Title: The rodent lateral orbitofrontal cortex as an arbitrator selecting between model-based and model-free learning systems

Marios C. Panayi *[1, 2], Mehdi Khamassi*[3], Simon Killcross[1]

[1] School of Psychology, The University of New South Wales, Australia.

[2] National Institutes on Drug Abuse, Intramural Research Program, Baltimore, Maryland, United States of America.

[3] Institute of Intelligent Systems and Robotics, Sorbonne Université, CNRS, F-75005 Paris, France

*Equal contribution and corresponding authors: m.panayi@unsw.edu.au, mehdi.khamassi@upmc.fr

Abstract

Our understanding of orbitofrontal cortex (OFC) function has progressed remarkably over the past decades in part due to theoretical advances in associative and reinforcement learning theories. These theoretical accounts of OFC function have implicated the region in progressively more psychologically refined processes from the value and sensory specific properties of expected outcomes to the representation and inference over latent state representations in cognitive maps of task space. While these accounts have been successful at modelling many of the effects of causal manipulation of OFC function in both rodents and primates, recent findings suggest that further refinement of our current models are still required. Here we briefly review how our understanding of OFC function has developed to understand two cardinal deficits following OFC dysfunction: reversal learning and outcome devaluation. We then consider recent findings that OFC dysfunction also significantly affects initial acquisition learning, often assumed to be intact. To account for these findings, we consider a possible role for the OFC in the arbitration and exploration between model-free and model-based learning systems, off-line updating of model-based representations, and attention. While the function of the OFC as a whole is still likely to be integral to the formation and use of a cognitive map of task space, these refinements suggest a way in which distinct orbital subregions, such as the rodent lateral OFC, might contribute to this overall function.

1. Introduction

1.1 Introduction

The orbitofrontal cortex (OFC) continues to attract research interest as a key region involved in flexible value-based decision making, a process fundamental to normal decision making and disorders such as addiction and schizophrenia (Kanahara et al., 2013; Schoenbaum et al., 2016; Schoenbaum & Shaham, 2008). Historically, OFC function has predominantly been understood in the context of the psychological constructs of associative learning theory (Delamater, 2007; Schoenbaum et al., 2009; Stalnaker et al., 2015). Recently, these ideas have been extended and incorporated into the computational models of reinforcement learning theory (RL) (Behrens et al., 2018; Bradfield & Hart, 2020; Niv, 2019; Wikenheiser & Schoenbaum, 2016; Wilson et al., 2014). To date these RL models have had the most success in accounting for the extant literature on OFC function.

In this article we will first present a brief and selective overview of OFC function in experimental research focusing on two cardinal deficits following OFC dysfunction (i.e. lesion, pharmacological, chemogenetic, optogenetic inactivation etc …): reversal learning deficits and outcome devaluation deficits. These deficits are remarkably consistent between rodents and primates (Boulougouris et al., 2007; Butter, 1969; Gallagher et al., 1999; Izquierdo et al., 2004; Izquierdo & Murray, 2004, 2005; Machado & Bachevalier, 2007; Panayi & Killcross, 2018; Pickens et al., 2003, 2005; Schoenbaum, Setlow, Nugent, et al., 2003; West et al., 2011) (but see also Rudebeck et al., 2013; Sallet et al., 2020) and must be accounted for by any theory of OFC function. Then, we will discuss how RL models of OFC function might account for recent findings that the OFC is involved in the initial acquisition of simple tasks, a process previously thought to be unaffected by OFC dysfunction (for a comprehensive review see Murray et al., 2007; Rudebeck & Murray, 2014; Stalnaker et al., 2015).

1.2 The orbitofrontal cortex

We shall consider evidence from non-human primates and rodent studies of OFC function as these have contributed to the majority of experimental evidence regarding OFC function. In primates regions that have been considered as OFC encompass a large number of structures spanning medial and lateral orbital sulci, including Walker areas 11, 12, 13, 14 and aspects of agranular insular cortex (Ongur & Price, 2000; Price, 2007; Rudebeck & Murray, 2011a; Sallet et al., 2020). In rodents the OFC also encompasses a large number of prefrontal structures along the entire orbital surface including medial, lateral, dorsolateral, ventral OFC, and often encompasses the lateral structures of the rostral agranular insular cortex (Krettek & Price, 1977; Price, 2006). While there is no clear consensus of homologous OFC regions between rodents and primates, there is good evidence to suggest that homologies can be established based on similar patterns of anatomical projections and functional properties (Roesch & Schoenbaum, 2006).

