Neuron
Viewpoint

Orbitofrontal Cortex as a Cognitive Map of Task Space

Robert C. Wilson,[1,*] Yuji K. Takahashi,[2] Geoffrey Schoenbaum,[2,3,4] and Yael Niv[1,4,*]
[1]Department of Psychology and Neuroscience Institute, Princeton University, Princeton, NJ 08544, USA
[2]Department of Anatomy and Neurobiology, University of Maryland School of Medicine, Baltimore MD 21201, USA
[3]Department of Psychiatry, University of Maryland School of Medicine, Baltimore MD 21201, USA
[4]These authors contributed equally to this work
*Correspondence: rcw2@princeton.edu (R.C.W.), yael@princeton.edu (Y.N.)
http://dx.doi.org/10.1016/j.neuron.2013.11.005

SUMMARY

Orbitofrontal cortex (OFC) has long been known to play an important role in decision making. However, the exact nature of that role has remained elusive. Here, we propose a unifying theory of OFC function. We hypothesize that OFC provides an abstraction of currently available information in the form of a labeling of the current task state, which is used for reinforcement learning (RL) elsewhere in the brain. This function is especially critical when task states include unobservable information, for instance, from working memory. We use this framework to explain classic findings in reversal learning, delayed alternation, extinction, and devaluation as well as more recent findings showing the effect of OFC lesions on the firing of dopaminergic neurons in ventral tegmental area (VTA) in rodents performing an RL task. In addition, we generate a number of testable experimental predictions that can distinguish our theory from other accounts of OFC function.

INTRODUCTION

Many studies have shown that orbitofrontal cortex (OFC) is important for learning and decision making (see reviews by Murray et al., 2007; Wallis, 2007; Padoa-Schioppa, 2011; Rushworth et al., 2011). Despite this progress, the exact role that the OFC plays in decision making is unclear. Even without an OFC, animals and humans can learn, unlearn, and even reverse previous associations, although they do so more slowly than their healthy counterparts. What role can the OFC be playing whose absence would cause such subtle, yet broadly permeating, deficits? We suggest that the OFC represents the animal's current location within an abstract cognitive map of the task (formally, the current state in a state space).

Our hypothesis links OFC function to the formal theory of reinforcement learning (RL). In recent years, RL has successfully accounted for a diverse set of findings from behavioral results in classical conditioning (Rescorla and Wagner, 1972) to the firing patterns of midbrain dopaminergic neurons (Schultz et al., 1997). At the heart of RL models is the concept of a "state representation," an abstract representation of the task that describes its underlying structure, the different states of the task, and the (possibly action-dependent) links between them. RL provides a set of algorithms by which one can learn a value for each state, $V(s)$, that approximates the total discounted future reward that can be expected when the current state is s. These values aid decision making in the service of harvesting reward and avoiding punishments.

In most RL models, it is assumed de facto that the animal magically knows the true state representation of the task. However, it is clear that an integral part of learning a new task is learning to represent it correctly (Gershman and Niv, 2010, 2013; Gershman et al., 2010; Wilson and Niv, 2011). The state representation can be as simple as the two states needed to model a Pavlovian conditioning experiment in which a single stimulus predicts reward (e.g., the states "light on" and "light off") or as intractably huge as the state space of a game of chess. The states can be tied to external stimuli (as in light on/off), or they can include internal information that is not available in the environment and must be retained in memory or inferred, such as one's previous actions or the context of the task (e.g., information about the opponent's style of play in chess). More formally, one way to distinguish between simple and complex tasks relates to whether states are fully or partially observable to the animal given perceptual information. In fully observable decision problems, states correspond to easily detectable features of the environment, making these problems much simpler to solve than partially observable problems, which are notoriously difficult to solve optimally (Kaelbling et al., 1998).

We hypothesize that OFC is critical for representing task states in such partially observable scenarios. We propose that OFC integrates multisensory perceptual input from cortical and subcortical areas, together with information about memories of previous stimuli, choices, and rewards, to determine the current state—an abstract label of a multitude of information akin to the current "location" in a "cognitive map" of the task. Importantly, although state representations most likely exist elsewhere in the brain as well, we hypothesize that the OFC is unique in its ability to disambiguate task states that are perceptually similar but conceptually different, for instance, by using information from working memory. Thus, impaired OFC function does not imply a complete loss of state information but rather that perceptually similar states can no longer be distinguished—an OFC-lesioned

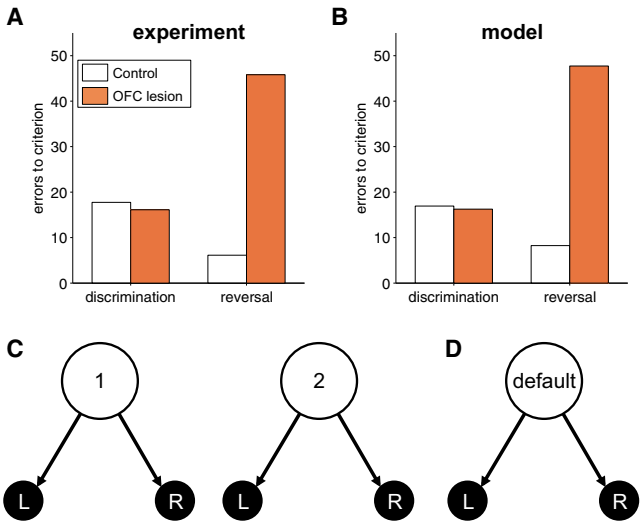


Figure 1. Reversal Learning

(A) Experimental results showing the mean errors to criterion in initial discrimination learning and final reversal for control and OFC-lesioned animals. Adapted from Butter (1969).

(B) Model simulations of the same task.

(C) State representation of the task used to model control animals, in which the state depends on both the action and outcome on the last trial.

(D) Stimulus-bound state representation modeling OFC-lesioned animals.

animal can still learn and perform basic tasks using RL, albeit using only observable (stimulus-bound) states based on current perceptual information. As a result, basic learning and decision making are possible without the OFC, but behavior becomes more and more impaired as tasks become abstract, and more of their states are partially observable.

RESULTS

Here, we show how our theory can account for a number of experimental findings. First, we consider the archetypal "OFC task" of reversal learning as well as delayed alternation, extinction, and devaluation before turning to neural findings that more directly reveal the contribution that the OFC might make to RL.

Reversal Learning

Perhaps the most classic behavioral deficit associated with OFC dysfunction is impaired reversal learning (Teitelbaum, 1964; Butter, 1969; Jones and Mishkin, 1972; Rolls et al., 1994; Dias et al., 1996; Meunier et al., 1997; McAlonan and Brown, 2003; Schoenbaum et al., 2002, 2003a; Chudasama and Robbins, 2003; Bohn et al., 2003; Izquierdo et al., 2004; Kim and Ragozzino, 2005). We illustrate our theory through a simulation of Butter (1969), although we stress that the model similarly accounts for reversal learning deficits in other animals and preparations.

In Butter (1969), monkeys displaced a plaque on either their left or on their right in order to receive food reward. Only one location was rewarded in each block, and its identity was reversed once the monkey reached a criterion of 90% correct. Reward contingencies were reversed five times. Figure 1A summarizes the results—whereas initial learning was spared, OFC-lesioned animals were impaired on reversals relative to sham-lesioned controls.

To model behavior in this task, we used a simple Q-learning algorithm (Sutton and Barto, 1998; Morris et al., 2004) that learns $Q(a, s_t)$, the value of taking action a in state s_t. This q value is updated every time an action is taken and a (possibly zero) reward r_{t+1} is observed according to

$$Q_{new}(a_t, s_t) = Q_{old}(a_t, s_t) + \alpha(r_{t+1} - Q_{old}(a_t, s_t)),$$

where α is a learning rate parameter and $[r_{t+1} - Q_{old}(a_t, s_t)]$ is the prediction error. We omit the value of the subsequent state from the prediction error (cf. Sutton and Barto, 1998) because, in this task, trials involve one state with no sequential contingencies. This renders our learning rule identical to Rescorla and Wagner (1972). Using the learned values, the probability of taking action α in state s_t is given by the softmax or Luce rule

$$p(a|s_t) = \frac{\exp(\beta Q(a, s_t))}{\sum_{a'} \exp(\beta Q(a', s_t))},$$

where β is an inverse-temperature parameter that affects the tradeoff between exploiting and exploring, and the sum in the denominator is over all possible actions. Unless mentioned otherwise, in all simulations we used $\alpha = 0.03$ and $\beta = 3$.

Our model proposes that all animals learned with this same model-free algorithm, but that the crucial difference between sham- and OFC-lesioned animals was in the states, s_t, about which they learned values. In particular, in concordance with the true structure of the task, for sham-lesioned animals, we modeled the task with two different states: state 1, in which choosing "right" yields reward and choosing "left" does not, and state 2, with the opposite reward contingencies (Figure 1C). In each state, the animal must learn values for the right and left actions. After an action is selected, the state transitions according to the chosen action and its outcome, and the next trial begins.

It is easy to see that such a state representation leads to rapid learning of reversals. When the reward is on the right, the model will be in state 1, and because a "right" choice from this state is most likely to be rewarded, the model develops a strong preference for the right action in this state. Similarly, after the reversal, the model transitions to state 2 and learns a strong preference for "left" from this state. Reversing back to the initial contingencies will not necessitate new learning, given that the action propensities learned in state 1 are left unaltered. If rewards and choices are deterministic, then the model will only take one trial to reverse its behavior after such a rereversal. In the face of decision noise, mistakes can occur at a rate determined by β.

The two states in the above model are defined by memory of the action and outcome of the last trial but are perceptually identical. Thus, according to our hypothesis, when the OFC is lesioned, these two states are no longer distinguishable, and the task reduces to one state (Figure 1D). As a result, the reversal of behavior after a reversal of reward contingency requires "unlearning" of the preference that was acquired in the previous

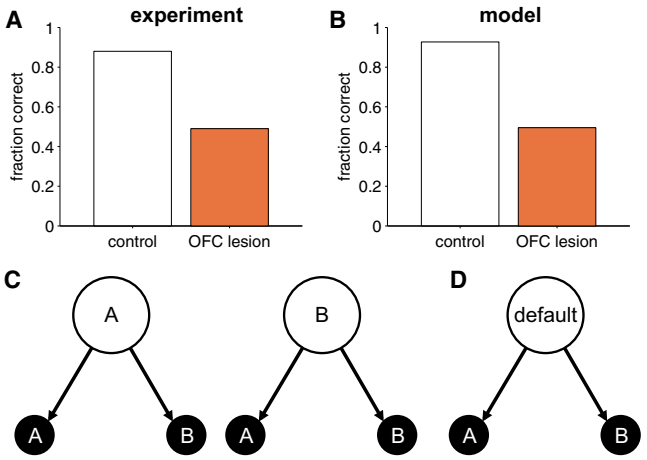


Figure 2. Delayed Alternation

(A) Experimental results showing the fraction of trials on which monkeys chose the correct option for control and OFC-lesioned animals.

(B) Model simulations on the same task.

(C) State representation used to model control animals, in which the state depends on the last action.

(D) Stimulus-bound state representation modeling the OFC-lesioned animals.

block, and, although initial learning is similarly slow for both the intact and the lesioned models, the lesioned model takes much longer to learn subsequent reversals (Figure 1B).

In general, the two states of our model of reversal learning can be seen as representing the two phases of the task ("reward more likely on left" and "reward more likely on right"). Thus, our representation generalizes to probabilistic reversal learning tasks (e.g., Tsuchida et al., 2010) in which the animal (and model) must infer what state it is in by using actions and outcomes from multiple previous trials (Gershman et al., 2010).

Delayed Alternation

The same reasoning can be applied to model the effect of OFC lesions on delayed alternation tasks (Mishkin et al., 1969; Miller and Orbach, 1972; Butters et al., 1973; Mishkin and Manning, 1978). In particular, we model Mishkin et al. (1969). In this task, monkeys made a series of choices between two options, one of which was paired with a reward. The rewarding option on the current trial was determined by the action on the previous trial such that reward was always made available for the action opposite to that on the previous trial. Thus, the monkeys had to learn to alternate their responses, which, due to a 5 s delay between trials, required memory of the last action. Control animals learned this task easily, ultimately performing at around 90% correct. However, monkeys with OFC lesions failed to perform better than chance even after 2,000 trials of training (Figure 2A).

