Published in final edited form as:
Neuron. 2014 January 22; 81(2): 267–279. doi:10.1016/j.neuron.2013.11.005.

Orbitofrontal cortex as a cognitive map of task space

Robert C. Wilson[1], **Yuji K. Takahashi**[2], **G. Schoenbaum**[#2,3], and **Yael Niv**[#1]

[1]Department of Psychology and Neuroscience Institute, Princeton University

[2]Department of Anatomy & Neurobiology, University of Maryland School of Medicine

[3]Department of Psychiatry, University of Maryland School of Medicine

[#] These authors contributed equally to this work.

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 new 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 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 a reinforcement learning 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, albeit 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 & 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 then aid decision making in the service of harvesting rewards 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 & Niv, 2010, 2013; Gershman et al., 2010; Wilson & 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 to 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 likely also exist elsewhere in the brain, 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 complete loss of state information, but rather that perceptually similar states can no longer be distinguished – an OFC-lesioned 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 their states partially observable.

Results

Here we show how our theory can account for a number of experimental findings. We first consider the archetypal ‘OFC task’ of reversal learning, as well as delayed alternation, extinction and devaluation, before turning to neural findings that reveal more directly 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 & Mishkin, 1972; Rolls et al., 1994; Dias et al., 1996; Meunier et al., 1997; McAlonan & Brown, 2003; Schoenbaum et al., 2002, 2003a; Chudasama & Robbins, 2003; Bohn et al., 2003; Izquierdo et al., 2004; Kim & Ragozzino, 2005). Here we illustrate our theory through 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 either on their left or on their right to receive food reward. Only one location was rewarded in each block, its identity reversed once the monkey reached a criterion of 90% correct. Reward contingencies were reversed five times. Figure 1A summarizes the results: while initial learning was spared, OFC-lesioned animals (grey) were impaired on reversals relative to sham-lesioned controls (orange).

To model behavior in this task we used a simple Q-learning algorithm (Sutton & 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

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

where α is a learning rate parameter and $[r_{t+1} - Q_{old}(a_t, s_t)]$ is the prediction error[1]. Using the learned values, the probability of taking action a in state s_t is given by the softmax or Luce rule:

$$p(a|s_t) = \frac{exp(\beta Q(a, s_t))}{\sum_{a'} exp(\beta Q(a', s_t))} \quad (2)$$

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 using 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 using 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. Following an action, 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 since a 'right' choice from this state is

[1]We omit the value of the subsequent state in equation 2 (cf. Sutton & Barto, 1998) as in this task trials involves one state with no sequential contingencies. This renders our learning rule identical to Rescorla and Wagner (1972).

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, as the action propensities learned in state 1 are left unaltered. If rewards and choices are deterministic, it is easy to see that the model will only take one trial to reverse its behavior after such a re-reversal. 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 'un-learning' of the preference that was acquired in the previous 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 & Orbach, 1972; Butters et al., 1973; Mishkin & 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 5s delay between trials, required memory of the last action. Control animals learned this task easily, ultimately performing at around 90% correct. Monkeys with OFC lesions, however, failed to perform better than chance even after 2000 trials of training (figure 2A).

We modeled behavior of control animals using the state representation in figure 2C, in which the current state is determined by the choice on the last trial (option A or option B). With this state representation, the model learns the task easily (figure 2B) with performance limited only 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 & 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 post-extinction 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 using 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 minutes of reinforced pressing, an extinction phase began: rewards were no longer available and extinction of responding was measured as the number of presses in successive ten-minute blocks. The results, shown in figure 3A, clearly demonstrate slower extinction for OFC-lesioned animals.

As previously, we modeled control animals (figure 3C) using a two-state model: the animal is in state ‘P1’ if the previous leverpress 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 one unit of reward during conditioning and −0.2 units in extinction, to represent 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, as extinction transitioned the animal into the P0 state where new action values for P and N were learned (starting from low initial values). The one-state model of the OFC-lesioned animals, on the other hand, 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 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).

Post-extinction predictions—To assess the effectiveness of extinction and to investigate what was learned during extinction, researchers often re-test 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 then 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 the longer the waiting time 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 then selected at random (for instance, with probability proportional to the time since it last occurred), 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 does truly 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 & Bouton, 1996), in which re-conditioning of a stimulus → outcome association occurs more rapidly after extinction than in the original learning. The two-state model predicts this phenomenon because re-conditioning 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 indeed 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 using 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 & Rescorla, 1985; Balleine & 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 the OFC. Consistent with this proposal, OFC lesions impair

performance in devaluation experiments, causing lesioned animals to respond equally to devalued and non-devalued 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 to associate a light cue with food. Subsequently, the food was devalued by pairing its consumption to injection of lithium chloride. Then, a testing session measured the amount of time spent at the food cup when the light was presented. 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, as the model transitioned from state s to state s' it computed a prediction error

$$\delta = r + V_{MF}(s') - V_{MF}(s) \quad (3)$$

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

$$V_{MF}(s) \leftarrow V_{MF}(s) + \alpha\delta \quad (4)$$

where $\alpha = 0.1$ 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 (as it only ever experiences the light paired with food) and, in the devaluation stage, a negative value for the food state. The model-based (MB) component, in contrast, uses the low value of the food state to update, even absent direct experience, the value of the light state through imagined simulation:

$$V_{MB}(\text{light}) = V_{MF}(\text{food})\, p(\text{food}|\text{light}) \quad (5)$$

where V_{MB}(light) is the model-based value of the light, V_{MF}(food) is the model-free value of the food state and p(food | 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 then a combination of the model-based and model-free values as in (Daw et al., 2005),

$$V(\text{light}) = \zeta V_{MB}(\text{light}) + (1-\zeta)\, V_{MF}(\text{light}) \quad (6)$$

where we used $\zeta = 0.2$ 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). The OFC-lesioned model, however, lacks model-based planning abilities ($\zeta = 0$) 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 over-expectation (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 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, the changes in neural firing patterns should be consistent with the loss of non-stimulus-bound states, but 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 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 is shown in figure 5A: In the 1st block of trials one well paid out one drop of juice after a short delay while 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 (2 drops of juice) and a small reward (1 drop of juice) and these contingencies reversed again in the fourth and final block of the session. The experiment 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). An odor is then presented signaling a forced left ('left' state), free choice ('free') or forced right ('right') trial. On forced right trials or free choice trials, if the rat chooses to go to the right fluid well, 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, 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. We thus 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, as 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 were compromised. We also assume that temporal information is not available externally, and thus OFC-lesioned rats cannot distinguish reward states that are separated only by the passage of time. Together, 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 as 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 VTA (Schultz et al., 1997). These signals convey the difference between predicted and actual outcomes (Sutton & Barto, 1998; see Supplementary Material 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, compared to the prediction before odor onset). Indeed, although behavior in both groups was equated due to 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 Supplementary Material.