However, it is also becoming increasingly apparent that there is significant functional heterogeneity within the OFC (see Barreiros et al., 2021 this issue). Here we will discuss evidence spanning several OFC subregions in both primates and rodents as the OFC region as a

whole appears to be involved in similar aspects of flexible behavioural control, and there is still a paucity of evidence differentiating functional differences within the OFC (for a recent review see Izquierdo, 2017b). Indeed, models of OFC function have often accounted for experimental findings spanning multiple OFC subregions and species in this manner (Rudebeck & Murray, 2014; Wilson et al., 2014), suggesting a similar organization of functional principles across OFC subregions. Acknowledging this important caveat, we will first focus on the function of the OFC as a whole, and then explore how recent data specifically from the rodent lateral OFC might be interpreted as a subordinate function of the OFC as a whole.

2. Cardinal features of OFC dysfunction

2.1 Reversal Learning

Modern experimental research interest in OFC function began with studies of reversal learning deficits in non-human primates (Butter, 1969; Butter et al., 1963; McEnaney & Butter, 1969). For example, in a discriminative conditioning task, subjects first learned the relationship between an object that lead to reward (A+) and an object that led to no-reward (B-) (Butter, 1969). Subjects with OFC lesions can learn to choose A+ and inhibit choices to B- at a rate comparable to control subjects. However, when these initial cue-reward contingencies are reversed (i.e. A-/B+), OFC lesions impair the ability to flexibly update behaviour and subjects show perseverative responding to the no longer rewarded A-. This deficit is also seen in extinction procedures where an initially rewarded cue (A+) is no longer rewarded (A-), i.e. presented in extinction. Again, OFC dysfunction results in persistent responding to A- in extinction (Butter, 1969; Lay et al., 2020; Panayi & Killcross, 2014).

One account of these reversal learning deficits is that the OFC is necessary for representing and updating the value of expected outcomes formed during Pavlovian cue-outcome learning. Population and single-unit neuronal firing in the OFC tracks many feature of reward value during learning, including firing to reward predictive cues i.e. expected value (Moorman & Aston-Jones, 2014; Roesch et al., 2010; Schoenbaum et al., 2009; Schoenbaum, Setlow, Saddoris, et al., 2003; Takahashi et al., 2013; van Duuren et al., 2008; van Wingerden et al., 2010). In reinforcement learning, the expected value accrued to a cue is thought to be fundamental to prediction errors (Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Sutton & Barto, 1998), the difference between expected and actual value of a reward. Indeed, OFC lesions have been found to disrupt normal mid-brain dopaminergic prediction error signals (Takahashi et al., 2011), which have been shown to drive learning (Nasser et al., 2017; Schultz et al., 1997; Steinberg et al., 2013).

If the OFC represents expected value, it makes sense that OFC lesions disrupt reversal learning since the reversal involves a significant change in outcome contingencies i.e. an expected reward is now omitted (A+ -> A-). Furthermore, the OFC appears to be necessary in other situations where expected value is necessary for updating learning such as Pavlovian overexpectation, a task in which combining the expected value of multiple cues leads to overestimation of reward and updating expected values accordingly (Lay et al., 2020; Lucantonio et al., 2015; Takahashi et al., 2009). However, OFC lesions do not disrupt initial acquisition (A+) where expected-value information for prediction-errors is also necessary for

learning. Therefore, the OFC cannot simply represent expected value necessary for calculating prediction errors, and other candidates such as the ventral striatum shall be considered (Khamassi et al., 2008) (for a discussion of negative vs. positive prediction-error representations within OFC see Stalnaker et al., 2015).

One solution to this problem is that, in addition to expected value, cues can come to predict multiple aspects of reward such as their sensory specific properties e.g. flavour, texture, location etc.… (Delamater, 2007, 2012; Delamater & Oakeshott, 2007; Hall, 2002; Killcross & Blundell, 2002; Wagner & Brandon, 1989). Therefore, in a reversal task, in addition to the value of the outcome changing at the point of reversal (A+ -> A-; i.e. High -> Low), the identity of the outcome also changes (sucrose -> nothing). Indeed, there is a rich history in associative learning theory of reward omission being considered a unique outcome (Delamater, 2004; Urcuioli, 2005; Westbrook & Bouton, 2010). Furthermore, expected outcome activity within the OFC encodes many of these aspects of the expected outcome identity in addition to value many features of reward outcomes (e.g. size, preference, identity, time, location, probability, certainty, salience (Delamater, 2007; Ogawa et al., 2013; Padoa-Schioppa, 2009; Sadacca et al., 2018; Stalnaker et al., 2014; Takahashi et al., 2013; Zhou et al., 2019). Therefore, if the OFC is necessary for representing the identity of expected outcomes, OFC lesions would disrupt only reversal learning and not initial acquisition because outcome identity is only relevant to task performance at the point of reversal. Clearer evidence for the functional role of the OFC in encoding sensory-specific outcome information comes from the second cardinal feature of OFC dysfunction: outcome devaluation deficits.