We modeled the behavior of control animals with the state representation in Figure 2C, in which the current state is determined by the choice on the last trial (option A or B). With this state representation, the model learns the task easily (Figure 2B), and performance is only limited by the degree of "random" responding mediated by the inverse-temperature parameter β. To model OFC-lesioned animals, we again removed states that require memory, resulting in only one (default) state. With this state representation, the model can never learn to solve an alternation task; hence, performance remained at 50% correct in the lesioned case.

A crucial result is that even OFC-lesioned animals could learn the alternation task if the delay was removed (Miller and Orbach, 1972). Thus, the ability to learn about the value of alternation was unimpaired when a stimulus-bound two-state representation could be constructed but grossly impaired when a short delay required a memory-based state representation to be constructed. This suggests that value learning itself is unimpaired in OFC-lesioned animals and that the deficit lies in encoding of latent variables within the state representation.

Extinction

Our model also captures deficits in extinction that are caused by OFC lesions and makes a number of easily testable experimental predictions about postextinction phenomena (Bouton, 2004). In extinction, a previously trained association between an outcome and a certain state or action is changed such that the outcome is no longer available. Theories suggest that extinction does not cause unlearning of the original association but rather results in learning of a new, competing association (Bouton, 2004; Redish et al., 2007). Consequently, similar to the model of reversal learning, we modeled extinction with a two-state system (see also Gershman et al., 2010).

In particular, we consider the experiment in Butter et al. (1963). Here, monkeys were trained to press a lever for food reward. After 30 min of reinforced pressing, an extinction phase began—rewards were no longer available, and the extinction of responding was measured as the number of presses in successive 10 min blocks. The results, shown in Figure 3A, clearly demonstrate slower extinction for OFC-lesioned animals.

As previously, we modeled control animals (Figure 3C) with a two-state model—the animal is in state "P1" if the previous lever press was rewarded and in "P0" if it was not. These states naturally distinguish the two contexts of reinforcement and extinction. We considered two possible actions—either the animal presses the lever (P) or it does not (N). In our simulation, pressing the lever led to 1 U of reward during conditioning and to -0.2 U in extinction (representing the cost of performing the action). Not pressing always yielded 0 reward. Again, OFC-lesioned animals were modeled as having an impoverished state representation that included only one memory-free state (Figure 3D).

The simulation results are shown in Figure 3B. As in the experimental data, extinction in the two-state model was fast, given that extinction transitioned the animal into the P0 state, wherein new action values for P and N were learned (starting from low initial values). On the other hand, the one-state model of the OFC-lesioned animals could only learn to stop pressing the lever by changing the action value for P from a high value to a low one, which necessitated more trials.

As with reversal learning, in the case of probabilistic reinforcement, animals would need to integrate outcomes from multiple trials in order to infer which state or context (conditioning or extinction) they were in. For an exposition of how this kind of integration might be achieved, see Gershman et al. (2010).

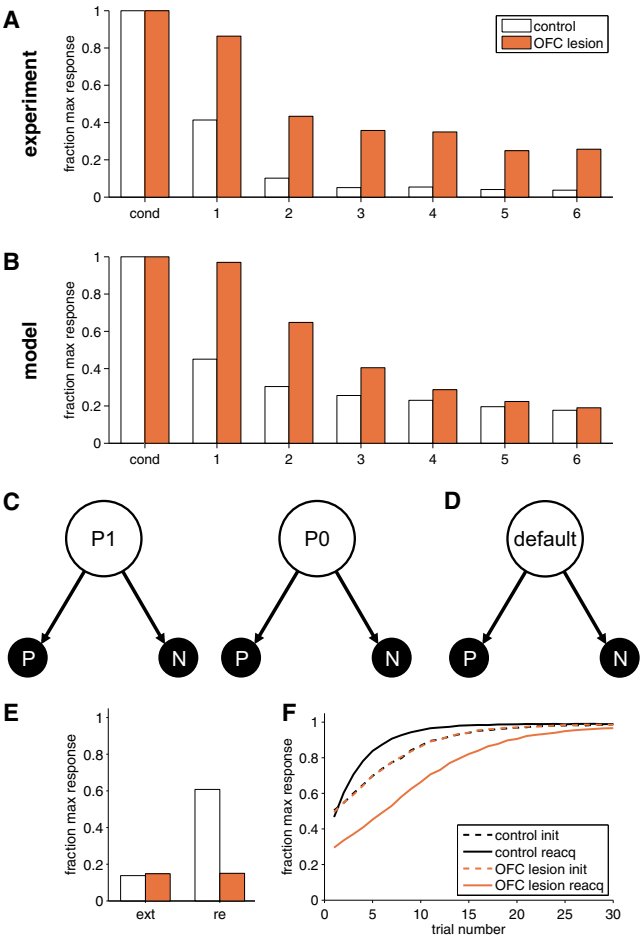


Figure 3. Extinction

(A) Experimental results. Lever press rates were normalized to the maximum response rate in conditioning. Adapted from Butter et al. (1963).

(B) Model results.

(C) State representation used to model the control group in which the state depends on the last outcome.

(D) State representation used to model the OFC lesion group with only a single state.

(E) Model predictions for extinction (ext) and spontaneous recovery (re).

(F) Model predictions for reacquisition. init, initial learning; reacq, reacquisition.

Postextinction Predictions

To assess the effectiveness of extinction and investigate what was learned during extinction, researchers often retest behavior after the extinction phase is completed. In particular, four classic effects— spontaneous recovery, reinstatement, rapid reacquisition, and renewal (Bouton, 2004)—have been taken as evidence that extinction training does not normally lead to permanent modification of the original association.

Our two-state model also exhibits these effects because the original associations between stimulus and reward are maintained in the P1 state and can be recovered when this state is reactivated. However, our one-state model predicts different results for OFC-lesioned animals because there the original association is, in fact, erased during extinction. For example, consider spontaneous recovery. Here, conditioning (cue or action → outcome) and extinction (action → no outcome) are performed. Then, after days or even weeks, animals undergo a test phase in which no outcome is available, and the propensity to perform the action is measured. Animals typically show recovery of responding at test to response levels that are greater than those at the end of extinction, with more recovery for longer waiting times between extinction and test.

Our two-state model accounts for this behavior if we assume that the passage of time causes the animal to be unsure whether it is in P1 or P0 at the start of testing. If a state is selected at random (for instance, with probability proportional to the time since it last occurred), then, on average, animals will respond more in the testing phase than at the end of the extinction phase. In contrast, when the OFC is lesioned (that is, in the one-state model) extinction truly does extinguish the original association, and, thus, our model predicts no spontaneous recovery (Figure 3E).

The model's predictions are even starker for rapid reacquisition (Napier et al., 1992; Ricker and Bouton, 1996), in which reconditioning of a stimulus → outcome association occurs more rapidly after extinction than in the original learning. The two-state model predicts this phenomenon, given that reconditioning will return the animal to the P1 state in which the old action preferences remain. However, we predict that OFC-lesioned animals will not show rapid reacquisition and may even show slightly slower reacquisition than original learning if there is a small cost associated with the response (Figure 3F).

Devaluation

The above tasks are predominantly explained with model-free RL (Daw et al., 2005). However, OFC is also thought to be important for model-based RL, in which animals use a learned model of reward contingencies to compute values. A prototypical example of such a model-based task is reinforcer devaluation (Colwill and Rescorla, 1985; Balleine and Dickinson, 1998). In this paradigm (Figure 4A), animals are trained to perform actions or associate cues with an outcome. When the outcome is devalued outside the context of the experiment, for example, by pairing its consumption with indigestion-inducing poison, actions that were trained with the devalued food are reduced at test, even if the test is performed in extinction conditions (that is, with no additional experience of the contingency between these actions and the devalued outcome). Such behavior indicates a capacity to "simulate" the consequences of actions within a cognitive model of the task and, thus, realize that a once valuable action would now lead to an unwanted outcome and, hence, should no longer be chosen. These mental simulations (Daw et al., 2005) involve taking imaginary paths through the states of the task, and we propose that these imagined (but not externally available) states are encoded in OFC. Consistent with this proposal, OFC lesions impair performance in devaluation experiments, causing lesioned animals to respond equally to devalued and nondevalued cues (Gallagher et al., 1999; Pickens et al., 2003; Izquierdo et al., 2004; but see Ostlund and Balleine, 2007).

We illustrate this effect through the results of Pickens et al. (2003), reproduced in Figure 4B. Here, rats were first taught

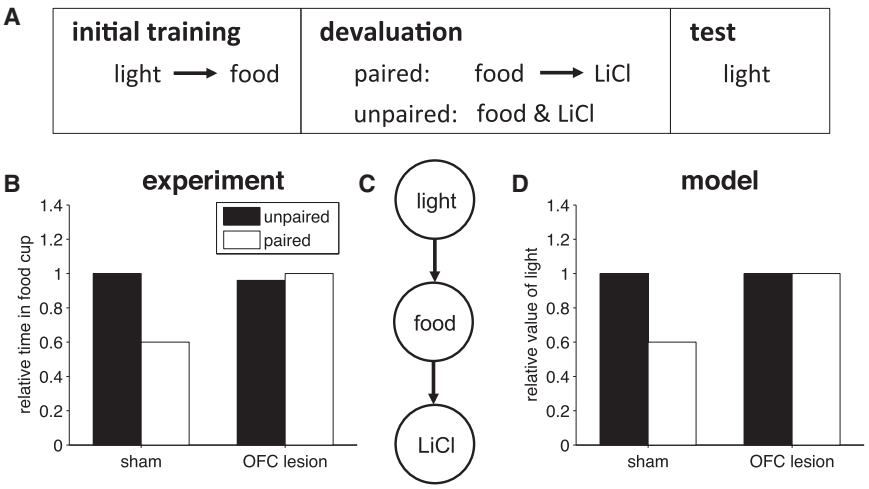


Figure 4. Devaluation

(A) First, animals are trained to associate a light with food. Then, the food is devalued by pairing it with an indigestion-inducing poison, LiCl. In a control condition, the food and LiCl are unpaired during devaluation. Finally, the extent of devaluation is indexed by measuring responding to the light.

(B) Experimental results from Pickens et al. (2003) showing relative responding to the food cup when the light is turned on for sham- and OFC-lesioned animals in the paired and unpaired condition.

(C) State representation of the devaluation task.

(D) Model results showing the relative value of the light for the sham- and OFC-lesioned models.

to associate a light cue with food. Subsequently, the food was devalued by pairing its consumption to the injection of lithium chloride. Then, a testing session measured the amount of time spent at the food cup when the light was presented. In order to establish a baseline level of responding, in a control condition, lithium chloride was administered in the second stage but was not paired with the food. Sham-lesioned animals showed reduced responding to the light in the paired condition relative to the unpaired condition, as if they were imagining the (never experienced) chain of events light → food → poison. OFC-lesioned animals showed no such change in behavior, as if they were incapable of such model-based reasoning.

We modeled the behavior of sham-lesioned animals using the state representation shown in Figure 4C. We assumed that sham-lesioned animals used a mixture of model-based and model-free learning to compute values. The model-free (MF) component learned a value, $V_{MF}(s)$, for each state s using standard temporal-difference prediction error learning. Specifically, when the model transitioned from state s to state s', it computed a prediction error

$$\delta = r + V_{MF}(s') - V_{MF}(s),$$

which was used to update the model-free value of state s

$$V_{MF}(s) \leftarrow V_{MF}(s) + \alpha\delta,$$

where α = 0.1 was the learning rate, and we assumed that the reward, r, was +1 during the initial learning phase and −1 after devaluation. Thus, the model-free component learns a positive value for the light state (given that it only ever experiences the light paired with food) and, in the devaluation stage, a negative value for the food state. In contrast, the model-based (MB) component uses the low value of the food state to update, even absent direct experience, the value of the light state through imagined simulation

$$V_{MB}(\text{light}) = V_{MF}(\text{food})p(\text{food}|\text{light}),$$

where $V_{MB}(\text{light})$ is the model-based value of the light, $V_{MF}(\text{food})$ is the model-free value of the food state, and $p(\text{food}|\text{light})$ is the estimated (learned) probability of the light state leading to the food state (set to 0.9 in our simulations). The total value of the light was a combination of the model-based and model-free values as in Daw et al. (2005),

$$V(\text{light}) = \zeta V_{MB}(\text{light}) + (1-\zeta)V_{MF}(\text{light}),$$

where we used ζ = 0.2 as the mixing fraction. According to this model, when the food is devalued, sham-lesioned animals compute a low value for the light (Figure 4D). However, the OFC-lesioned model lacks model-based planning abilities (ζ = 0) and, thus, shows no effect of devaluation.