Figure 6 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 (see Supplementary Material 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 rewards 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 rewards on a seemingly random half of the trials, and delayed rewards on the other half. The model does, however, 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 2 to 3, 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.'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. While the timing of this manually applied third drop was not recorded, examination of the 14 spike raster plots in which the response of individual neurons to each drop is clearly visible (for an example see supplementary figure 1) 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). A prediction of the model, then, 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 not change throughout a block, as 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 (supplementary figure 2). 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 orbitofrontal cortex 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 post-extinction phenomena in animals with OFC lesions. We now 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. The question thus 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.
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, while 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 & Eichenbaum, 1995; Sul et al., 2010; the latter also found these variables in dlPFC and ACC). In a probabilistic reversal-learning task, Hampton and colleagues (Hampton, Bossaerts, & O'Doherty, 2006) showed that BOLD 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 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 & Eichenbaum, 2000; Lara et al., 2009; the latter study is especially interesting as 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 & Brown, 2012).

Specificity—Addressing the specificity condition is more difficult as it is 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 Eichenbaum and colleagues (Schoenbaum & Eichenbaum, 1995; Ramus & Eichenbaum, 2000; reviewed in Schoenbaum et al., 2003b). In the first task, 4 of 8 odors predicted that a response at a nearby fluid well would be rewarded. In the second task, 8 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, but in the first task the identity of the odor was critical for predicting reward, while 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 non-match.

Simmons et al. (2008) also demonstrated that small changes in a task can cause significant changes to OFC representations. In their task, monkeys were rewarded after 1, 2 or 3 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, while in a ‘random cue’ condition there was no relation between background color and number of trials to reward. As a result, only in the random cue condition the outcome of the previous trial was informative for reward prediction, as after a rewarded trial the next trial would be rewarded only on one third of the cases (a 1-correct trial requirement), while after an unrewarded trial the next trial would be rewarded on one half of the cases (a 2-correct or a 3-correct 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%). We further 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 & 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 & Murray, 2011a, 2011b; Noonan, Kolling, Walton, & Rushworth, 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, and thus more work is needed in order to precisely localize the representation of task states within OFC subregions.

Interspecies differences in the OFC

We have not distinguished between rats and monkeys, treating what is defined as 'the 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, with OFC in rats having 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 analogues in the rat (Wise, 2008; but see Preuss, 1995).

Interestingly, despite these differences, there are strong inter-species similarities at the level of connectivity (Carmichael & Price, 1996; Price, 2007), neural activity, and function. This is particularly true for the 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 & Mishkin, 1972; Rolls et al., 1994; Dias et al., 1996; Meunier et al., 1997; McAlonan & Brown, 2003; Schoenbaum et al., 2002, 2003a; Chudasama & Robbins, 2003; Bohn et al., 2003; Izquierdo et al., 2004; Kim & Ragozzino, 2005), extinction (Butter, 1969; McEnaney & 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 & Eichenbaum 1995; Critchley & Rolls, 1996a, 1996b; Schoenbaum et al., 1999; Gottfried et al., 2003; O'Doherty et al., 2002; Morrison & Salzmann, 2009). We suggest that OFC encodes the current task state in all of these species, with animals such as rodents perhaps being limited in the complexity of the state that can be represented in their relatively small OFC, while humans, who have a much more developed OFC, being 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 (SNc), where reward input from areas such as the lateral habenula, hypothalamus and the 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 & Barto, 1998; Daw et al., 2005), which makes use of learned relationships between states 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 post-extinction renewal (Ji & Maren, 2007). However, this is true only when states are defined according to spatial location. Hippocampal lesions seem to have no effect on non-spatial reversal learning, while OFC lesions generally affect all types of reversal (Teitelbaum, 1964).

Seo and colleagues (2007) proposed that the dorsolateral prefrontal cortex (dlPFC) encodes task states, based on neural recordings that showed that choices, stimuli and rewards were encoded in dlPFC neurons. Indeed it seems clear that dlPFC satisfies the representation condition, however, this area is less able to satisfy the specificity condition as 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 using the variables found to be most relevant to the current task (Otto et al, 2009).

There is also a large literature on rule-based behavior that does not explicitly mention state representations but is clearly related. 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), while tasks requiring non-trivial assignment of rewards in a complex state space elicit activations in the lateral OFC (Noonan et al. 2011). Further, Buckley and colleagues (2009) found differential effects of lesions to the OFC and the dlPFC in a monkey analogue of the Wisconsin card sorting task: OFC lesions diminished monkeys' ability to learn new reward associations, consistent with an impaired representation of state, while dlPFC lesions decreased the ability to use a previously learned rule.

Finally, one might argue that 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, 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, to compute the current state, due to 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

hippocampus and has strong reciprocal connections with subcortical regions such as striatum and amygdala that are critical to the representation of reward (Carmichael & Price 1995a, Carmichael & Price 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 & 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, Holland, & Gallagher, 2008), satiety (Araujo et al., 2006), uncertainty (Kepecs et al., 2008) and taste (Padoa-Schioppa & 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, while 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, as 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 as it points to a key difference between the theories: our theory predicts that in tasks in which stimulus-bound states suffice, OFC lesions will not appear to cause a deficit in credit assignment. 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Araujo IE, de Gutierrez R, Oliveira-Maia AJ, Pereira A Jr, Nicolelis MAL, Simon SA. Neural ensemble coding of satiety states. Neuron. 2006; 51(4):483–94. [PubMed: 16908413]

Balleine BW, Dickinson A. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. Neuropharmacology. 1998; 37:407–419. [PubMed: 9704982]

Banich MT, Milham MP, Atchley RA, Cohen NJ, Webb A, Wszalek T, Kramer AF, Liang Z, Barad V, Gullett D, Shah C, Brown C. The prefrontal regions play a predominant role in imposing an attentional 'set': evidence from fMRI. Cogn Brain Res. 2000; 10:1–9.

Bohn I, Giertler C, Hauber W. Orbital prefrontal cortex and guidance of instrumental behaviour in rats under reversal conditions. Behav Brain Res. 2003; 143:49–56. [PubMed: 12842295]

Bouton ME. Context and behavioral processes in extinction. Learn Mem. 2004; 11:485–94. [PubMed: 15466298]

Brass, M.; Derrfuss, J.; Cramon, Y. The Role of the Posterior Frontolateral Cortex in Task Related Control.. In: Bunge, SA.; Wallis, JD., editors. Neuroscience of rule-guided behavior. Oxford University Press; 2008.

Brown TI, Ross RS, Keller JB, Hasselmo ME, Stern CE. Which way was I going? Contextual retrieval supports the disambiguation of well learned overlapping navigational routes. J Neurosci. 2010; 30:7414–22. [PubMed: 20505108]

Buckley MJ, Mansouri FA, Hoda H, Mahboubi M, Browning PGF, Kwok SC, Phillips A, Tanaka K. Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. Science. 2009; 325:52–58. [PubMed: 19574382]

Butter CM. Perseveration in extinction and in discrimination reversal tasks following selective frontal ablations in macaca mulatta. Physiol Behav. 1969; 4:163–171.

Butter CM, Mishkin M, Rosvold HE. Conditioning and extinction of a food rewarded response after selective ablations of frontal cortex in rhesus monkeys. Exp Neurol. 1963; 7:65–75. [PubMed: 14017412]

Butters N, Butter C, Rosen J, Stein D. Behavioral effects of sequential and one-stage ablations of orbital prefrontal cortex in the monkey. Exp Neurol. 1973; 39:204–14. [PubMed: 4634005]

Carmichael ST, Price JL. Limbic connections of the orbital and medial prefrontal cortex in macaque monkeys. J Comp Neurol. 1995a; 363:615–41. [PubMed: 8847421]

Carmichael ST, Price JL. Sensory and premotor connections of the orbital and medial prefrontal cortex of macaque monkeys. J Comp Neurol. 1995b; 363:642–64. [PubMed: 8847422]

Carmichael ST, Price JL. Connectional networks within the orbital and medial prefrontal cortex of macaque monkeys. J Comp Neurol. 1996; 371:179–207. [PubMed: 8835726]

Chudasama Y, Robbins TW. 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. 2003; 23:8771–8780. [PubMed: 14507977]

Colwill RM, Rescorla RA. Postconditioning devaluation of a reinforcer affects instrumental responding. J Exp Psych. 1985; 11:120–132.