2.2 Outcome Devaluation

The second characteristic feature of OFC dysfunction is a deficit in outcome devaluation procedures (Gallagher et al., 1999; Izquierdo & Murray, 2000; Murray et al., 2015; Panayi & Killcross, 2018; Pickens et al., 2003, 2005). In a typical Pavlovian version of the procedure, subjects first learn about a specific cue-outcome (CS-US) relationship e.g. a 10s light predicts the delivery of a lemon flavoured sucrose reward. In a subsequent second stage, the value of this specific outcome is devalued, often by eating the outcome to satiety (sensory specific satiety) or pairing consumption with illness (via injection of Lithium Chloride) to establish a specific taste-aversion. Importantly, this new learning that the outcome is no longer valuable is done independently of the predictive light CS. Next, the subjects are presented with the light CS to assess whether the subjects will continue to respond for the outcome that has now been devalued. Control subjects will appropriately reduce responding to the CS predicting the now devalued outcome relative to a non-devalued control condition (either a different non-devalued group or a different non-devalued CS-US relationship within the same subject).

Subjects with OFC dysfunction are significantly impaired on outcome devaluation tests and will continue to respond to the devalued CS as if the outcome had not been devalued. Notably, OFC dysfunction does not appear to disrupt the initial acquisition of the CS-US relationship, or the outcome devaluation manipulation (specific satiety consumption or taste-aversion learning) (Gallagher et al., 1999; Izquierdo & Murray, 2000; Murray et al., 2015; Panayi & Killcross, 2018; Pickens et al., 2003, 2005). Therefore, it is only at test when the specific identity information

about the expected outcome is relevant to adaptive behaviour that OFC dysfunction is detected. This supports the theoretical account of the OFC as the neural locus of the outcome-specific properties of expected outcomes (Delamater, 2007; Roesch & Schoenbaum, 2006; Schoenbaum et al., 2009). Informally, in a devaluation test, subjects with OFC dysfunction know that the light predicts a rewarding outcome, but they do not know that the reward is specifically, say, the lemon flavoured sucrose solution (which is now no longer very rewarding).

More recently, model-based reinforcement learning (RL) theories of OFC function have proposed a complementary class of function to the OFC: the representation or use of latent states (Wilson et al., 2014). In a task such as Pavlovian conditioning, where a cue predicts an outcome, the task can be split into distinct observable physical states e.g. “cue absent”, “cue present”, and “reward”. However, after learning this task there may also be learning of latent states which are signalled by partially observable information and recalled into working memory such as reinforcement history. Together, these observable and latent state representations have been proposed as a cognitive map of task structure (Behrens et al., 2018; Wikenheiser & Schoenbaum, 2016; Wilson et al., 2014). The OFC is thought to represent these latent states, and OFC lesions are thought to disrupt learning or behaviour that involves making inferences over latent states (Bradfield & Hart, 2020; Niv, 2019; Sharpe et al., 2019). Examples of OFC deficits in latent state inferences include extinction and reversal learning (reinforcement history no longer matches current reinforcement contingencies) (Boulougouris et al., 2007; Panayi & Killcross, 2014; Rudebeck & Murray, 2011b; Schoenbaum et al., 2002), outcome devaluation (the value of the predicted outcome changes)(Gallagher et al., 1999; Panayi & Killcross, 2018; Pickens et al., 2003, 2005; West et al., 2011), Pavlovian overexpectation (combining the predicted value of multiple cues)(Takahashi et al., 2009), sensory preconditioning (inferring the future sequence of neutral events)(Hart et al., 2020; Jones et al., 2012). This RL account of OFC function is the most successful theoretical framework to date in accounting for the extant OFC literature.