This line of reasoning can also be used to explain other recent findings that are thought to reflect the role of OFC in model-based RL, such as sensory preconditioning (Jones et al., 2012), identity unblocking (McDannald et al., 2011), and Pavlovian overexpectation (Takahashi et al., 2009). In each case, OFC-dependent behavior or learning requires a form of mental simulation with the appropriate imagined (but not externally available) states.

Insights into the Role of OFC from Dopamine Firing Patterns

If the OFC is involved in RL, then, in addition to changes in behavior, lesions to the OFC should cause changes in the neural substrates of RL. Moreover, if our hypothesis is correct, then the changes in neural firing patterns should be consistent with the loss of non-stimulus-bound states but the preservation of all other RL processes. Motivated by this idea, in Takahashi et al. (2011), we investigated the effects of unilateral OFC lesions on prediction error signals in the ventral tegmental area (VTA) (Schultz et al., 1997).

In this experiment, described in detail in Takahashi et al. (2011), after a light came on, rats initiated a trial by entering an odor port, where they were presented with one of three odors. One odor indicated that the left fluid well would be paying out a reward on this trial (henceforth, a forced left trial), a second odor indicated that the rat must go right to get a reward (forced right), and the third odor indicated that both wells were paying out (free choice).

Critically, the amount and delay of the reward offered at each fluid well changed every 60 trials as shown in Figure 5A. In the

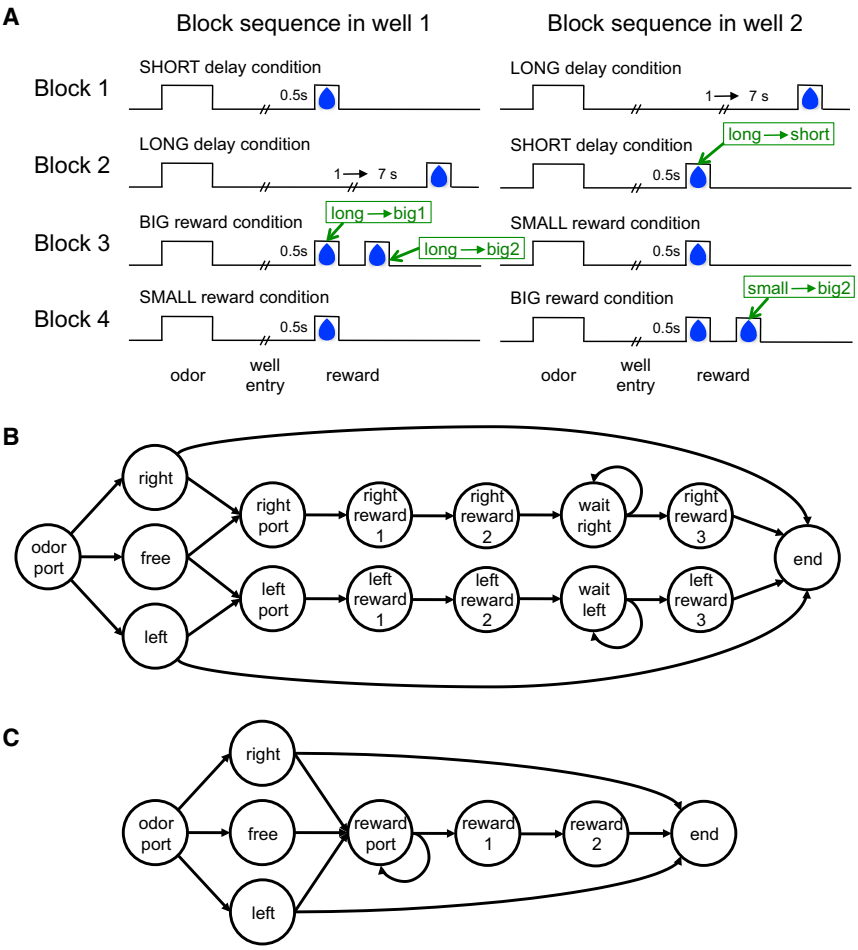


Figure 5. Task Design and State Representations for the Odor-Guided Choice Task

(A) Time course of reward for the different blocks. Times associated with positive prediction errors caused by unexpected reward are labeled in green. Figure adapted from Takahashi et al. (2011).

(B) State representation used to model sham-lesioned controls.

(C) State representation used to model OFC-lesioned animals.

first block of trials, one well paid out one drop of juice after a short delay, whereas the other paid out one drop after a longer delay. In the second block, these reward contingencies were reversed. In the third block, the two wells offered a big reward (two drops of juice) and a small reward (one drop of juice), and these contingencies reversed again in the fourth and final block of the session. The experiment was repeated with similar sessions daily.

State Representations of the Task

We modeled both the rats' behavior and the firing of dopaminergic VTA neurons. The true generative state representation of the task (that is, the representation that accords with the experimenter-defined reward contingencies) is depicted in Figure 5B. A trial begins when the rat moves to the odor port (indicated by the "odor port" state). Then, an odor is presented, signaling a forced left ("left" state), free choice ("free"), or forced right ("right") trial. In forced right trials or free choice trials, if the rat chooses to go to the right fluid well, then it arrives at the "right port" state. Over time, the state changes to "right reward 1," which denotes the time of juice delivery in blocks in which a small or short reward is delivered as well as the time of the first drop of juice if a big reward is to be delivered. The state continues to transition to "right reward 2," the time of the second drop in big reward trials, "wait right," a state that represents the unpredictable delay before reward on long reward trials, "right reward 3," which is the reward delivery time in long reward trials, and finally the "end" state. In contrast, if the rat chooses to go to the left fluid well on a right trial, then the task transitions (without reward) to the end state, signifying the end of the trial. A similar sequence of states occurs for the left reward arc. Through repeated experience with the task, it is reasonable to assume that rats learned this correct representation of the task contingencies or at least the breakdown of the task into fairly well-delineated states. Thus, we assumed this representation when modeling the sham-lesioned group.

Although a straightforward description of the task, some states in this sequence are not directly tied to fully observable stimuli. For instance, the "right port" state does not correspond directly to the physical right port, given that going to that same physical port on a forced left trial will not lead to this state. Moreover, we assume that the two physical food ports are relatively indistinguishable from the vantage point of a rat waiting for reward with its nose in the port. Of course, remembering the previous odor and action will uniquely identify the state. However, this is precisely the type of information that we hypothesized would be missing from the state representation if OFC function was compromised. We also assume that temporal information is not available externally, and, thus, OFC-lesioned rats cannot distinguish reward states that are only separated by the passage of time. Altogether, these assumptions define the OFC-lesioned state representation depicted in Figure 5C, which involves a single "reward port" state and two rather than four states in the reward arc ("reward 1" representing the first drop of juice, and "reward 2" representing the second drop on big trials, externally distinguishable from reward 1 because it is immediately preceded by a drop of juice).

Prediction Errors

Our goal was to understand OFC-lesion-induced changes in prediction error signals recorded from dopaminergic neurons in the

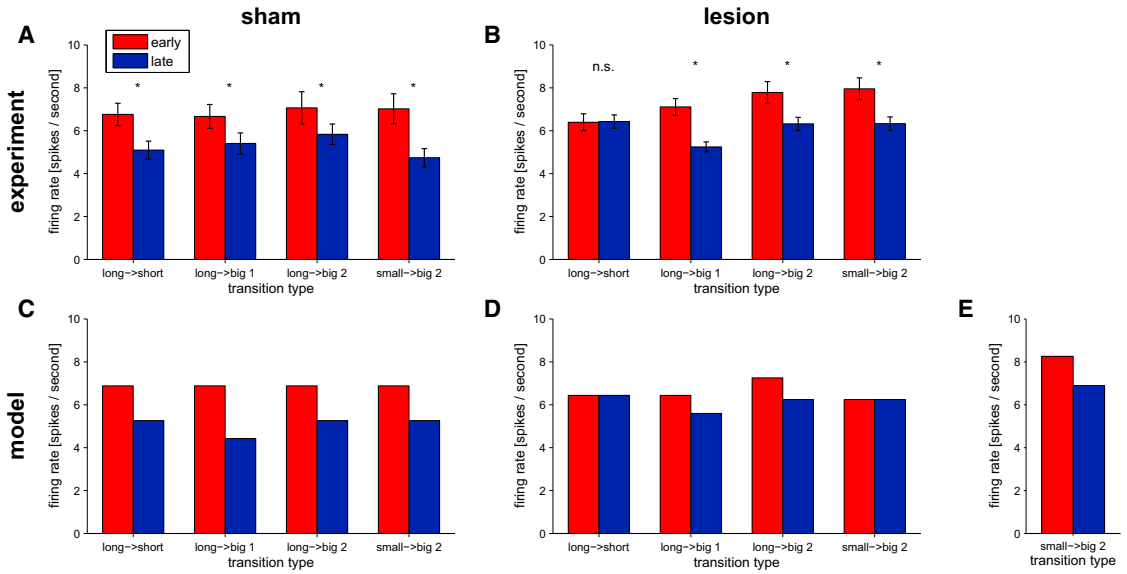


Figure 6. Firing of Dopaminergic VTA Neurons at the Time of Unexpected Reward Early and Late in a Block

Unlike in Takahashi et al. (2011), where neural responses were averaged over the different types of unexpected reward delivery, here we divided the data into the four different cases, indicated by the green annotations in Figure 5A: the short reward after the long to short transition between blocks 1 and 2 (long → short), the arrival of the first (long → big1) and second (long → big2) drops of reward after the long to big transition between blocks 2 and 3, and the second drop of the small to big transition between blocks 3 and 4 (small → big2). Early, first two trials; late, last five trials.

(A) Experimental data for sham-lesioned controls (n = 30 neurons; error bars represent SEM).

(B) Experimental data for the OFC-lesioned group (n = 50 neurons; error bars represent SEM).

(C) Model predictions for the sham-lesioned animals.

(D) Model predictions for OFC-lesioned animals.

(E) Model predictions for the small → big2 transition when taking into account the variable third drop of juice.

VTA (Schultz et al., 1997). These signals convey the difference between predicted and actual outcomes (Sutton and Barto, 1998; see the Supplemental Information for a detailed description) and, in theory, should depend strongly on how the task is parsed into states.

There are two points in a trial in which we can expect prediction errors—the time of reward (if the reward obtained is different from the expected reward) and the time of odor presentation (where prediction errors are due to the difference between the reward predicted after sampling the odor in comparison to the prediction before odor onset). Indeed, although behavior in both groups was equated because of the lesion being unilateral, Takahashi et al. (2011) observed small but clear differences between the firing of dopaminergic neurons on the side of the lesion in sham- and OFC-lesioned animals, the specific pattern of which was captured by our model. Here, we look more closely at these differences at the time of reward. Results at the time of the odor are presented in the Supplemental Information.

Figure 6 shows the firing of VTA neurons at the time of unexpected reward. These rewards are unexpected at the start of a block, after reward contingencies have changed unexpectedly, but, given learning with the correct state representation, should be predicted by the end of the block. Thus, we compared the first two (early) trials to the last five (late) trials of a block in order to test for effects of learning (see the Supplemental Information for additional details).

Sham-lesioned animals (Figure 6A) showed a decrease in prediction error firing between early and late trials in all cases ($p < 0.05$). Importantly, there was no effect of transition type on the difference between early and late prediction errors. These findings are consistent with the predictions of the intact RL model (Figure 6C).

In contrast, in the OFC-lesioned animals, the difference in firing between early and late trials was wholly absent ($p = 0.74$) in the "long" to "short" transition at the beginning of the second block (Figure 6B). The lesioned model predicts the lack of elevated prediction errors at the beginning of this block. This is because the lesioned model cannot learn different predictions for reward on the left and right ports but, rather, learns to predict the average reward in the block. For the lesioned model, both blocks involve early reward on a seemingly random half of the trials and delayed reward on the other half. However, the model does predict positive prediction errors on block switches in which the average reward, over both options, increases. This can be seen in the data for the "long" to "big" transition from block two to three, both for the first drop (previously delayed on half the trials and now surprisingly reliably early) and the second drop (which did not appear before and now appears on half the trials).

The lesioned model also predicts no change in prediction errors for the "small" to "big2" transition at the beginning of the fourth block, a prediction seemingly not borne out in the data. However, in Takahashi et al. (2011)'s experiment, on

some trials in the fourth block, an extra third drop of water was added to "big" trials if the rat appeared to be losing interest in the task. Although the timing of this manually applied third drop was not recorded, examination of the spike raster plots in which the response of individual neurons to each drop is clearly visible (for an example, see Figure S1 available online) shows the third drop in 13 of the 14 examples. Adding this third drop indeed changes the average available reward, aligning the lesioned model's predictions with the experimental results (Figure 6E). Therefore, a prediction of the model is that, without the third drop, this difference in firing between early and late trials for the "small → big2" transition would disappear.