Critchley HD, Rolls ET. Hunger and satiety modify the responses of olfactory and visual neurons in the primate orbitofrontal cortex. J Neurophysiol. 1996a; 75:1673–1686. [PubMed: 8727405]

Critchley HD, Rolls ET. Olfactory neuronal responses in the primate orbitofrontal cortex: analysis in an olfactory discrimination task. J Neurophysiol. 1996b; 75:1659–1672. [PubMed: 8727404]

Croxson PL, Johansen-Berg H, Behrens TEJ, Robson MD, Pinsk MA, Gross CG, Richter W, Richter C, Kastner S, Rushworth MF. Quantitative investigation of connections of the prefrontal cortex in the human and macaque using probabilistic diffusion tractography. J Neurosci. 2005; 25:8854–66. [PubMed: 16192375]

Damasio, AR. Descartes' error: Emotion, reason, and the human brain. Putnam; New York: 1994.

Daw ND, Niv Y, Dayan P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci. 2005; 8:1704–11. [PubMed: 16286932]

Dias R, Robbins TW, Roberts AC. Dissociation in prefrontal cortex of affective and attentional shifts. Nature. 1996; 380:69–72. [PubMed: 8598908]

Feierstein CE, Quirk MC, Uchida N, Sosulski DL, Mainen ZF. Representation of spatial goals in rat orbitofrontal cortex. Neuron. 2006; 51:495–507. [PubMed: 16908414]

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

Furuyashiki T, Holland PC, Gallagher M. Rat orbitofrontal cortex separately encodes response and outcome information during performance of goal-directed behavior. J Neurosci. 2008; 28:5127–38. [PubMed: 18463266]

Fuster, JM. The prefrontal cortex. Lippin-Ravencott; New York: 1997.

Gallagher M, McMahan RW, Schoenbaum G. Orbitofrontal cortex and representation of incentive value in associative learning. J Neurosci. 1999; 19:6610–6614. [PubMed: 10414988]

Gershman SJ, Blei DM, Niv Y. Context, learning, and extinction. Psych Rev. 2010; 117:197–209.

Gershman SJ, Niv Y. Learning latent structure: carving nature at its joints. Curr Opin Neurobiol. 2010; 20:251–6. [PubMed: 20227271]

Gershman SJ, Niv Y. Perceptual estimation obeys Occam's razor. Front Psychol. 2013; 4:623. [PubMed: 24137136]

Gottfried JA, O'Doherty J, Dolan RJ. Encoding predictive reward value in human amygdala and orbitofrontal cortex. Science. 2003; 301:1104–1107. [PubMed: 12934011]

Hampton AN, Bossaerts P, O'Doherty JP. The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. J Neurosci. 2006; 26:8360–8367. [PubMed: 16899731]

Izquierdo A, Suda RK, Murray EA. Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. J Neurosci. 2004; 24:7540–8. [PubMed: 15329401]

Ji J, Maren S. Hippocampal involvement in contextual modulation of fear extinction. Hippocampus. 2007; 17:749–58. [PubMed: 17604353]

Jones B, Mishkin M. Limbic lesions and the problem of stimulus–reinforcement associations. Exp Neurol. 1972; 36:362–77. [PubMed: 4626489]

Jones JL, Esber GR, McDannald MA, Gruber AJ, Alex Hernandez AM, Schoenbaum G. Orbitofrontal cortex supports behavior and learning using inferred but not cached values. Science. 2012; 338:953–956. [PubMed: 23162000]

Kaelbling LP, Littman MI, Cassandra AR. Planning and acting in partially observable stochastic domains. Artif Intell. 1998; 101:99–134.

Kahnt T, Chang LJ, Park SQ, Heinzle J, Haynes J-D. Connectivity-based parcellation of the human orbitofrontal cortex. J Neurosci. 2012; 32:6240–50. [PubMed: 22553030]

Kepecs A, Uchida N, Zariwala H, Mainen ZF. Neural correlates, computation and behavioural impact of decision confidence. Nature. 2008; 455(7210):227–31. [PubMed: 18690210]

Kim J, Ragozzino ME. The involvement of the orbitofrontal cortex in learning under changing task contingencies. Neurobiol Learn Mem. 2005; 8:125–133. [PubMed: 15721796]

Lara AH, Kennerley SW, Wallis JD. Encoding of gustatory working memory by orbitofrontal neurons. J Neurosci. 2009; 29:765–774. [PubMed: 19158302]

MacDonald AW III, Cohen JD, Stenger VA, Carter CS. Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. Science. 2000; 288:1835–1838. [PubMed: 10846167]

McAlonan K, Brown VJ. Orbital prefrontal cortex mediates reversal learning and not attentional set shifting in the rat. Behav Brain Res. 2003; 146:97–103. [PubMed: 14643463]

McDannald MA, Lucantonio F, Burke KA, Niv Y, Schoenbaum G. Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. J Neurosci. 2011; 31:2700–5. [PubMed: 21325538]

McEnaney KW, Butter CM. Perseveration of responding and nonresponding in monkeys with orbital frontal ablations. J Comp Phys Psych. 1969; 4:558–561.

Meunier M, Bachevalier J, Mishkin M. Effects of orbital frontal and anterior cingulate lesions on object and spatial memory in rhesus monkeys. Neuropsychologia. 1997; 35:999–1015. [PubMed: 9226661]

Miller MH, Orbach J. Retention of spatial alternation following frontal lobe resections in stump-tailed macaques. Neuropsychologia. 1972; 10:291–298. [PubMed: 4628085]

Mishkin M, Manning FJ. Non-spatial memory after selective prefrontal lesions in monkeys. Brain Res. 1978; 143:313–23. [PubMed: 415803]

Mishkin M, Vest B, Waxler M, Rosvold HE. A re-examination of the effects of frontal lesions on object alternation. Neuropsychologia. 1969; 7:357–363.

Morris G, Arkadir D, Nevet A, Vaadia E, Bergman H. Coincident but distinct messages of midbrain dopamine and striatal tonically active neurons. Neuron. 2004; 43:133–143. [PubMed: 15233923]

Morrison SE, Salzman CD. The convergence of information about rewarding and aversive stimuli in single neurons. J Neurosci. 2009; 29:11471–11483. [PubMed: 19759296]

Murray EA, O'Doherty JP, Schoenbaum G. What we know and do not know about the functions of the orbitofrontal cortex after 20 years of cross-species studies. J Neurosci. 2007; 27:8166–9. [PubMed: 17670960]

Murray, EA.; Wise, SP.; Rhodes, SEV. What Can Different Brains Do with Reward?. In: Gottfried, JA., editor. Neurobiology of sensation and reward. CRC Press; 2011.