2.3 Acquisition learning

We have briefly introduced the two cardinal experimental features of OFC dysfunction, reversal learning and outcome devaluation deficits, and how they relate to predicted outcome representations within the OFC. We now focus on the lack of effect of OFC dysfunction on initial acquisition learning in these tasks, a critical null effect that must also be considered. This null effect has been replicated in many studies of OFC dysfunction (Izquierdo, 2017b; Murray et al., 2007; Murray & Rudebeck, 2018; Stalnaker et al., 2015), except in tasks with complex probabilistic cue-outcome relationships (Walton et al., 2010), or tasks in which correct responding depends on the identity of the predicted outcome (McDannald et al., 2005)). For example, OFC lesions disrupt the ability to discriminate quickly between two options that lead to unique outcomes compared to a common outcome (the differential outcome effect (McDannald et al., 2005; Ramirez & Savage, 2007; Trapold & Overmier, 1972)).

This has led to the implicit assumption that, in a simple task such as Pavlovian conditioning with a single deterministic cue-outcome relationship (e.g. a 10s light always predicts delivery of the same sucrose reward), the OFC is not involved in initial learning. Indeed, computational

modelling of OFC dysfunction might even suggest that the representations underlying initial acquisition are intact in animals with OFC lesions (Wilson et al., 2014). A superficial interpretation of these accounts would be that the OFC is not involved in initial acquisition at all. It is only when some established learning needs to be modified/updated that the OFC plays a role in learning and behaviour.

However, theoretical accounts of OFC function predict that the nature of this initial learning should be impoverished in some way e.g. missing sensory specific information or an incomplete representation of the underlying task structure (Schoenbaum et al., 2009; Wilson et al., 2014). Therefore, the simple modelling of no deficits during initial acquisition must be considered a practical simplification and not a prediction of these models. Here, we highlight that, while it is often considered a null behavioural result, OFC dysfunction during acquisition should disrupt the associatively evoked representations formed during acquisition.

Unsurprisingly, there has been very little focus on the effect of OFC dysfunction on initial acquisition in simple single CS-US Pavlovian acquisition. One exception to this is studies of the role of the OFC Pavlovian sign- and goal-tracking behaviour in rodents. In a typical rodent sign-tracking task (Boakes, 1977), a typical Pavlovian CS-US relationship is established by pairing the insertion of a lever with a food pellet reward. Initially, rats will approach the magazine site where the reward is delivered (goal-tracking), but over the course of acquisition rats will engage with the lever cue that signals the reward (sign-tracking). This sign-tracking behaviour has been conceptualized as the attribution of motivational value to the lever cue and the dominant influence of a feature-model-free learning system (Lesaint et al., 2014). OFC lesions and inactivation have been found to disrupt sign-tracking behaviour and shift responding towards goal-tracking (Chudasama & Robbins, 2003; Panayi & Killcross, 2018; Stringfield et al., 2017)(but see Chang, 2014). This suggests that OFC dysfunction can indeed disrupt some aspect of initial acquisition learning in simple Pavlovian CS-US procedures.

Surprisingly, we have recently found that lateral OFC lesions in rats significantly disrupt simple single CS-US Pavlovian acquisition (Panayi & Killcross, 2020). Furthermore, whereas pre-training lesions significantly enhanced acquisition after extended training, post-training lesions and functional inactivation impaired subsequent acquisition. As discussed above, these effects are implied by RL models of OFC function but have not been explicitly predicted. Therefore, these results reveal a fundamental aspect of OFC function that must be accounted for by current RL model formulations. Next, we will first introduce the RL modelling framework that has been applied to understanding OFC function. Then we will consider what assumptions might be necessary to update our current models and accommodate these effects of OFC dysfunction on acquisition learning. Specifically, we will consider the role of the rodent lateral OFC as an arbitrator between model-free and model-based learning systems.

3. Reinforcement learning systems

3.1 Preliminary considerations

Reinforcement learning theory. Modelling value-based decision-making and learning mechanisms involving the OFC's subcircuits is often envisioned through the prism of the reinforcement learning (RL) theory (Sutton & Barto, 1998), in close interaction with economics models based on the notion of expected utility, such as in game theory (Daw & O'Doherty, 2014; Rustichini & Padoa-Schioppa, 2015; Schultz et al., 2017). Here, we will adopt an RL perspective and try to relate a series of experimental results with RL computational mechanisms.