Importantly, these neural results are inconsistent with prominent ideas according to which the OFC contributes to RL by directly encoding expected value. As detailed in Takahashi et al. (2011), an inability to learn or represent values would predict that dopaminergic firing at the time of reward would not change throughout a block, because obtained rewards would be completely unpredictable—a prediction clearly inconsistent with the data. Observed differences in firing at the time of the odor are also inconsistent with this idea that OFC encodes value (Figure S2). Altogether, the behavioral and neural results suggest that, rather than representing values per se, the OFC is involved in representing unobservable states, which are often essential for learning or calculation of accurate values.

DISCUSSION

We have proposed a role for the OFC in encoding the current state in a cognitive map of task space and shown how this role would manifest in associative learning and decision making tasks known to depend on the OFC. Specifically, we have proposed that the OFC is necessary for disambiguating states that are not perceptually distinct. Our theory explains classic findings in reversal learning, delayed alternation, extinction, and devaluation, along with neural results from a recent lesion experiment (Takahashi et al., 2011) and makes easily testable experimental predictions about postextinction phenomena in animals with OFC lesions. Now, we turn to discuss the implications of our theory and relate it to other results and models of OFC function.

Neural Activity in OFC

According to our theory, we ought to be able to see state-related signals in the activity of OFC neurons. Thus, the question arises: what is the neural signature of a state representation for RL? We propose two conditions that should be satisfied by a brain region encoding states: (1) representation—all the variables that comprise the current state, as it is defined for the purpose of RL, are encoded in the brain area—and (2) specificity—irrelevant variables that are not part of the current state are not encoded in the area. The first condition ensures that all relevant variables are at least present in the area, whereas the second condition rules out areas whose encoding is not task specific. Our theory predicts that neural representations in the OFC would satisfy these two conditions across tasks and, specifically, that variables that are not necessarily perceptually available (such as memory for previous actions or outcomes) would be represented in the OFC, but only if they are required for the current task.

Representation

Although no experiments have explicitly tested these neural predictions, several results are consistent with the first condition—in particular, in tasks in which relevant variables are not externally available. For instance, our model implies that both the previous choice and the previous outcome should be encoded in OFC in reversal learning tasks, which has been found (Schoenbaum and Eichenbaum, 1995; Sul et al., 2010; the latter also found these variables in dorsolateral prefrontal cortex [dlPFC] and anterior cingulate cortex [ACC]). In a probabilistic RL task, Hampton et al. (2006) showed that fMRI activation in ventromedial prefrontal cortex close to OFC was correlated with the underlying task state in a Bayesian model.

A related experiment is the "shift-stay" paradigm (Tsujimoto et al., 2009, 2011), in which monkeys choose between two options with a strategy cue, presented at the start of a trial, instructing them as to whether the rewarded response is to "stay" with their last choice or "switch" to the other option. Such a task is readily solved with two states that combine the last choice and strategy. Intriguingly, Tsujimoto et al. (2009, 2011) found neural correlates of these variables in OFC.

Similarly, in delayed match-to-sample tasks, OFC encodes the remembered sample, a critical component of the state (Ramus and Eichenbaum, 2000; Lara et al., 2009; the latter study is especially interesting becasue it included "distractor" drops of water that did not elicit OFC firing), and, in fMRI studies, OFC activity has been associated with context-dependent disambiguation of navigational routes (Brown et al., 2010) and task rules (Nee and Brown, 2012).

Specificity

Addressing the specificity condition is more difficult, due to it being hard to know exactly what state representation an animal is using in any given task. However, one could look for differences in OFC representations in tasks with similar stimuli but different underlying states. If OFC encodes the states of the task, even subtle changes in the task should lead to changes in OFC firing. This was indeed shown in two tasks by Schoenbaum and Eichenbaum (1995) and Ramus and Eichenbaum (2000) (reviewed in Schoenbaum et al., 2003b). In the first task, four of eight odors predicted that a response at a nearby fluid well would be rewarded. In the second task, eight odors were used in the same apparatus, but reward on a given trial was not predicated on odor identity but, rather, on whether the odor on the current trial was different from that presented on the previous trial. In both cases, the odor was relevant for performance. However, in the first task, the identity of the odor was critical for predicting reward, whereas, in the latter task, whether or not the odors on consecutive trials matched was critical. Intriguingly, approximately 77% of OFC neurons were odor selective when odor identity was relevant, whereas only 15% of OFC neurons were odor selective in the task in which match, but not identity, predicted reward. Furthermore, in that latter task, 63% of OFC neurons encoded whether the odor was a match or a nonmatch.

Simmons and Richmond (2008) also demonstrated that small changes in a task can cause significant changes to OFC representations. In their task, monkeys were rewarded after

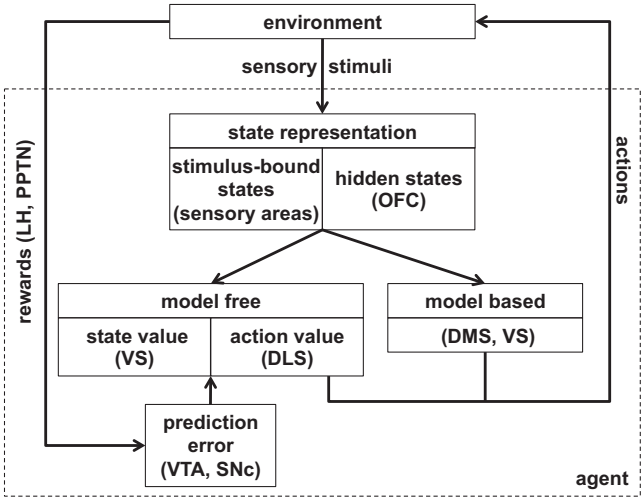


Figure 7. Schematic of Neural RL with Hypothesized Mapping of Functions to Brain Areas

The environment provides rewards and sensory stimuli to the brain. Rewards, represented in areas such as the lateral habenula (LH) and the pedunculopontine nucleus (PPTN) are used to compute prediction error signals in ventral tegmental area (VTA) and substantia nigra pars compacta (SNc). Sensory stimuli are used to define the animal's state within the current task. The state representation might involve both a stimulus-bound (externally observable) component, which we propose is encoded in both OFC and sensory areas, and a hidden (unobservable) component, which we hypothesize is uniquely encoded in OFC. State representations are used as scaffolding for both model-free and -based RL. Model-free learning of state and action values occurs in ventral striatum (VS) and dorsolateral striatum (DLS), respectively, whereas model-based learning occurs in dorsomedial striatum (DMS) as well as VS.

one, two, or three correct trials in a row, a number selected randomly after each reward. In a "valid cue" condition, background color indicated to the monkey the number of trials before the next reward, whereas, in a "random cue" condition, there was no relation between background color and number of trials to reward. As a result, the outcome of the previous trial was informative for reward prediction only in the random cue condition, because after a rewarded trial, the next trial would be rewarded only on one-third of the cases (a one-correct-trial requirement), whereas, after an unrewarded trial, the next trial would be rewarded on one-half of the cases (a two-correct- or a three-correct-trial requirement). Indeed, far fewer neurons encoded the last reward in the valid cue condition (25%), where it was not informative regarding task state, than in the random cue condition (50%). Furthermore, we predict that OFC encoding of background color should be different across the two conditions in this task.

Subdivisions of the OFC

The OFC is not a single, homogeneous region—connectivity analyses suggest a division into distinct medial and lateral networks in monkeys (Carmichael and Price, 1996), humans (Croxson et al., 2005; Kahnt et al., 2012), and rats (Price, 2007). Recent results implicate medial OFC in encoding economic value and lateral OFC in more complex functions, such as credit assignment and model-based RL (Noonan et al., 2010; Rudebeck and Murray, 2011a, 2011b; Noonan et al., 2012). It seems likely that our theory pertains more to the lateral than the medial OFC, although the lesion studies we discussed typically targeted the entire OFC. Thus, more work is needed in order to precisely localize the representation of task states within OFC subregions.

Interspecies Differences in OFC

We have not distinguished between rats and monkeys, treating what is defined as "OFC" in these very different species as essentially the same area. However, it is important to note that there are large differences in anatomy across species, and OFC in rats has a very different cytoarchitecture than OFC in monkeys and humans (Wise, 2008; Wallis, 2012). These stark anatomical differences have lead some researchers to question whether many of the frontal structures found in primates, including OFC, have analogs in the rat (Wise, 2008; but see Preuss, 1995).

Interestingly, despite these differences, there are strong interspecies similarities at the level of connectivity (Carmichael and Price, 1996; Price, 2007), neural activity, and function. This is particularly true for OFC, perhaps more so than any other prefrontal region (Preuss, 1995). For example, lesions to OFC cause similar deficits in reversal learning (Teitelbaum, 1964; Butter, 1969; Jones and Mishkin, 1972; Rolls et al., 1994; Dias et al., 1996; Meunier et al., 1997; McAlonan and Brown, 2003; Schoenbaum et al., 2002, 2003a; Chudasama and Robbins, 2003; Bohn et al., 2003; Izquierdo et al., 2004; Kim and Ragozzino, 2005), extinction (Butter, 1969; McEnaney and Butter 1969), and devaluation (Gallagher et al., 1999; Gottfried et al., 2003; Izquierdo et al., 2004) across species, and neural firing in different species in these tasks is also very similar (Thorpe et al., 1983; Schoenbaum and Eichenbaum 1995; Critchley and Rolls, 1996a, 1996b; Schoenbaum et al., 1999; Gottfried et al., 2003; O'Doherty et al., 2002; Morrison and Salzman, 2009). We suggest that OFC encodes the current task state in all of these species. Animals such as rodents are perhaps limited in the complexity of the state that can be represented in their relatively small OFC, whereas humans, who have a much more developed OFC, are able to deal with highly complex tasks that involve many hidden states.

Interaction with Other Brain Areas

Figure 7 illustrates how our theory of OFC fits into a larger model of RL in the brain. In particular, we propose that OFC encodes task states, drawing on both stimulus-bound (externally available) and memory-based (or internally inferred) information. These states provide scaffolding for model-free RL in a network involving ventral striatum (encoding state values $V(s)$) and dorsolateral striatum (encoding state-action values $Q(a,s)$). This system is trained by prediction errors computed in VTA and substantia nigra pars compacta, where reward input from areas such as the lateral habenula, hypothalamus, and pedunculopontine nucleus is compared to predicted values from the ventral and dorsolateral striatum. State information in OFC is also critical for model-based RL (Sutton and Barto, 1998; Daw et al., 2005), which makes use of learned relationships between

states in order to plan a course of action through mental simulation of imagined states.

In parallel, we propose that a purely stimulus-bound state representation encoded in sensory areas can also be used for learning and decision making. These stimulus-bound states are the sole basis for RL when OFC is lesioned but may also be used for learning in intact animals. For instance, concurrent use of a suboptimal stimulus-bound state representation could account for some erroneous credit assignment seen even in sham-lesioned control animals, as evidenced in Walton et al. (2010).

Other Areas that Might Encode Task States

Several other areas have been proposed to encode task states. Perhaps chief among these is the hippocampus. Like OFC, lesions in hippocampus cause deficits in spatial reversal learning (Teitelbaum, 1964) and prevent postextinction renewal (Ji and Maren, 2007). However, this is true only when states are defined according to spatial location. Hippocampal lesions seem to have no effect on nonspatial reversal learning, whereas OFC lesions generally affect all types of reversal (Teitelbaum, 1964).

On the basis of neural recordings that showed that choices, stimuli, and rewards were encoded in neurons in the dlPFC, Seo et al. (2007) proposed that dlPFC encodes task states. Indeed it seems clear that dlPFC satisfies the representation condition; however, this area is less able to satisfy the specificity condition, given that dlPFC seems to encode combinations of task relevant and task irrelevant stimuli. An intriguing possibility is that dlPFC encodes a reservoir of candidate state variables from which OFC constructs the current state with the variables found to be most relevant to the current task (Otto et al., 2009).