Napier RM, Macrae M, Kehoe EJ. Rapid reacquisition in conditioning of the rabbits nictitating membrane response. J Exp Psych. 1992; 18:182–192.

Nee DE, Brown JW. Rostral–caudal gradients of abstraction revealed by multi-variate pattern analysis of working memory. Neuroimage. 2012; 63:1285–1294. [PubMed: 22992491]

Noonan MP, Kolling N, Walton ME, Rushworth MFS. Re-evaluating the role of the orbitofrontal cortex in reward and reinforcement. Eur J Neurosci. 2012; 35:997–1010. [PubMed: 22487031]

Noonan MP, Mars RB, Rushworth MFS. Distinct roles of three frontal cortical areas in reward-guided behavior. J Neurosci. 2011; 31:14399–412. [PubMed: 21976525]

Noonan MP, Walton ME, Behrens TEJ, Sallet J, Buckley MJ, Rushworth MFS. Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. P Natl Acad Sci USA. 2010; 107:20547–52.

O'Doherty J, Deichmann R, Critchley HD, Dolan RJ. Neural responses during anticipation of a primary taste reward. Neuron. 2002; 33:815–826. [PubMed: 11879657]

Ostlund SB, Balleine BW. Orbitofrontal cortex mediates outcome encoding in Pavlovian but not instrumental conditioning. J Neurosci. 2007; 27:4819–25. [PubMed: 17475789]

Otto AR, Gureckis TM, Markman AB, Love BC. Navigating through abstract decision spaces: evaluating the role of state generalization in a dynamic decision-making task. Psychon B Rev. 2009; 16:957–63.

Padoa-Schioppa C. Neurobiology of Economic Choice : A Good-Based Model. Ann Rev Neurosci. 2011; 34:333–359. [PubMed: 21456961]

Padoa-Schioppa C, Assad JA. Neurons in the orbitofrontal cortex encode economic value. Nature. 2006; 441:223–226. [PubMed: 16633341]

Padoa-Schioppa C, Assad JA. The representation of economic value in the orbitofrontal cortex is invariant for changes of menu. Nat Neurosci. 2008; 11:95–102. [PubMed: 18066060]

Petrides, M. Mapping prefrontal cortical systems for the control of cognition.. In: Toga, AW.; Mazziotta, JC., editors. Brain mapping: the systems. Academic Press; San Diego: 2000. p. 159-176.

Pickens CL, Saddoris MP, Setlow B, Gallagher M, Holland PC, Schoenbaum G. Different roles for orbitofrontal cortex and basolateral amygdala in a reinforcer devaluation task. J Neurosci. 2003; 23:11078–11084. [PubMed: 14657165]

Preuss TM. Do rats have prefrontal cortex? The Rose-Woolsey-Akert program reconsidered. J Comp Neurol. 1995; 7:1–24.

Price JL. Definition of the orbital cortex in relation to specific connections with limbic and visceral structures and other cortical regions. Ann N Y Acad Sci. 2007; 1121:54–71. [PubMed: 17698999]

Ramus SJ, Eichenbaum H. Neural correlates of olfactory recognition memory in the rat orbitofrontal cortex. J Neurosci. 2000; 20:8199–8208. [PubMed: 11050143]

Redish AD, Jensen S, Johnson A, Kurth-Nelson Z. Reconciling Reinforcement Learning Models With Behavioral Extinction and Renewal: Implications for Addiction, Relapse, and Problem Gambling. Psych Rev. 2007; 114:784–805.

Rescorla, RA.; Wagner, AR. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement.. In: Black, AH.; Prokasy, WF., editors. Classical conditioning II: Current research and theory. Appleton-Century-Crofts; New York: 1972. p. 64-99.

Ricker ST, Bouton ME. Reacquisition following extinction in appetitive conditioning. Anim Learn Behav. 1996; 24:423–436.

Roesch MR, Taylor AR, Schoenbaum G. Encoding of time-discounted rewards in orbitofrontal cortex is independent of value representation. Neuron. 2006; 51:509–20. [PubMed: 16908415]

Rolls E, Hornak J, Wade D, McGrath J. Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage. J Neurophysiol. 1994; 57:1518–1524.

Rudebeck PH, Murray EA. Balkanizing the primate orbitofrontal cortex: distinct subregions for comparing and contrasting values. Ann N Y Acad Sci. 2011a; 1239:1–13. [PubMed: 22145870]

Rudebeck PH, Murray EA. Dissociable effects of subtotal lesions within the macaque orbital prefrontal cortex on reward-guided behavior. J Neurosci. 2011b; 31:10569–78. [PubMed: 21775601]

Rushworth MFS, Noonan MP, Boorman ED, Walton ME, Behrens TE. Frontal cortex and reward-guided learning and decision-making. Neuron. 2011; 70:1054–69. [PubMed: 21689594]

Schoenbaum G, Chiba AA, Gallagher M. Neural encoding in orbitofrontal cortex and basolateral amygdala during olfactory discrimination learning. J Neurosci. 1999; 19:1876–1884. [PubMed: 10024371]

Schoenbaum G, Eichenbaum H. Information coding in the rodent prefrontal cortex. i. single- neuron activity in orbitofrontal cortex compared with that in pyriform cortex. J Neurophys. 1995; 74:733–750.

Schoenbaum G, Nugent SL, Saddoris MP, Setlow B. Orbitofrontal lesions in rats impair reversal but not acquisition of go, no-go odor discriminations. Neuroreport. 2002; 13:885–890. [PubMed: 11997707]

Schoenbaum G, Setlow B, Nugent SL, Saddoris MP, Gallagher M. Lesions of orbitofrontal cortex and basolateral amygdala complex disrupt acquisition of odor-guided discriminations and reversals. Learn Mem. 2003a; 10:129–140. [PubMed: 12663751]

Schoenbaum G, Setlow B, Ramus SJ. A systems approach to orbitofrontal cortex function: recordings in rat orbitofrontal cortex reveal interactions with different learning systems. Behav Brain Res. 2003b; 146:19–29. [PubMed: 14643456]

Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997; 275:1593–1599. [PubMed: 9054347]

Seo H, Barraclough DJ, Lee D. Dynamic signals related to choices and outcomes in the dorsolateral prefrontal cortex. Cereb Cortex. 2007; 17:110–7.

Simmons JM, Richmond BJ. Dynamic changes in representations of preceding and upcoming reward in monkey orbitofrontal cortex. Cereb Cortex. 2008; 18:93–103. [PubMed: 17434918]

Sul JH, Kim H, Huh N, Lee D, Jung MW. Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. Neuron. 2010; 66:449–60. [PubMed: 20471357]

Sutton, RS.; Barto, AG. Reinforcement learning: An introduction. MIT Press; 1998.