In the RL framework, the task is usually modelled as a succession of discrete *Markovian* states *s* taken from a finite set of states: $s \in \mathcal{S}$ (*e.g.,* state s_1: the agent is in the middle of the conditioning chamber; state s_2: the agent is near the lever; state s_3: the agent is near the lever and a food pellet has been delivered in the magazine, etc.). These states are called *Markovian* because we assume (for mathematical simplicity) that the Markov hypothesis is verified: being in a given state of the task is sufficient to determine what the consequence of the action will be; in other words, the effect of the present action does not depend on a remembered event from the past. Nevertheless, this does not prevent the agent from sometimes pausing its decision process about what action to perform in the real world in order to replay some elements in episodic memory so as to re-estimate an action's value before deciding. We will also see cases where this *off-line* value update process (off-line because it occurs while the agent is momentarily suspending its interactions with the real-world) can be employed to mentally (virtually) simulate the anticipated consequences of an action (or of a sequence of actions) using a cognitive map (O'Keefe & Nadel, 1978) in order to re-evaluate the action before deciding (Johnson & Redish, 2007).

Action values. The decisions made by the agent rely on the comparison of action values, which represent their respective expected utilities. Specifically, the agent can choose among a finite set of actions $a \in \mathcal{A}$. The learned *value Q(s,a)* of an action *a* in a given state *s* informs the agent about how good this action was on average during past experience, and thus how desirable it is now. We neglect here action values estimated from instruction rather than from experience (Erev & Haruvy, 2016). Nevertheless, one could straightforwardly generalize the *off-line* value update process proposed here to also cover mental computations using information from task instructions.

Reward model. In the RL context, the value is considered to reflect the agent's sole motivation to try and maximize the amount of reward it can get from the environment. The reward is basically modelled as a positive scalar value $r \in \mathbb{R}^+$ when the agent reaches a rewarding state (*e.g.,* reaching the food pellet in a magazine), and zero in other states. Moreover, unless a reward devaluation (Dickinson & Balleine, 1995) occurs, we assume for the sake of simplicity that the agent's drive for the reward is constant throughout the task. Despite such a simplicity, the same RL principles generalize to more complex reward functions, such as those enabling to cope with both reward and punishment (*i.e.,* negative reward/positive punishment) (Palminteri et al., 2015), multidimensional reward functions were each dimension represents a particular need for homeostatic regulation (food, water, temperature) (Keramati & Gutkin, 2014; Konidaris & Barto, 2006), and even models where some reward dimensions represent information obtained from the environment (Genzel et al., 2019), which roughly corresponds to the notion of epistemic value used in the active inference framework (Friston et al., 2017).

Learning action values so as to maximize reward. Now the central question for the agent is: How to acquire a behaviour which enables it to maximize reward from the environment? In reinforcement learning models applied to neuroscience and psychology, such a behavioural output of the agent can either be an instrumental action (*i.e.,* pressing a lever) in operant conditioning tasks (Daw et al., 2005), a Pavlovian response such as an approach in Pavlovian conditioning paradigms (Dayan et al., 2006; Lesaint et al., 2014), or even a movement following a cardinal direction in navigation experiments (Dollé et al., 2018a). In all these cases, the estimated state-action value *Q(s,a)* represents the mathematical expectation $\mathbb{E}[.]$ of the sum of future rewards *r* after performing this action: $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$, where $\gamma (\gamma < 1)$ is a discount factor which assigns weaker weights to long-term rewards than to short-term rewards. Rather than focusing on the immediate reward following action *a*, this equation takes into account long-term consequences of the action, which is important in tasks where a sequence of actions is required to get rewarded, such as in the "two-step" task (Daw et al., 2011).

Learning state/stimulus values. Interestingly, RL models not only permit to learn action values, but also state (or stimulus) values. For instance, in the Actor-Critic model (Joel et al., 2002; Khamassi et al., 2005; Sutton & Barto, 1987), the reward is used to reinforce both the probability to perform certain actions in the Actor, and the value of states in the Critic. In the case of a navigation task where a state can represent a particular position within a cognitive map, a state value learning process can be used to model conditioned place preference (Arleo & Gerstner, 2000). In the general case, a state does not necessarily represent an allocentric position in space, but can also represent any state where a meaningful event occurs for the task, such as the presentation of a stimulus or the delivery of a food pellet (Daw et al., 2005; Khamassi & Humphries, 2012; Wilson et al., 2014). While the story can become a bit more complicated when stimulus values are learned without relying on the notion of state (Schmajuk et al., 1996), or when some models learn the value of specific features of stimuli (texture, shape, colour, etc.) (Niv et al., 2015), in the following we will simply consider that the appearance of a stimulus triggers a state change (*i.e.,* from stimulus off to stimulus on).

