftci EO, Averbeck BB (2019) Reinforcement learning in artificial and biological systems. Nat Mach Intell 1:133–143.

Olejnik S, Algina J (2000) Measures of effect size for comparative studies:

applications, interpretations, and limitations. Contemp Educ Psychol 25:241–286.
Padoa-Schioppa C, Assad JA (2006) Neurons in the orbitofrontal cortex encode economic value. Nature 441:223–226.
Padoa-Schioppa C, Cai X (2011) The orbitofrontal cortex and the computation of subjective value: consolidated concepts and new perspectives. Ann N Y Acad Sci 1239:130–137.
Puterman ML (1994) Markov decision processes: discrete stochastic dynamic programming. New York: Wiley.
Rhodes SE, Murray EA (2013) Differential effects of amygdala, orbital prefrontal cortex, and prelimbic cortex lesions on goal-directed behavior in rhesus macaques. J Neurosci 33:3380–3389.
Rigotti M, Barak O, Warden MR, Wang XJ, Daw ND, Miller EK, Fusi S (2013) The importance of mixed selectivity in complex cognitive tasks. Nature 497:585–590.
Rudebeck PH, Murray EA (2014) The orbitofrontal oracle: cortical mechanisms for the prediction and evaluation of specific behavioral outcomes. Neuron 84:1143–1156.
Rudebeck PH, Mitz AR, Chacko RV, Murray EA (2013a) Effects of amygdala lesions on reward-value coding in orbital and medial prefrontal cortex. Neuron 80:1519–1531.
Rudebeck PH, Saunders RC, Prescott AT, Chau LS, Murray EA (2013b) Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. Nat Neurosci 16:1140–1145.
Rudebeck PH, Saunders RC, Lundgren DA, Murray EA (2017a) Specialized representations of value in the orbital and ventrolateral prefrontal cortex: desirability versus availability of outcomes. Neuron 95:1208–1220.e5.
Rudebeck PH, Ripple JA, Mitz AR, Averbeck BB, Murray EA (2017b) Amygdala contributions to stimulus–reward encoding in the macaque medial and orbital frontal cortex during learning. J Neurosci 37:2186–2202.
Schoenbaum G, Setlow B, Saddoris MP, Gallagher M (2003a) Encoding predicted outcome and acquired value in orbitofrontal cortex during cue sampling depends upon input from basolateral amygdala. Neuron 39: 855–867.
Schoenbaum G, Setlow B, Nugent SL, Saddoris MP, Gallagher M (2003b) Lesions of orbitofrontal cortex and basolateral amygdala complex disrupt acquisition of odor-guided discriminations and reversals. Learn Mem 10:129–140.
Schoenbaum G, Roesch MR, Stalnaker TA, Takahashi YK (2009) A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. Nat Rev Neurosci 10:885–892.
Schuck NW, Cai MB, Wilson RC, Niv Y (2016) Human orbitofrontal cortex represents a cognitive map of state space. Neuron 91:1402–1412.
Seo M, Lee E, Averbeck BB (2012) Action selection and action value in frontal-striatal circuits. Neuron 74:947–960.
Simmons JM, Richmond BJ (2008) Dynamic changes in representations of preceding and upcoming reward in monkey orbitofrontal cortex. Cereb Cortex 18:93–103.
Stalnaker TA, Cooch NK, Schoenbaum G (2015) What the orbitofrontal cortex does not do. Nat Neurosci 18:620–627.
Starkweather CK, Gershman SJ, Uchida N (2018) The medial prefrontal cortex shapes dopamine reward prediction errors under state uncertainty. Neuron 98:616–629.e6.
Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. Cambridge, MA: MIT.
Wallis JD, Miller EK (2003) Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. Eur J Neurosci 18:2069–2081.
Willenbockel V, Sadr J, Fiset D, Horne GO, Gosselin F, Tanaka JW (2010) Controlling low-level image properties: the SHINE toolbox. Behav Res Methods 42:671–684.
Wilson RC, Takahashi YK, Schoenbaum G, Niv Y (2014a) Orbitofrontal cortex as a cognitive map of task space. Neuron 81:267–279.
Wilson RC, Geana A, White JM, Ludvig EA, Cohen JD (2014b) Humans use directed and random exploration to solve the explore–exploit dilemma. J Exp Psychol Gen 143:2074–2081.
Wittmann BC, Daw ND, Seymour B, Dolan RJ (2008) Striatal activity underlies novelty-based choice in humans. Neuron 58:967–973.
Zajkowski WK, Kossut M, Wilson RC (2017) A causal role for right frontopolar cortex in directed, but not random, exploration. eLife 6:e27430.