Takahashi Y, Roesch MR, Stalnaker TA, Haney RZ, Calu DJ, Taylor AR, Burke KA, Schoenbaum G. The orbitofrontal cortex and ventral tegmental area are necessary for learning from unexpected outcomes. Neuron. 2009; 62:269–280. [PubMed: 19409271]

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

Teitelbaum H. A Comparison of Effects of Orbitofrontal and Hippocampal Lesions Upon Discrimination Learning and Reversal in the Cat. Exper Neurol. 1964; 9:452–462. [PubMed: 14188532]

Thorpe SJ, Rolls ET, Maddison S. The orbitofrontal cortex: neuronal activity in the behaving monkey. Exper Brain Res. 1983; 49:93–115. [PubMed: 6861938]

Tsuchida A, Doll BB, Fellows LK. Beyond reversal: a critical role for human orbitofrontal cortex in flexible learning from probabilistic feedback. J Neurosci. 2010; 30:16868–75. [PubMed: 21159958]

Tsujimoto S, Genovesio A, Wise SP. Monkey orbitofrontal cortex encodes response choices near feedback time. J Neurosci. 2009; 29:2569–74. [PubMed: 19244532]

Tsujimoto S, Genovesio A, Wise SP. Comparison of strategy signals in the dorsolateral and orbital prefrontal cortex. J Neurosci. 2011; 31:4583–92. [PubMed: 21430158]

Wallis JD. Orbitofrontal Cortex and Its Contribution to Decision-Making. Annu Rev Neurosci. 2007; 30:31–56. [PubMed: 17417936]

Wallis JD. Cross-species studies of orbitofrontal cortex and value-based decision-making. Nat Neuro. 2012; 15:13–9.

Walton ME, Behrens TEJ, Buckley MJ, Rudebeck PH, Rushworth MFS. Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. Neuron. 2010; 65:927–39. [PubMed: 20346766]

Wilson RC, Niv Y. Inferring relevance in a changing world. Front Hum Neurosci. 2011; 5:189. [PubMed: 22291631]

Wise SP. Forward frontal fields: phylogeny and fundamental function. Trends Neurosci. 2008; 31:599–608. [PubMed: 18835649]

Highlights

- We propose that OFC encodes the current, abstract, state of a task for the purposes of reinforcement learning.
- Diverse inputs to the OFC allow for states to include information that may not be perceptually observable.
- State information is used for both model-based and model-free reinforcement learning.
- We use the theory to account for results from OFC lesion experiments, and to make testable experimental predictions

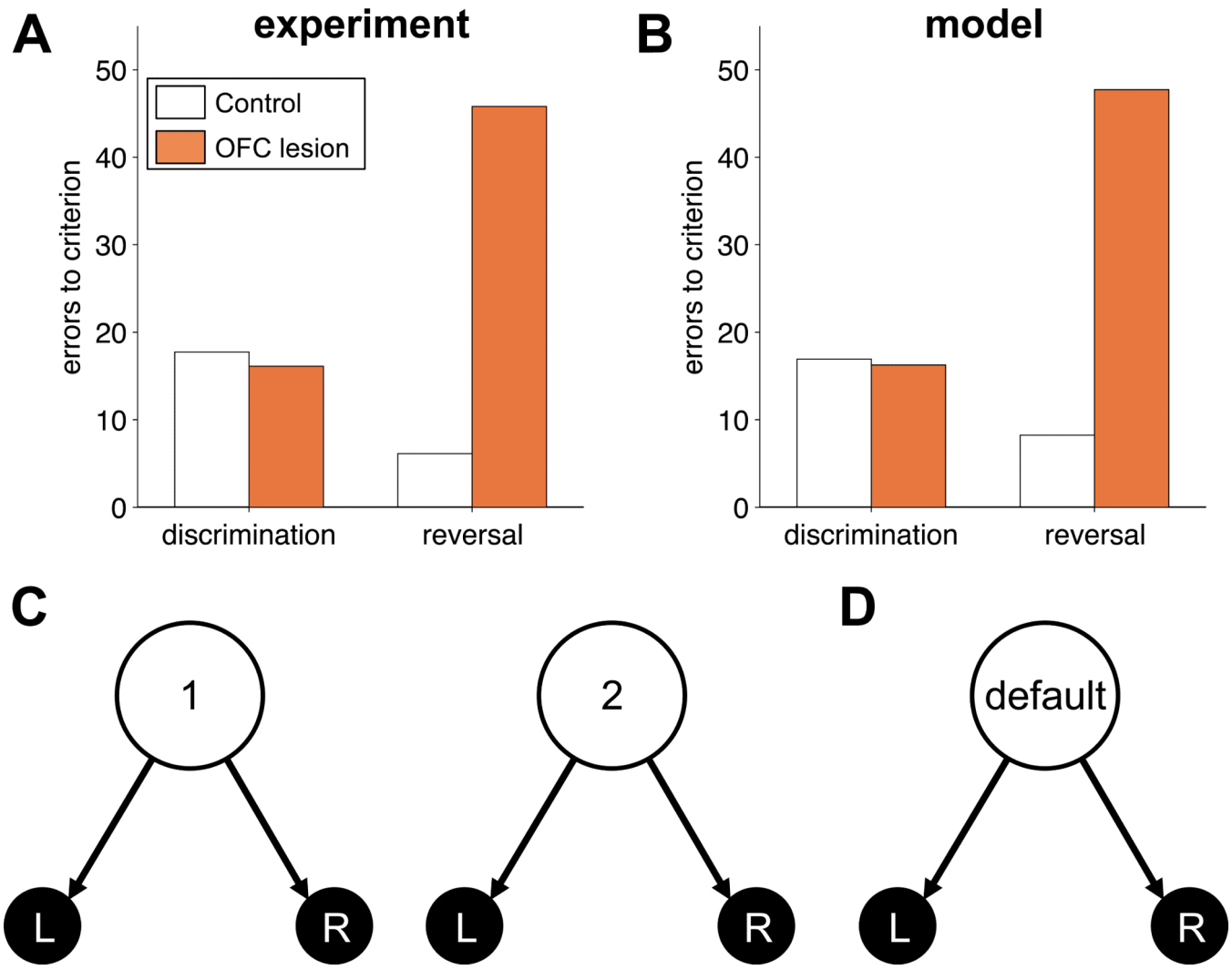

Figure 1.

Reversal learning. (A) Experimental results showing the mean errors to criterion in initial discrimination learning and final reversal for control (grey) and OFC-lesioned (orange) animals. (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.