There is also clearly related literature on rule-based behavior that does not explicitly mention state representations. Indeed, the outcome of learning with a sophisticated state representation is a set of action values that essentially determine rules for the task by specifying the most rewarding action in each state. Such rule-based behavior has long been thought to depend on dlPFC (Banich et al., 2000; MacDonald et al., 2000; Petrides, 2000), and recent imaging studies have further localized this function to the inferior frontal sulcus and inferior frontal junction (Brass et al., 2008). However, it is important to distinguish between a state, which is an abstract representation of the current location in a task, and a rule, which specifies a mapping from conditions to actions. These two functions may be associated with different brain areas, consistent with neuroimaging results in which tasks involving the implementation of explicit rules invoke dlPFC activity (Banich et al., 2000; MacDonald et al., 2000; Petrides, 2000), whereas tasks requiring nontrivial assignment of reward in a complex state space elicit activations in the lateral OFC (Noonan et al., 2011). Furthermore, Buckley et al. (2009) found differential effects of lesions to the OFC and the dlPFC in a monkey analog of the Wisconsin card sorting task—OFC lesions diminished monkeys' ability to learn new reward associations, consistent with an impaired representation of state, whereas dlPFC lesions decreased the ability to use a previously learned rule.

Finally, one might argue that the encoding of state information is too general a function to be ascribed to a single brain region and that these representations are widely distributed, perhaps over the entire prefrontal cortex. However, this seems at odds with the specificity of deficits that occur as a result of OFC lesions (Buckley et al., 2009)—if the encoding of state were more distributed, then one might expect that lesions to other prefrontal areas would cause similar deficits. Furthermore, the OFC might be uniquely well placed to integrate disparate pieces of information, including sensory information and latent variables such as memories, in order to compute the current state because of its afferent connectivity, which is different from that of other prefrontal areas. For instance, the OFC is the only prefrontal area to receive sensory input from all sensory modalities; it has strong connections to areas such as dlPFC, ACC, and the hippocampus, and it has strong reciprocal connections with subcortical regions such as striatum and amygdala, which are critical to the representation of reward (Carmichael and Price 1995a, 1995b; Murray et al., 2011).

Relation to Other Theories of OFC Function

Over the years, many hypotheses of OFC function have been put forth. For example, that the OFC inhibits prepotent responses (Ferrier, 1876; Fuster, 1997) or that it represents bodily markers for affective state (Damasio, 1994). Here, we discuss two popular recent accounts that also relate OFC function to RL.

OFC Encodes Economic Value

Perhaps the dominant theory of OFC function in the past few years has been the idea that OFC encodes economic value (Padoa-Schioppa and Assad, 2006). Interpreted in the language of RL, this essentially implies that OFC encodes state values, $V(s)$.

Recent studies have begun to cast doubt on this account. In particular, some patterns of firing in OFC neurons are hard to interpret as a pure value signal. For instance, OFC neurons have been found to encode variables such as spatial location (Roesch et al., 2006; Feierstein et al., 2006; Furuyashiki et al., 2008), satiety (de Araujo et al., 2006), uncertainty (Kepecs et al., 2008), and taste (Padoa-Schioppa and Assad, 2008). Indeed, our own results, specifically the preservation of VTA firing at the time of the odor after OFC lesions (Takahashi et al., 2011), are inconsistent with the view that OFC provides values to the computation of prediction errors in dopamine neurons.

A more recent idea is that, rather than storing learned values, OFC computes values in a model-based way to enable flexible economic decision making and choices among many different options in many different situations without explicitly storing a previously learned value for each (Padoa-Schioppa, 2011). This account fits well with our theory. In particular, although it is not yet clear whether OFC itself is involved in computing model-based values, we propose that the OFC provides the state information that allows these computations to occur and is thus essential to such economic decision making.

OFC Takes Part in Solving the Credit Assignment Problem

Our theory is closely related to a recent proposal that OFC (in particular, lateral OFC) acts to solve the credit assignment problem; i.e., to decide which reward should be attributed to which action for learning (Walton et al., 2010; Noonan et al., 2012). This idea shares many properties with our state-representation

hypothesis, given that correctly keeping track of the current state allows credit to be assigned appropriately. However, in our theory, credit assignment itself is not damaged by the loss of the OFC, but, rather, the states to which credit is assigned are changed. This subtle distinction is an important one because it points to a key difference between the theories: our theory predicts that OFC lesions will not appear to cause a deficit in credit assignment in tasks in which stimulus-bound states suffice. Moreover, the credit-assignment hypothesis suggests that past actions should always be represented in OFC for credit assignment, whereas we predict that past actions will only be encoded when they are important for determining the states of the task.

More generally, our theory accounts for the role for OFC in a wide range of tasks, not only reversal learning, delayed alternation, and extinction, but also devaluation, sensory preconditioning, and so on. Indeed, it predicts involvement in any situation where task states are not stimulus bound. As such, our theory provides a unifying account of OFC function that can be tested (and disproved) in a variety of different tasks.

SUPPLEMENTAL INFORMATION

Supplemental Information contains Supplemental Experimental Procedures and two figures and can be found with this article online at http://dx.doi.org/10.1016/j.neuron.2013.11.005.

ACKNOWLEDGMENTS

This work was supported in part by T32 training grant 5T32MH065214 in quantitative neuroscience from the NIMH and NIDA (R.C.W.), by NIMH grant 1R01MH098861 (Y.N.), and by the Intramural Research Program at the National Institute on Drug Abuse (G.S. and Y.T.). The opinions expressed in this article are the authors' own and do not reflect the views of the NIH/DHHS.

Accepted: November 5, 2013
Published: January 22, 2014

REFERENCES

Balleine, B.W., and Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology *37*, 407–419.

Banich, M.T., Milham, M.P., Atchley, R.A., Cohen, N.J., Webb, A., Wszalek, T., Kramer, A.F., Liang, Z., Barad, V., Gullett, D., et al. (2000). Prefrontal regions play a predominant role in imposing an attentional 'set': evidence from fMRI. Brain Res. Cogn. Brain Res. *10*, 1–9.

Bohn, I., Giertler, C., and Hauber, W. (2003). Orbital prefrontal cortex and guidance of instrumental behaviour in rats under reversal conditions. Behav. Brain Res. *143*, 49–56.

Bouton, M.E. (2004). Context and behavioral processes in extinction. Learn. Mem. *11*, 485–494.

Brass, M., Derrfuss, J., and von Cramon, Y. (2008). The Role of the Posterior Frontolateral Cortex in Task Related Control. In Neuroscience of Rule-Guided Behavior, S.A. Bunge and J.D. Wallis, eds. (New York: Oxford University Press).

Brown, T.I., Ross, R.S., Keller, J.B., Hasselmo, M.E., and Stern, C.E. (2010). Which way was I going? Contextual retrieval supports the disambiguation of well learned overlapping navigational routes. J. Neurosci. *30*, 7414–7422.

Buckley, M.J., Mansouri, F.A., Hoda, H., Mahboubi, M., Browning, P.G.F., Kwok, S.C., Phillips, A., and Tanaka, K. (2009). Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. Science *325*, 52–58.

Butter, C.M. (1969). Perseveration in extinction and in discrimination reversal tasks following selective frontal ablations in macaca mulatta. Physiol. Behav. *4*, 163–171.

Butter, C.M., Mishkin, M., and Rosvold, H.E. (1963). Conditioning and extinction of a food-rewarded response after selective ablations of frontal cortex in rhesus monkeys. Exp. Neurol. *7*, 65–75.

Butters, N., Butter, C., Rosen, J., and Stein, D. (1973). Behavioral effects of sequential and one-stage ablations of orbital prefrontal cortex in the monkey. Exp. Neurol. *39*, 204–214.

Carmichael, S.T., and Price, J.L. (1995a). Limbic connections of the orbital and medial prefrontal cortex in macaque monkeys. J. Comp. Neurol. *363*, 615–641.

Carmichael, S.T., and Price, J.L. (1995b). Sensory and premotor connections of the orbital and medial prefrontal cortex of macaque monkeys. J. Comp. Neurol. *363*, 642–664.

Carmichael, S.T., and Price, J.L. (1996). Connectional networks within the orbital and medial prefrontal cortex of macaque monkeys. J. Comp. Neurol. *371*, 179–207.

Chudasama, Y., and Robbins, T.W. (2003). Dissociable contributions of the orbitofrontal and infralimbic cortex to pavlovian autoshaping and discrimination reversal learning: further evidence for the functional heterogeneity of the rodent frontal cortex. J. Neurosci. *23*, 8771–8780.

Colwill, R.M., and Rescorla, R.A. (1985). Postconditioning devaluation of a reinforcer affects instrumental responding. J. Exp. Psychol. *11*, 120–132.

Critchley, H.D., and Rolls, E.T. (1996a). Hunger and satiety modify the responses of olfactory and visual neurons in the primate orbitofrontal cortex. J. Neurophysiol. *75*, 1673–1686.

Critchley, H.D., and Rolls, E.T. (1996b). Olfactory neuronal responses in the primate orbitofrontal cortex: analysis in an olfactory discrimination task. J. Neurophysiol. *75*, 1659–1672.

Croxson, P.L., Johansen-Berg, H., Behrens, T.E.J., Robson, M.D., Pinsk, M.A., Gross, C.G., Richter, W., Richter, M.C., Kastner, S., and Rushworth, M.F. (2005). Quantitative investigation of connections of the prefrontal cortex in the human and macaque using probabilistic diffusion tractography. J. Neurosci. *25*, 8854–8866.

Damasio, A.R. (1994). Descartes' error: Emotion, reason, and the human brain. (New York: Putnam).

Daw, N.D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat. Neurosci. *8*, 1704–1711.

de Araujo, I.E., Gutierrez, R., Oliveira-Maia, A.J., Pereira, A., Jr., Nicolelis, M.A.L., and Simon, S.A. (2006). Neural ensemble coding of satiety states. Neuron *51*, 483–494.

Dias, R., Robbins, T.W., and Roberts, A.C. (1996). Dissociation in prefrontal cortex of affective and attentional shifts. Nature *380*, 69–72.

Feierstein, C.E., Quirk, M.C., Uchida, N., Sosulski, D.L., and Mainen, Z.F. (2006). Representation of spatial goals in rat orbitofrontal cortex. Neuron *51*, 495–507.

Ferrier, D. (1876). The functions of the brain. (New York: G. P. Putnam's Sons).

Furuyashiki, T., Holland, P.C., and Gallagher, M. (2008). Rat orbitofrontal cortex separately encodes response and outcome information during performance of goal-directed behavior. J. Neurosci. *28*, 5127–5138.

Fuster, J.M. (1997). The prefrontal cortex. (New York: Lippin-Ravencott).

Gallagher, M., McMahan, R.W., and Schoenbaum, G. (1999). Orbitofrontal cortex and representation of incentive value in associative learning. J. Neurosci. *19*, 6610–6614.

Gershman, S.J., and Niv, Y. (2010). Learning latent structure: carving nature at its joints. Curr. Opin. Neurobiol. *20*, 251–256.

Gershman, S.J., and Niv, Y. (2013). Perceptual estimation obeys Occam's razor. Front. Psychol. *4*, 623.

Gershman, S.J., Blei, D.M., and Niv, Y. (2010). Context, learning, and extinction. Psychol. Rev. *117*, 197–209.

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

Hampton, A.N., Bossaerts, P., and O'Doherty, J.P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J. Neurosci. *26*, 8360–8367.

Izquierdo, A., Suda, R.K., and Murray, E.A. (2004). Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. J. Neurosci. *24*, 7540–7548.

Ji, J., and Maren, S. (2007). Hippocampal involvement in contextual modulation of fear extinction. Hippocampus *17*, 749–758.

Jones, B., and Mishkin, M. (1972). Limbic lesions and the problem of stimulus—reinforcement associations. Exp. Neurol. *36*, 362–377.

Jones, J.L., Esber, G.R., McDannald, M.A., Gruber, A.J., Hernandez, A., Mirenzi, A., and Schoenbaum, G. (2012). Orbitofrontal cortex supports behavior and learning using inferred but not cached values. Science *338*, 953–956.

Kaelbling, L.P., Littman, M.I., and Cassandra, A.R. (1998). Planning and acting in partially observable stochastic domains. Artif. Intell. *101*, 99–134.

Kahnt, T., Chang, L.J., Park, S.Q., Heinzle, J., and Haynes, J.-D. (2012). Connectivity-based parcellation of the human orbitofrontal cortex. J. Neurosci. *32*, 6240–6250.

Kepecs, A., Uchida, N., Zariwala, H.A., and Mainen, Z.F. (2008). Neural correlates, computation and behavioural impact of decision confidence. Nature *455*, 227–231.

Kim, J., and Ragozzino, M.E. (2005). The involvement of the orbitofrontal cortex in learning under changing task contingencies. Neurobiol. Learn. Mem. *83*, 125–133.