3.2 Different value learning systems

Importantly, in the machine learning literature, there are different possible learning strategies to estimate action values and state values. In particular, two famous ones, called *model-based* and *model-free* RL, turn out to be relevant for the study of OFC functions (Bradfield & Hart, 2020; Niv, 2019; Sharpe et al., 2019; Wilson et al., 2014). Before diving into their description, it is important to arm the reader with a few precautions. As with any computationally-grounded distinction, there comes the risk of oversimplification (Collins & Cockburn, 2020). Indeed, considering two different learning mechanisms to update action values does not imply that the two underlying learning systems are completely disjoint nor in total competition. Instead, there can be cooperation between them so that a sequence of actions can rely on the alternation between decisions of each one (Dollé et al., 2010); there can be mutual help between MB and MF processes through learning by observing each other's output (Dollé et al., 2018a); there can be bootstrapping of MF learning through MB offline replay (Cazé et al., 2018; Mattar & Daw, 2018), etc.. Nevertheless, we argue here that the MB/MF distinction still represents a useful clarification

of distinct computational mechanisms for value update, which generate distinct hypotheses and predictions that can guide future experiments. This can help to better understand value-based decision-making in a variety of contexts, such as economic choices (Lee et al., 2014), instrumental conditioning (Daw et al., 2005; Keramati et al., 2011), Pavlovian conditioning (Lesaint et al., 2014), or even navigation (Khamassi & Humphries, 2012; Pezzulo et al., 2013; Van Der Meer et al., 2012).

This computational distinction also recently turned out useful in understanding how the balance between different learning strategies evolves through development (Decker et al., 2016), or how it varies between different individuals in Pavlovian conditioning paradigms, such as sign- versus goal-tracking behaviours (Cinotti, Marchand, et al., 2019; Lesaint et al., 2015). Finally, it is worth noting that this distinction is currently also a hot topic in machine learning and robotics (Khamassi, 2020; Kober et al., 2014; Wang et al., 2019), so that upcoming breakthroughs in these disciplines can later on fertilize computational neuroscience models of learning and decision-making.

While MB and MF learning processes have been extensively described in the literature (e.g., see (Daw et al., 2005; Keramati et al., 2011; Khamassi & Humphries, 2012)), here we briefly recount the main computational distinctions between the two so as to derive clear distinctive interpretations of experimental results in the next sections.

Model-based learning. A *model-based* agent is an agent which manipulates an internal model of the world to make decisions (Sutton & Barto, 1998). In the case of navigation, where states represent different allocentric positions within the environment, such an internal model takes the form of a *cognitive map* (O'Keefe & Nadel, 1978). The cognitive map hypothesis of OFC function (Wilson et al., 2014) thus relies on the notion of internal models. We will nowdescribe how internal models are built and manipulated by a model-based agent in the RL theory.

Conventionally, this model is a set of two mathematical functions: a reward function $R(s,a)\colon(\mathcal{S},\mathcal{A}) \longrightarrow \mathbb{R}$, and a transition function $T(s,a,s')\colon(\mathcal{S},\mathcal{A},\mathcal{S}) \longrightarrow [0;1]$. The former represents the agent's memory of how much reward can ultimately be obtained from the environment when performing an action *a* in a state *s*. If we consider discrete states and actions (as is the case here), this can be represented as a table where each (state,action) couple has an associated average reward value. In practice, each (state,action) couple can be associated with a probability distribution over reward magnitude if the agent obtains variable quantities of reward (*e.g.,* 45mg food pellet, 42mg, 46mg, 0mg, etc.). Conversely, the latter transition function represents the agent's estimation of all possible transition probabilities between couples of states *(s,s')* of the environment, given a performed action *a*. One can conceive this transition estimation as a simple count of the frequencies of each encountered *(s,a,s')* transition. For instance, if the agent has performed 10 times action a_1 (*e.g.,* press the lever) in state s_1 (*e.g.,* being close to the lever with the light CS on) and observed 8 times out of 10 the state s_2 where a reward has been delivered, while 2/10 times the agent remained in state s_1 (thus no reward delivered), then we have the estimated transition probabilities $p(s_2|s_1,a_1) = 8/10$, $p(s_1|s_1,a_1) = 2/10$, and $p(s_i|s_1,a_1) = 0$ for all other states $s_i, i \notin \{1,2\}$.