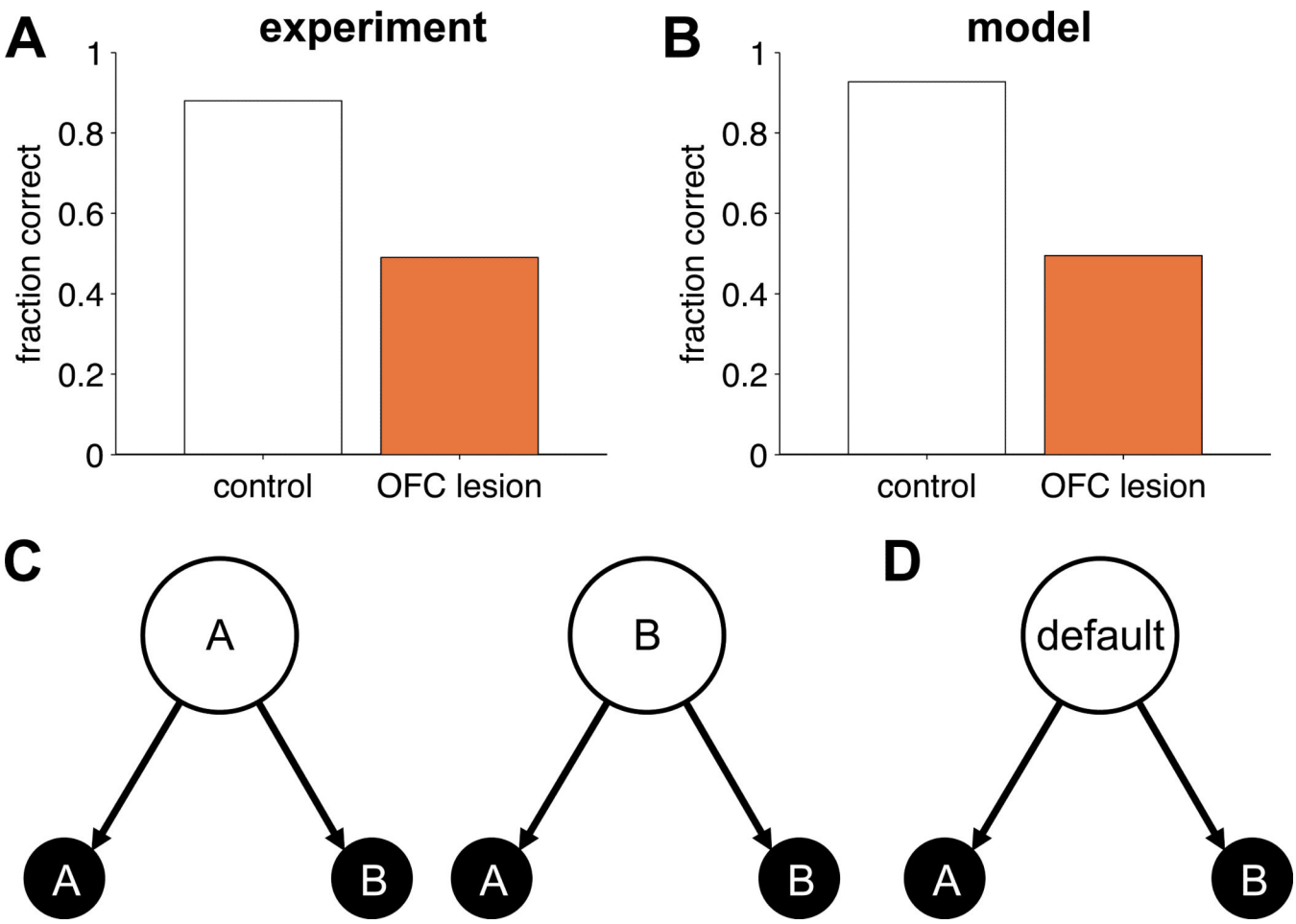

Figure 2.

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

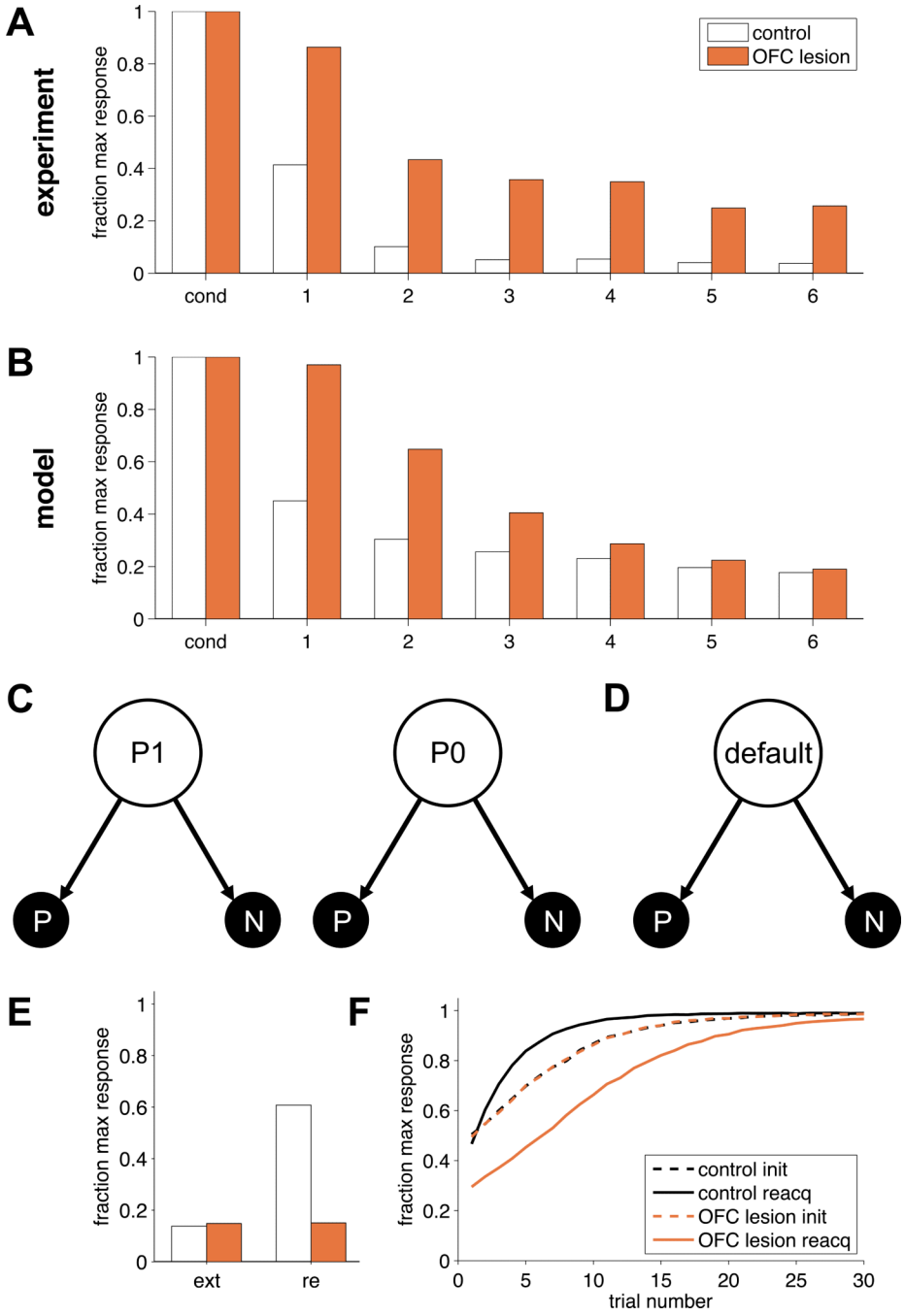

Figure 3.

Extinction. (A) Experimental results. Lever-press rates were normalized to the maximum response rate in conditioning. (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.