Lara, A.H., Kennerley, S.W., and Wallis, J.D. (2009). Encoding of gustatory working memory by orbitofrontal neurons. J. Neurosci. *29*, 765–774.

MacDonald, A.W., 3rd, Cohen, J.D., Stenger, V.A., and Carter, C.S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. Science *288*, 1835–1838.

McAlonan, K., and Brown, V.J. (2003). Orbital prefrontal cortex mediates reversal learning and not attentional set shifting in the rat. Behav. Brain Res. *146*, 97–103.

McDannald, M.A., Lucantonio, F., Burke, K.A., Niv, Y., and Schoenbaum, G. (2011). Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. J. Neurosci. *31*, 2700–2705.

McEnaney, K.W., and Butter, C.M. (1969). Perseveration of responding and nonresponding in monkeys with orbital frontal ablations. J. Comp. Physiol. Psychol. *68*, 558–561.

Meunier, M., Bachevalier, J., and Mishkin, M. (1997). Effects of orbital frontal and anterior cingulate lesions on object and spatial memory in rhesus monkeys. Neuropsychologia *35*, 999–1015.

Miller, M.H., and Orbach, J. (1972). Retention of spatial alternation following frontal lobe resections in stump-tailed macaques. Neuropsychologia *10*, 291–298.

Mishkin, M., and Manning, F.J. (1978). Non-spatial memory after selective prefrontal lesions in monkeys. Brain Res. *143*, 313–323.

Mishkin, M., Vest, B., Waxler, M., and Rosvold, H.E. (1969). A re-examination of the effects of frontal lesions on object alternation. Neuropsychologia *7*, 357–363.

Morris, G., Arkadir, D., Nevet, A., Vaadia, E., and Bergman, H. (2004). Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. Neuron *43*, 133–143.

Morrison, S.E., and Salzman, C.D. (2009). The convergence of information about rewarding and aversive stimuli in single neurons. J. Neurosci. *29*, 11471–11483.

Murray, E.A., O'Doherty, J.P., and Schoenbaum, G. (2007). What we know and do not know about the functions of the orbitofrontal cortex after 20 years of cross-species studies. J. Neurosci. *27*, 8166–8169.

Murray, E.A., Wise, S.P., and Rhodes, S.E.V. (2011). What Can Different Brains Do with Reward? In Neurobiology of Sensation and Reward, J.A. Gottfried, ed. (Boca Raton: LCRC Press).

Napier, R.M., Macrae, M., and Kehoe, E.J. (1992). Rapid reacquisition in conditioning of the rabbits nictitating membrane response. J. Exp. Psychol. *18*, 182–192.

Nee, D.E., and Brown, J.W. (2012). Rostral-caudal gradients of abstraction revealed by multi-variate pattern analysis of working memory. Neuroimage *63*, 1285–1294.

Noonan, M.P., Walton, M.E., Behrens, T.E.J., Sallet, J., Buckley, M.J., and Rushworth, M.F.S. (2010). Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. Proc. Natl. Acad. Sci. USA *107*, 20547–20552.

Noonan, M.P., Mars, R.B., and Rushworth, M.F.S. (2011). Distinct roles of three frontal cortical areas in reward-guided behavior. J. Neurosci. *31*, 14399–14412.

Noonan, M.P., Kolling, N., Walton, M.E., and Rushworth, M.F.S. (2012). Re-evaluating the role of the orbitofrontal cortex in reward and reinforcement. Eur. J. Neurosci. *35*, 997–1010.

O'Doherty, J.P., Deichmann, R., Critchley, H.D., and Dolan, R.J. (2002). Neural responses during anticipation of a primary taste reward. Neuron *33*, 815–826.

Ostlund, S.B., and Balleine, B.W. (2007). Orbitofrontal cortex mediates outcome encoding in Pavlovian but not instrumental conditioning. J. Neurosci. *27*, 4819–4825.

Otto, A.R., Gureckis, T.M., Markman, A.B., and Love, B.C. (2009). Navigating through abstract decision spaces: evaluating the role of state generalization in a dynamic decision-making task. Psychon. Bull. Rev. *16*, 957–963.

Padoa-Schioppa, C. (2011). Neurobiology of economic choice: a good-based model. Annu. Rev. Neurosci. *34*, 333–359.

Padoa-Schioppa, C., and Assad, J.A. (2006). Neurons in the orbitofrontal cortex encode economic value. Nature *441*, 223–226.

Padoa-Schioppa, C., and Assad, J.A. (2008). The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. Nat. Neurosci. *11*, 95–102.

Petrides, M. (2000). Mapping prefrontal cortical systems for the control of cognition. In Brain mapping: the systems, A.W. Toga and J.C. Mazziotta, eds. (San Diego: Academic Press), pp. 159–176.

Pickens, C.L., Saddoris, M.P., Setlow, B., Gallagher, M., Holland, P.C., and Schoenbaum, G. (2003). Different roles for orbitofrontal cortex and basolateral amygdala in a reinforcer devaluation task. J. Neurosci. *23*, 11078–11084.

Preuss, T.M. (1995). Do rats have prefrontal cortex? The rose-woolsey-akert program reconsidered. J. Cogn. Neurosci. *7*, 1–24.

Price, J.L. (2007). Definition of the orbital cortex in relation to specific connections with limbic and visceral structures and other cortical regions. Ann. N Y Acad. Sci. *1121*, 54–71.

Ramus, S.J., and Eichenbaum, H. (2000). Neural correlates of olfactory recognition memory in the rat orbitofrontal cortex. J. Neurosci. *20*, 8199–8208.

Redish, A.D., Jensen, S., Johnson, A., and Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. Psychol. Rev. *114*, 784–805.

Rescorla, R.A., and Wagner, A.R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Classical conditioning II: Current research and theory, A.H. Black and W.F. Prokasy, eds. (New York: Appleton-Century-Crofts), pp. 64–99.

Ricker, S.T., and Bouton, M.E. (1996). Reacquisition following extinction in appetitive conditioning. Anim. Learn. Behav. *24*, 423–436.

Roesch, M.R., Taylor, A.R., and Schoenbaum, G. (2006). Encoding of time-discounted rewards in orbitofrontal cortex is independent of value representation. Neuron *51*, 509–520.

Rolls, E.T., Hornak, J., Wade, D., and McGrath, J. (1994). Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. J. Neurol. Neurosurg. Psychiatry *57*, 1518–1524.

Rudebeck, P.H., and Murray, E.A. (2011a). Balkanizing the primate orbitofrontal cortex: distinct subregions for comparing and contrasting values. Ann. N Y Acad. Sci. *1239*, 1–13.

Rudebeck, P.H., and Murray, E.A. (2011b). Dissociable effects of subtotal lesions within the macaque orbital prefrontal cortex on reward-guided behavior. J. Neurosci. *31*, 10569–10578.

Rushworth, M.F.S., Noonan, M.P., Boorman, E.D., Walton, M.E., and Behrens, T.E. (2011). Frontal cortex and reward-guided learning and decision-making. Neuron *70*, 1054–1069.

Schoenbaum, G., and Eichenbaum, H. (1995). Information coding in the rodent prefrontal cortex. I. Single-neuron activity in orbitofrontal cortex compared with that in pyriform cortex. J. Neurophysiol. *74*, 733–750.

Schoenbaum, G., Chiba, A.A., and Gallagher, M. (1999). Neural encoding in orbitofrontal cortex and basolateral amygdala during olfactory discrimination learning. J. Neurosci. *19*, 1876–1884.

Schoenbaum, G., Nugent, S.L., Saddoris, M.P., and Setlow, B. (2002). Orbitofrontal lesions in rats impair reversal but not acquisition of go, no-go odor discriminations. Neuroreport *13*, 885–890.

Schoenbaum, G., Setlow, B., Nugent, S.L., Saddoris, M.P., and Gallagher, M. (2003a). Lesions of orbitofrontal cortex and basolateral amygdala complex disrupt acquisition of odor-guided discriminations and reversals. Learn. Mem. *10*, 129–140.

Schoenbaum, G., Setlow, B., and Ramus, S.J. (2003b). A systems approach to orbitofrontal cortex function: recordings in rat orbitofrontal cortex reveal interactions with different learning systems. Behav. Brain Res. *146*, 19–29.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599.

Seo, H., Barraclough, D.J., and Lee, D. (2007). Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. Cereb. Cortex *17* (*Suppl 1*), i110–i117.

Simmons, J.M., and Richmond, B.J. (2008). Dynamic changes in representations of preceding and upcoming reward in monkey orbitofrontal cortex. Cereb. Cortex *18*, 93–103.

Sul, J.H., Kim, H., Huh, N., Lee, D., and Jung, M.W. (2010). Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. Neuron *66*, 449–460.

Sutton, R.S., and Barto, A.G. (1998). Reinforcement learning: An introduction. (Cambridge: MIT Press).

Takahashi, Y.K., Roesch, M.R., Stalnaker, T.A., Haney, R.Z., Calu, D.J., Taylor, A.R., Burke, K.A., and Schoenbaum, G. (2009). The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes. Neuron *62*, 269–280.

Takahashi, Y.K., Roesch, M.R., Wilson, R.C., Toreson, K., O'Donnell, P., Niv, Y., and Schoenbaum, G. (2011). Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. Nat. Neurosci. *14*, 1590–1597.

Teitelbaum, H. (1964). A comparison of effects of orbitofrontal and hippocampal lesions upon discrimination learning and reversal in the cat. Exp. Neurol. *9*, 452–462.

Thorpe, S.J., Rolls, E.T., and Maddison, S. (1983). The orbitofrontal cortex: neuronal activity in the behaving monkey. Exp. Brain Res. *49*, 93–115.

Tsuchida, A., Doll, B.B., and Fellows, L.K. (2010). Beyond reversal: a critical role for human orbitofrontal cortex in flexible learning from probabilistic feedback. J. Neurosci. *30*, 16868–16875.

Tsujimoto, S., Genovesio, A., and Wise, S.P. (2009). Monkey orbitofrontal cortex encodes response choices near feedback time. J. Neurosci. *29*, 2569–2574.

Tsujimoto, S., Genovesio, A., and Wise, S.P. (2011). Comparison of strategy signals in the dorsolateral and orbital prefrontal cortex. J. Neurosci. *31*, 4583–4592.

Wallis, J.D. (2007). Orbitofrontal cortex and its contribution to decision-making. Annu. Rev. Neurosci. *30*, 31–56.

Wallis, J.D. (2012). Cross-species studies of orbitofrontal cortex and value-based decision-making. Nat. Neurosci. *15*, 13–19.

Walton, M.E., Behrens, T.E.J., Buckley, M.J., Rudebeck, P.H., and Rushworth, M.F.S. (2010). Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. Neuron *65*, 927–939.

Wilson, R.C., and Niv, Y. (2011). Inferring relevance in a changing world. Front Hum Neurosci *5*, 189.

Wise, S.P. (2008). Forward frontal fields: phylogeny and fundamental function. Trends Neurosci. *31*, 599–608.

Neuron, Volume *81*

Supplemental Information

Orbitofrontal Cortex as a Cognitive Map of Task Space

Robert C. Wilson, Yuji K. Takahashi, Geoffrey Schoenbaum, and Yael Niv

Odor guided choice task

Lesions and recording

All experimental procedures were in accordance with the University of Maryland School of Medicine Animal Care and Use Committee and US National Institutes of Health guidelines. For full details of lesions see Takahashi et al. (2011). Recording electrodes were surgically implanted under stereotaxic guidance in the one hemisphere of VTA. Some rats (n = 7) also received neurotoxic lesions of ipsilateral OFC. Controls (n = 6) received sham lesions in which burr holes were drilled and the pipette tip lowered into the brain but no solution delivered. 30 neurons were recorded from the sham-lesioned group and 50 from the OFC-lesioned group.