At each timestep *t*, the agent can either interact with the world or use its current internal model of the world to infer what is its current estimation of the value $Q_{MB}(s,a)$ of a given (state,action) couple. The former can be done by performing an action *a* in the current state *s* and observing the resulting state *s'* and subesequent reward *r* in order to update the model (*i.e.,* update the transition and reward functions). The latter can be achieved by picking a (state,action) couple *(s,a)* (for the moment, let's say randomly, but we wil see in the off-line learning subsection that the agent can choose to "replay" specific sequences of (state,action) couples) and, using the transition and reward functions then perform a *value iteration* (Sutton & Barto, 1998) process:

$$Q_{MB}^{(t+1)}(s,a) = R(s,a) + \gamma \sum_{s'} T(s,a,s') \max_{k \in \mathcal{A}} Q_{MB}^{(t)}(s',k) \quad \textbf{(1)}$$

Cognitive map theories of OFC function suggests that the OFC represents the states and the link between states (*i.e.,* state transitions) that make up an internal model of the world (Wilson et al., 2014). Thus, following OFC dysfunction, in a given state *s* following a given action *a*, an organism is unable to infer the identity and value of future state *s'* and reward *r* without actually performing the action and observing the consequence in the environment.

Model-free learning. In contrast to a model-based agent, a model-free agent does not have access to a model of the world. Instead, it has to iteratively update its model-free estimate of value function $Q_{MF}(s,a)$ through interaction with world:

$$Q_{MF}^{(t+1)}(s,a) = Q_{MF}^{(t)}(s,a) + \alpha \left(r_t + \gamma \max_{k \in \mathcal{A}} Q_{MF}^{(t)}(s',k) - Q_{MF}^{(t)}(s,a) \right) \eta_t(s) \quad \textbf{(2)}$$

where $\alpha \in [0;1]$ is the learning rate, $\eta_t(s) \in [0;1]$ is the current attention level paid to a particular state (or stimulus), and the term between parentheses, often written δ_t, is called the *temporal-difference error* in machine learning (Sutton & Barto, 1998) or the *reward prediction error* (Schultz et al., 1997) in neuroscience. The only difference between this equation and standard model-free reinforcement learning is the attention level. This simply captures the fact that the agent may pay more attention to a stimulus or to a state than to another, so that learning will be modulated by the attention level (Lesaint et al., 2014; Niv et al., 2015), a common consideration in learning models (Mackintosh, 1975; Pearce & Hall, 1980). In the extreme case, learning will occur only for the attended stimulus ($\eta_t(s) = 1$) but not for unattended stimuli ($\eta_t(s_i) = 0, i \neq s$), which can occur when initial learning with stimulus s alone results in overshadowing when other stimuli are presented concomitantly.

Decision-making. Each time the agent is in a state *s* and wants to decide which action *a* to perform next, no matter if the agent is model-free or model-based, the agent will have to normalize the values of all possible actions in this state, so that they sum to one, thus representing a probability distribution over actions, and so that it can then pick an action within this probability distribution. Practically, this action probability distribution is computed using a Boltzmann softmax function:

$$P^{(t)}(a|s) = \frac{exp^{\beta Q_x^{(t)}(s,a)}}{\sum_{k \in \mathcal{A}} exp^{\beta Q_x^{(t)}(s,k)}} \quad \textbf{(3)}$$

where $x = MB \; or \; MF$, and β is the inverse temperature which tunes the *random exploration* level (Cinotti, Fresno, et al., 2019): β close to 0 means that the action probability distribution will be nearly flat, so that all actions are equiprobable (exploration); when β is high, or even tends

towards infinity, the probability of performing the action with the highest value will be close to 1 (exploitation).

Off-line learning. In addition to learning through the direct interaction with the environment, we will call *off-line learning* any update process that occurs while the agent is immobile (*e.g.,* quiet wakefulness or sleep). During such an immobility, the agent could do some mental simulations of action sequences that would update their model-based values through Equation 1 before moving to the next decision (Johnson & Redish, 2007). Alternatively, the agent may replay some previously performed actions (and the memorized resulting states and rewards) in order to consolidate memory, which can be captured by updating these actions' model-free values through Equation 2 (Cazé et al., 2018). Finally, during long periods of inactivity where the agent does not actively update action values, these action values may be progressively forgotten (Kato & Morita, 2016; Khamassi et al., 2015; Niv et al., 2015):

$$Q_{MF}^{(t+1)}(s,a) = Q_{MF}^{(t)}(s,a) + \kappa\left(Q_{MF}^{(0)}(s,a) - Q_{MF}^{(t)}(s,a)\right) \textbf{(4)}$$

where $Q_{MF}^{(0)}(s,a)$ is the initial value of this (state,action) couple (*e.g.,* 0) and $\kappa \in [0;1]$ is the forgetting rate. Importantly, day-to-day forgetting can be (at least partly) compensated by doing model-based mental simulation, so that action values are preserved. As we will argue later on, the increased day-to-day forgetting observed in OFC-inactivated animals (Panayi & Killcross, 2014), may be due to the impairment of such a model-based off-line compensation mechanism.