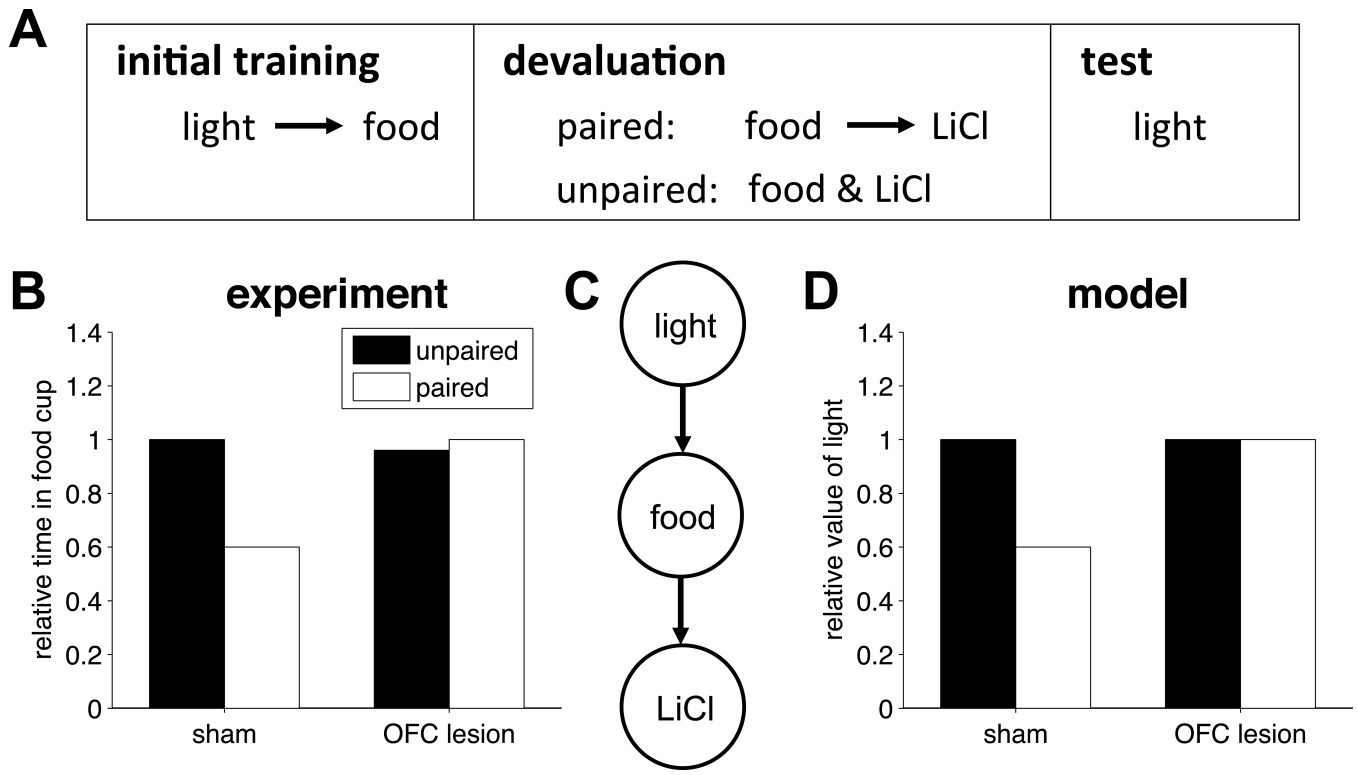

Figure 4.

Devaluation. (A) Animals are first 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.

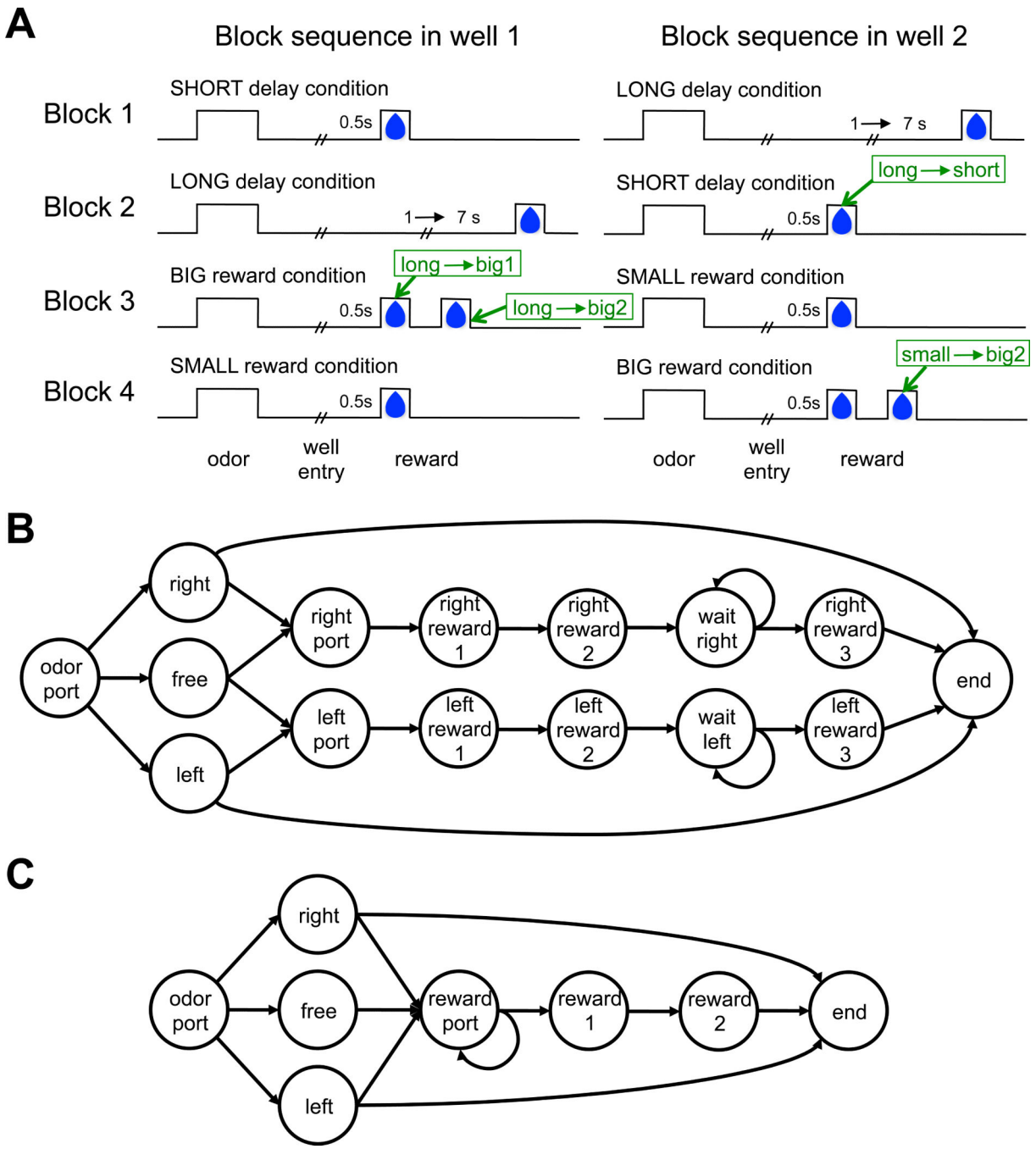

Figure 5.

Task design and state representations for Takahashi et al's (2011) odor guided choice task. (A) Time course of rewards for the different blocks. Times associated with positive prediction errors caused by unexpected rewards are labeled in green. (B) State representation used to model sham-lesioned controls. (C) State representation, used to model OFC-lesioned animals.