Model

Prediction errors

The temporal difference prediction error, δ_t, at time t (Sutton & Barto, 1998) is given by

$$\delta_t = r_t + \gamma V(s_t) - V(s_{t-1}) \tag{1}$$

where r_t is the current reward, γ is the discount factor, s_t is the state at time t and $V(s_t)$ is the value of that state. This prediction error is used to update the state values of all eligible states according to

$$V_{new}(s) = V_{old}(s) + \alpha e(s)\delta_t \tag{2}$$

where α is the learning rate and $e(s)$ is the eligibility trace that determines which states are eligible for update. In particular, at each time step of the model, a new state is visited (and a reward potentially obtained), after which all values are updated as above, and then $e(s)$ are updated to reflect the new state visited, according to

$$e_{new}(s) = \begin{cases} 1 + e_{old}(s) & \text{if } s = s_t \\ \lambda e_{old}(s) & \text{otherwise} \end{cases} \tag{3}$$

In contrast to many RL models, here we are not interested in modeling the fine scale dynamics of the learning process, but rather, we focus on the overall change that occurs across each block. Assuming that α is sufficiently large to ensure that learning goes to completion, and given that we know the behavior (which is identical in both groups, see Takahashi et al., 2011), we can write down expressions for the steady state values in each block. These values can then be used to compute prediction errors both at the end of a block, when learning has gone to completion, and at the start of a block, before learning has significantly changed the values. In particular we can write the early and late prediction errors for moving into state s_t in block b as

$$\begin{aligned} \delta^b_{early} &= r^b_t + \gamma V^{b-1}(s_t) - V^{b-1}(s_{t-1}) \\ \delta^b_{late} &= r^b_t + \gamma V^b(s_t) - V^b(s_{t-1}) \end{aligned} \tag{4}$$

where $V^b(s_t)$ is the value of state s_t in block b, and $r^b(s_t)$ is the reward associated with moving into state s_t, in block b. The expressions for the average values, $V^b(s_t)$, are given below.

State Values

Assuming that learning goes to completion in each block, we can compute the average state values at the end of each block. This is done by considering the average of the update equation, equation 2, over all possible choices and odors, i.e.:

$$\langle V_{new}(s) \rangle = \langle V_{old}(s) \rangle + \alpha \langle e(s)\delta_t \rangle \tag{5}$$

where $\langle \cdot \rangle$ denotes taking the average. When learning has gone to completion in block b we have $\langle V_{new}(s) \rangle = \langle V_{old}(s) \rangle = V^b(s)$ and thus at long times ($t \to \infty$),

$$\langle e(s)\delta_t \rangle = 0 \tag{6}$$

This condition enables us to compute the fixed point values in each block. For the sham model, these values are, for the right reward arc,

$$
\begin{aligned}
V^b(\text{right rew } 3) &= 0 \\
V^b(\text{wait right}) &= \frac{1-p_{wait}}{1-\gamma p_{wait}} r_r^b(3) \\
V^b(\text{right rew } 2) &= \gamma V^b(\text{wait right}) \\
V^b(\text{right rew } 1) &= \gamma V^b(\text{right rew } 2) + r_r^b(2) \\
V^b(\text{right port}) &= \gamma V^b(\text{right rew } 1) + r_r^b(1)
\end{aligned}
\tag{7}
$$

where p_{wait} is the probability of staying in the wait states (wait right or wait left) and $r_r^b(i)$ is the amount of reward delivered in the right reward port in block b at the ith reward point, i = 1 for the short, small and first drop of the big reward, i = 2 for the second drop of the big reward, i = 3 for the long reward. Similarly for the left reward arc we have

$$
\begin{aligned}
V^b(\text{left rew } 3) &= 0 \\
V^b(\text{wait left}) &= \frac{1-p_{wait}}{1-\gamma p_{wait}} r_l^b(3) \\
V^b(\text{left rew } 2) &= \gamma V^b(\text{wait left}) \\
V^b(\text{left rew } 1) &= \gamma V^b(\text{left rew } 2) + r_l^b(2) \\
V^b(\text{left port}) &= \gamma V^b(\text{left rew } 1) + r_l^b(1)
\end{aligned}
\tag{8}
$$

where $r_l^b(i)$ is the reward in the left reward port. Finally at the odor and odor port,

$$
\begin{aligned}
V^b(\text{right}) &= p(\text{correct} \mid \text{forced}) \gamma V^b(\text{right port}) \\
V^b(\text{left}) &= p(\text{correct} \mid \text{forced}) \gamma V^b(\text{left port}) \\
V^b(\text{free}) &= p(\text{left} \mid \text{free}, b) \gamma V^b(\text{left port}) + p(\text{right} \mid \text{free}, b) \gamma V^b(\text{right port}) \\
V^b(\text{odor port}) &= \frac{\gamma}{3}\left[V^b(\text{left}) + V^b(\text{free}) + V^b(\text{right}) \right]
\end{aligned}
\tag{9}
$$

where $p(\text{correct} \mid \text{forced})$ is the probability of making the correct choice on a forced trial (e.g. going right on a forced right trial instead of left), $p(\text{left} \mid \text{free}, b)$ is the probability of going left on a free choice trial in block b and $p(\text{right} \mid \text{free}, b)$ is the probability of going right.

For the lesioned model, as the state representation does not correspond to the underlying structure of the task, we have to be more careful when taking the average. In particular,

we must consider the long-short and small-big cases separately as their paths through the state space are different.

The long-short condition is complicated because the model will spend different amounts of time in the port state depending on whether the reward is short or long, spending just one time step there in the short condition and multiple time steps there in the long condition.

$$
\begin{aligned}
V^b(\text{reward 2}) &= 0 \\
V^b(\text{reward 1}) &= 0 \\
V^b(\text{port}) &= \frac{\dfrac{p(\text{long})\lambda^3(1-p_{wait})}{1-\lambda p_{wait}} + p(\text{short})}{p(\text{short}) + p(\text{long})\left((1-\gamma)(1+\lambda+\lambda^2) + \dfrac{\lambda^3(1-p_{wait})}{1-\lambda p_{wait}}\right)}
\end{aligned}
\tag{10}
$$

where $p(\text{short})$ is the average probability of encountering the short option,

$$
p(\text{short}) = \frac{1}{3}\big(p(\text{correct} \mid \text{forced}) + p(\text{short} \mid \text{free}, b)\big) \tag{11}
$$

Note that $p(\text{short} \mid \text{free}, b) = p(\text{right} \mid \text{free}, b)$ when the short reward is on the right and vice versa for the left. $p(\text{long})$ is the average probability of encountering the long option

$$
p(\text{long}) = \frac{1}{3}\big(p(\text{correct} \mid \text{forced}) + p(\text{long} \mid \text{free}, b)\big) \tag{12}
$$

where, similarly, $p(\text{long} \mid \text{free}, b) = p(\text{left} \mid \text{free}, b)$ when the long reward is on the left.

At the odor and odor port, we have

$$
\begin{aligned}
V^b(\text{short}) &= p(\text{correct}\,|\,\text{forced})\Big((\gamma-\lambda)V^b(\text{port})+\lambda\Big)\\
V^b(\text{long}) &= p(\text{correct}\,|\,\text{forced})\times\\
&\left(V^b(\text{port})(\gamma+(\gamma-1)\lambda(1+\lambda+\lambda^2)+\lambda^4\left(\gamma V^b(\text{reward 1})+1-V^b(\text{port})\right)\frac{1-p_{wait}}{1-\lambda p_{wait}}\right) \quad (13)\\
V^b(\text{free}) &= \frac{p(\text{short}\,|\,\text{free})}{p(\text{correct}\,|\,\text{forced})}V^b(\text{short})+\frac{p(\text{long}\,|\,\text{free})}{p(\text{correct}\,|\,\text{forced})}V^b(\text{long})\\
V^b(\text{odor}) &= \frac{\gamma}{3}(V^b(\text{left})+V^b(\text{free})+V^b(\text{right}))
\end{aligned}
$$

where we have introduced $V^b(\text{short})$ and $V^b(\text{long})$ to denote the value of the odor corresponding to forced trials to the short and long rewards respectively. Note that $V^b(\text{long})=V^b(\text{left})$ when the long reward is on the left.

The small-big condition is more straightforward as, regardless of the trial type, the model only spends one time step in the port state. Thus we have

$$
\begin{aligned}
V^b(\text{reward 2}) &= 0\\
V^b(\text{reward 1}) &= \frac{p(l)r_l(2)+p(r)r_r(2)}{p(l)+p(r)}\\
V^b(\text{port}) &= 1+\gamma V^b(\text{reward 1})\\
V^b(\text{right}) &= p(\text{correct}\,|\,\text{forced})\Big((\gamma-\lambda)V^b(\text{port})+\lambda(\gamma-\lambda)V^b(\text{reward 1})+\lambda r_r(1)+\lambda^2 r_r(2)\Big) \quad (14)\\
V^b(\text{left}) &= p(\text{correct}\,|\,\text{forced})\Big((\gamma-\lambda)V^b(\text{port})+\lambda(\gamma-\lambda)V^b(\text{reward 1})+\lambda r_l(1)+\lambda^2 r_l(2)\Big)\\
V^b(\text{free}) &= \frac{p(\text{right}\,|\,\text{free})}{p(\text{correct}\,|\,\text{forced})}V^b(\text{right})+\frac{p(\text{left}\,|\,\text{free})}{p(\text{correct}\,|\,\text{forced})}V^b(\text{left})\\
V^b(\text{odor}) &= \frac{\gamma}{3}(V^b(\text{left})+V^b(\text{free})+V^b(\text{right}))
\end{aligned}
$$

Converting Prediction Errors to Firing Rate in VTA

To convert the computed prediction errors into neural firing rates, f, we used a simple linear transformation

$$f = B + k\delta \quad (15)$$

where B is the baseline firing rate and k is the scale factor. Based on experimental findings (Schultz et al., 1997) and in line with previous work (Niv, Duff, & Dayan, 2005), we used a different scale factor for positive prediction errors (PPEs) and negative prediction errors (NPEs).

Fitting Model Parameters to Neural Data

To set the model's free parameters, we fit the model to the dopaminergic firing data at the time of the reward and the time of the odor. The model had six free parameters: the discount factor γ, the eligibility trace decay rate λ, the baseline firing rate before an odor cue (B in equation 15), the baseline firing rate before the reward, the scaling of positive prediction errors (k_+ in equation 15) and the scaling for negative prediction errors (k_- in equation 15). We fit the sham-lesioned and OFC-lesioned groups separately. To fit the parameters, we minimized the mean squared error between the model's predicted firing rates and the average firing rates measured in the experiment. The best fit values of the parameters were:

Parameter	Sham lesion	OFC lesion
Discount factor, γ	0.89	0.47
Eligibility trace decay, λ	1.0	1.0
Baseline before odor (Hz)	5.3	5.6
Baseline before reward (Hz)	3.7	4.2
PPE scaling, k_+	1.6	1.7
NPE scaling, k_-	2.0	1.6

Fit parameter values for the sham- and OFC-lesioned animals

Interestingly, for the most part, the parameter values were similar between the sham- and OFC-lesioned groups, suggesting that indeed basic reinforcement learning processes are intact in OFC-lesioned rats. The exception to this is the discount factor, which was reduced in the OFC-lesioned group. That is, the best-fit parameters suggest that the OFC-lesioned group discounts rewards more steeply than the sham-lesioned controls. One interpretation of this 'steeper discounting', is that due to the inaccurate (and specifically, non-Markov) state space representation of the task in the OFC-lesioned model, reward

information cannot 'travel back' reliably to reward predictive states. That is, the existence of aliased states that represent more than one state in the true state space of the task (for instance, the 'reward port' state), means that information about rewards cannot effectively be attributed to early states that lead to that reward. This effect could disguise itself as heavier discounting of rewards, as those future rewards do not exert their full effect on earlier state values. Moreover, this result is consistent with there being a smaller difference between the dopaminergic response to the forced high and forced low odors in the OFC-lesioned group, as compared to the sham-lesioned group.

Extra Drop for the Big Reward Option in Fourth Block

In figure 6 in the main text we discussed the effect of a third, discretionary drop of juice that was occasionally applied for the big reward in the fourth block. Unfortunately, the delivery of this drop was not recorded in the dataset and so it is impossible to determine at this point on what trials it was delivered. However, an echo of this extra reward can be seen in firing patterns of a subset of our VTA neurons.

To illustrate this, Figure S1 shows the firing of one of these neurons when the animal is at the high valued reward port in the fourth block. This neuron clearly shows elevated firing following the first two rewards (presented at 0.5 and 1 seconds following reward port entry). It also shows a clear elevation in firing rate at around 1.7 seconds indicative of an extra drop of juice presented at 1.5 seconds.

Similar visual examination of the firing of individual neurons suggests that at least 14 neurons were recorded in the presence of a third drop. For 13 of these neurons (as in our example neuron) the third drop appears from the start of the fourth block.

Firing at the Time of Reward

Figure 6 in the main text shows the firing of VTA neurons at the time of unexpected rewards. These rewards are unexpected at the start of a block, after reward contingencies have changed unexpectedly, but given learning with the correct state representation, should be predicted by the end of the block. Thus we compared the first two ('early') trials to the last five ('late') trials of a block to test for effects of learning. To maximize

statistical power, in Figure 6A,B we included both forced and free choice trials, as RL theory suggests that reward expectations should be similar regardless of the type of trial. Indeed, there were no qualitative differences in the results if only forced trials were considered (since the animals did not choose the low option often enough, we could not analyze free choice trials alone). In the model (Figure 6C-E), assuming that learning goes to completion in each block, we could compute these early and late prediction errors analytically, in a way that is robust to the choice of learning rate parameters

Firing at the Time of the Odor

In addition to the time of the reward, which we concentrated on in the main text, we can expect to see prediction error signals at the time of odor presentation. This is because the identity of the odor can change the rats' expectation of future reward. For example, smelling the 'forced high' odor can lead the rat to expect a higher than average reward, and thus cause a positive prediction error. Conversely, the 'forced low' odor leads to the expectation of a lower than average reward and should cause a negative prediction error.

In Supplementary Figure 2 we show the predicted and actual firing rates at the time of the odor, on forced trials. In particular we focus on the last 5 trials of each block when we assume that learning has gone to completion. The sham-lesion data (Supplementary Figure 2A) show increased firing when the rat expects a high-valued reward (e.g. forced left when the high reward is on the left), and lower firing rates when it expects the less-valued reward. These differences are significant for the 1st (short vs long), 2nd (long vs short) and 4th (small vs big) blocks while there is a trend in the same direction ($p = 0.11$) for the 3rd block (big vs small). These firing patterns are in line with standard RL accounts of dopaminergic firing, and our model is in close agreement with the data (Supplementary Figure 2C).

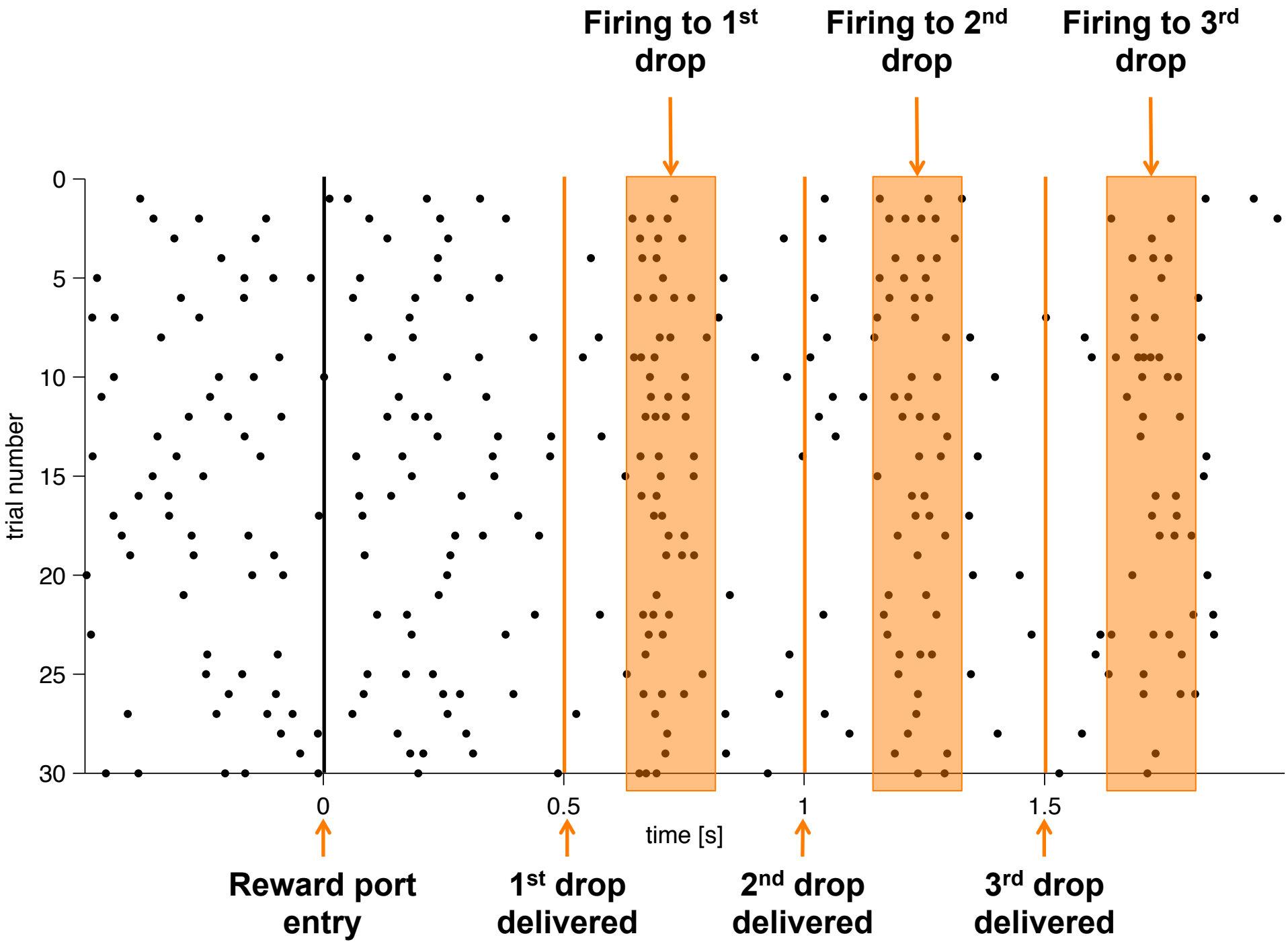
Intriguingly, data from OFC-lesioned rats (Supplementary Figure 2B) show a similar pattern of firing, even though the firing at the time of reward in this group was quite different from that of control rats, as discussed above. This is especially interesting for block 2 (long vs short): at the time of the odor there is a significant difference in firing between the odor cue predicting the short reward (forced high) and that predicting the

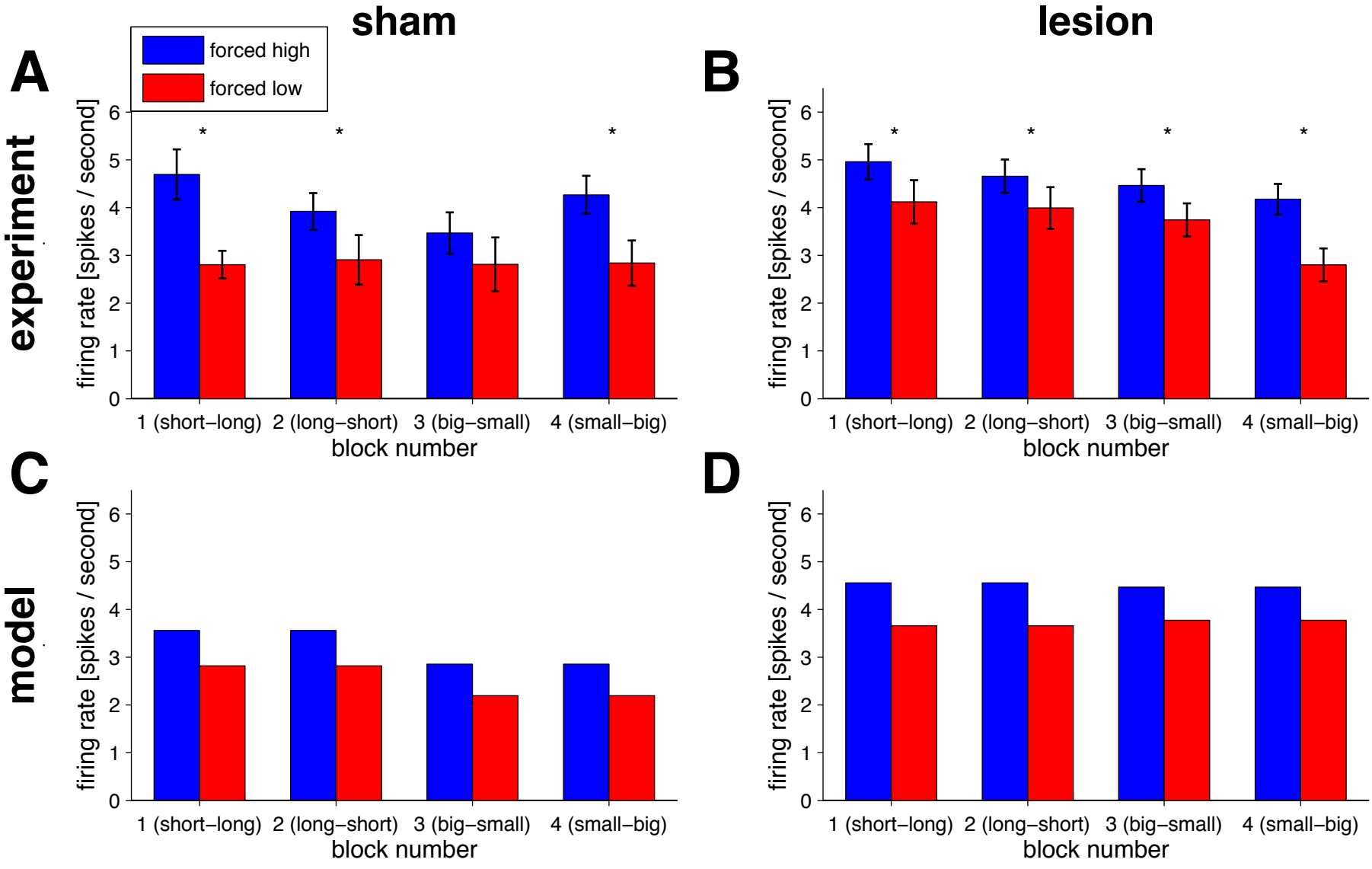
long reward (forced low), despite the fact that firing at the time of reward does not indicate accurate expectations of reward (Figure 6B). Thus, the predicted values at the time of the odor cues seem inconsistent with those at the time of reward. However, this seeming inconsistency is predicted by our model: because learning in the model uses eligibility traces, when a forced left odor is presented, for example, the model enters the left state and this state becomes eligible for update. When the animal then encounters a reward later on in the trial, the ensuing prediction error is used to update the value of the left state. Thus over the course of learning the animal learns to associate the short reward with forced left odor and the long reward with the forced right odor and the pattern of prediction errors at the time of the odor (Supplementary Figure 2D) is preserved in the lesioned model. This is despite the fact that the model cannot learn differential predictions for the left and right port states because the food port states are shared between both trial types, and thus the change from block 1 to 2 has no consequences for the average prediction errors at the time of reward (Figure 6B).

Supplemental Figures

Figure S1 – Firing in an example neuron from an OFC-lesioned rat evidencing the extra (third) reward drop in the fourth block. Trial 1 marks the first trial of the fourth block. Evidence of the manually-delivered reward is apparent throughout the whole block, though towards the end of the block learning seems to have attenuated the prediction error response to this drop.

Figure S2 – Measured and predicted dopaminergic firing at the time of the odor on forced trials in Takahashi et al. (2011). (A, B) Experimental data correspond to the average firing rate in the 500ms after the odor. Blue: trials in which the better of the two possible outcomes in this block was expected; Red: trials in which the worse of the two outcomes was expected. Stars denote significance at $p < 0.05$. (C, D) In the model, the time of the odor corresponds to entering either the left, free or right states. The left plots (A, C) show data from the sham-lesioned group and the full state space (see Figure 5B), and in the right plots (B, D) are data from the OFC-lesioned group and the reduced state space (Figure 5C).


Firing to 1st drop
Firing to 2nd drop
Firing to 3rd drop
trial number
0
5
10
15
20
25
30
0
0.5
1
1.5
time [s]
Reward port entry
1st drop delivered
2nd drop delivered
3rd drop delivered


A
sham
lesion
B
experiment
forced high
forced low
firing rate [spikes / second]
0
1
2
3
4
5
6
*
1 (short–long)
2 (long–short)
3 (big–small)
4 (small–big)
block number
C
D
model

Supplemental References

Niv, Y., Duff, M. O., & Dayan, P. (2005). Dopamine, uncertainty and TD learning. Behav Brain Funct, 1, 6.

Schultz, W., Dayan, P., & Montague, P.R. (1997). A neural substrate of prediction and reward. Science 275, 1593-1599.

Sutton, R.S., & Barto, A.G. (1998). Reinforcement learning: An introduction. MIT Press.

Takahashi, Y.K., Roesch, M.R., Wilson, R.C., Toreson, K., O'Donnell, P., Niv, Y., Schoenbaum, G. (2011). Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex. Nat Neuro 14(12), 1590-1597.