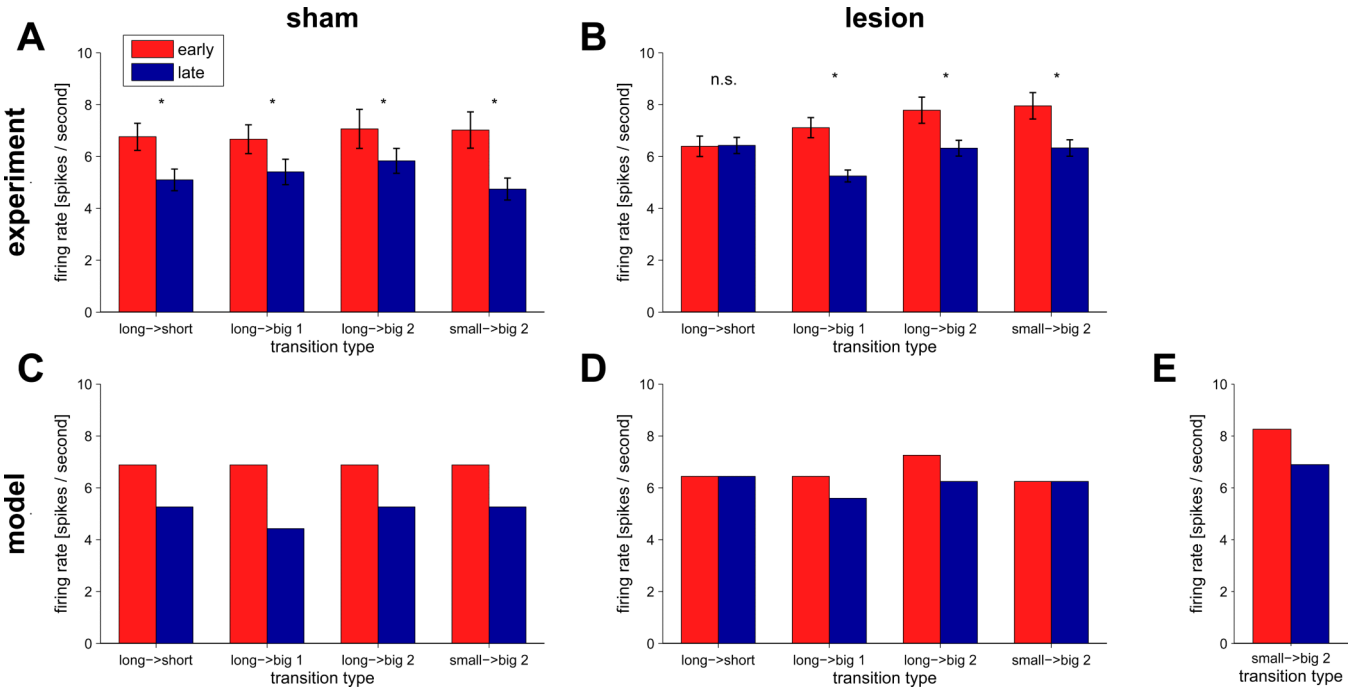

Figure 6.

Firing of dopaminergic VTA neurons at the time of unexpected reward early (first two trials, red) and late (last five trials, blue) 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). (A) Experimental data for sham-lesioned controls (n = 30 neurons). (B) Experimental data for the OFC-lesioned group (n = 50 neurons). (C) Model predictions for the sham-lesioned animals. (D) Model predictions for OFC-lesioned animals. (E) Model predictions for the small to big transition (small → big2) taking into account the variable third drop of juice.

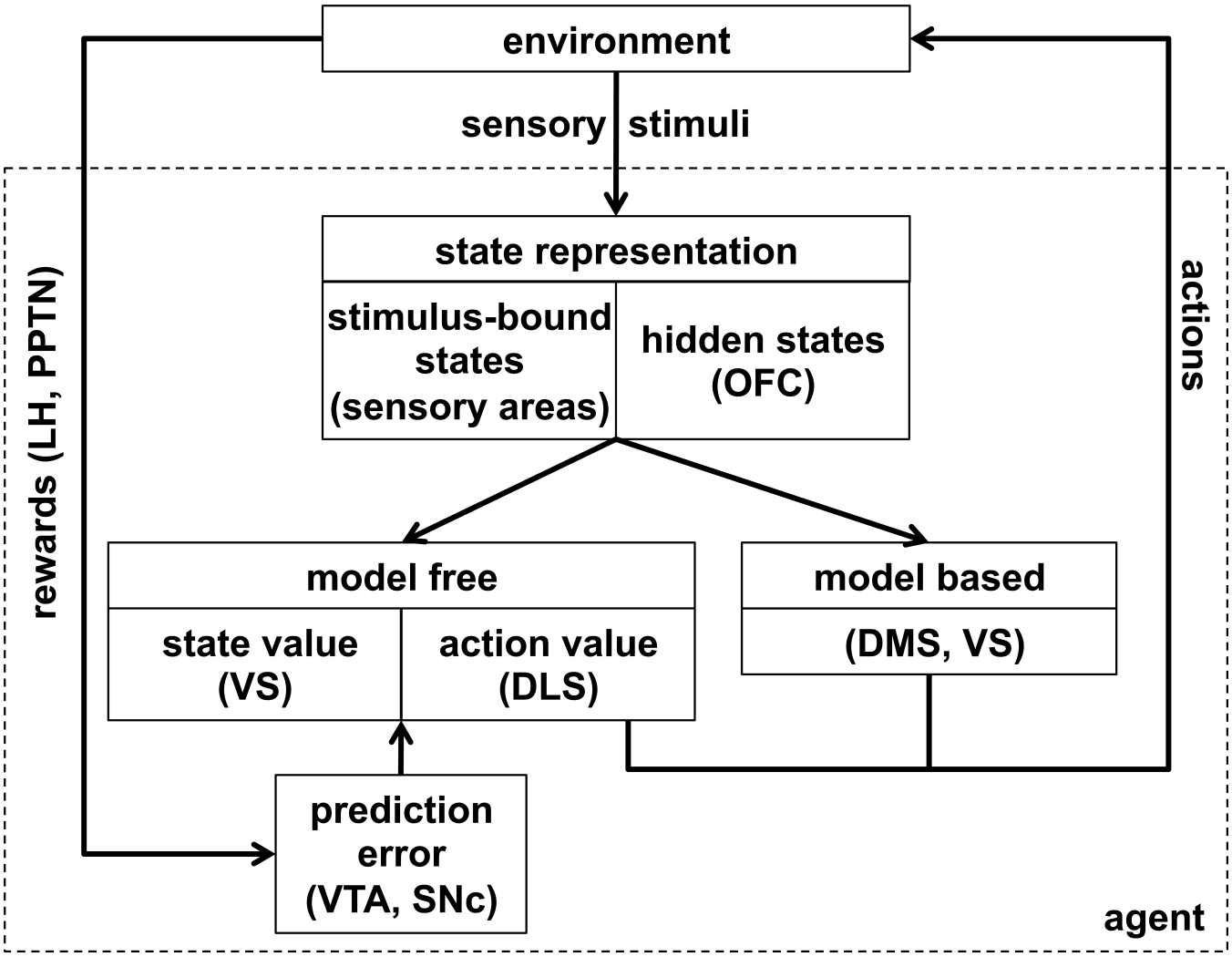

Figure 7.

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 both in OFC and in sensory areas, and a hidden (unobservable) component which we hypothesize is uniquely encoded in OFC. State representations are then used as scaffolding for both model-free and model-based RL. Model-free learning of state and action values occurs in ventral striatum (VS) and dorsolateral striatum (DLS), respectively, while model-based learning occurs in dorsomedial striatum (DMS) as well as VS.