Arbitration between learning systems. A classical way of arbitrating between MB and MF learning systems is to orchestrate an uncertainty-based competition (Daw et al., 2005; O'Doherty et al., 2020): the most certain system makes decisions, while both can learn from the outcome of the other system's decisions (Dollé et al., 2010). When the learning systems are implemented as approximate Bayesian learners, the imprecision (or spread) of the distributions over estimated action values can be used as a marker of uncertainty (Daw et al., 2005; Keramati et al., 2011). Under some conditions, alternative measures of uncertainty can give similar proxies to uncertainty at a lower computational expense, such as squared prediction errors (Lee et al., 2014), absolute variations of action values (Cazé et al., 2018), or even the systems relative choice uncertainty in simple stationary tasks (Viejo et al., 2015). Finally, in some models arbitration is performed by a third system, called the *meta-controller*, which learns through reinforcement which system to select in each state of the environment (Dollé et al., 2008, 2010, 2018b). Here, because our goal is not to propose a new model but rather to illustrate how impaired arbitration may mimic some experimental results under OFC inactivation, we will show model simulations using choice uncertainty for simple acquisition tasks (Experiments 1 and 2), and the following combination of choice uncertainty and absolute variations of actions values in tasks where acquisition is followed by extinction (Experiment 3):

$$U_x^{(t)}(s) = H_x^{(t)}(s) + \frac{\sum_{j,k}|\Delta Q_x(j,k)|}{\Delta_{max}} \textbf{(5)}$$

where $x = MB\ or\ MF$, $H_x^{(t)}(s,a)$ is the entropy of the action probability distribution for system x computed with **Equation 3**, and Δ_{max} is the maximum possible variation of action values in the task. The meta-controller then decides which system $e \in \mathcal{E}$ (e for "expert" (Caluwaerts et al.,

2012; O'Doherty et al., 2020)) to rely on for the next action choice by comparing systems' uncertainty:

$$P^{(t)}(e|s) = \frac{exp^{\lambda U_x^{(t)}(s)}}{\sum_{k\in\mathcal{E}} exp^{\lambda U_k^{(t)}(s)}} \textbf{(6)}$$

where λ is the meta-controller's inverse temperature.

Random exploration system. Importantly, as in Dollé et al. (2018b), here the meta-controller does not choose between two systems only (MB or MF), but rather between three systems (MB, MF, EXP), where EXP is a random exploration generator. This has the advantages of avoiding the need to accumulate random exploration in both MB and MF system, and to produce clear decisions to explore rather than simply relying on an uncontrolled decisional noise. In any state of the tasks considered here, because we will always consider two alternative actions (magazine entry versus not moving), the EXP system always outputs a flat [0.5 0.5] action probability distribution.

In the following sections we will present four experimental findings from rodent lateral OFC that do not fit *a priori* model predictions derived from current MB cognitive map theories of overall OFC function (e.g. Wilson et al., 2014). We will then demonstrate how some of the RL model modifications we have described i.e. arbitration between learning systems, off-line learning, and attentional bias, can account for these experimental findings.

4. Experiment 1: Pre-training lateral OFC lesions enhance simple Pavlovian acquisition

4.1 Experimental results

While OFC lesions in rodents have often been reported to have no effect on simple Pavlovian acquisition, these studies have often stopped initial acquisition training after approximately around 9-12 days (Burke et al., 2008; Gallagher et al., 1999; McDannald et al., 2011; Ostlund & Balleine, 2007; Panayi & Killcross, 2018), and proceed with an experimental manipulation e.g. devaluation. In this initial acquisition period, it is not always clear that behaviour has reached asymptote. We have recently found that after training rats for 21 days to reach a stable behavioural asymptote, lateral OFC lesions significantly enhanced performance relative to sham-operated control animals (Figure 1A). Consistent with previous reports, there were no significant differences between groups over the first 9-12 days acquisition. These findings are not what might be predicted by current theories of OFC function (Delamater, 2007; Rudebeck & Murray, 2014; Wilson et al., 2014). Specifically, the experimental protocol involved a simple single auditory CS always followed by a pellet US, this CS-US contingency was stable, the value of the US did not change, and the identity of the predicted outcome was irrelevant to task performance.

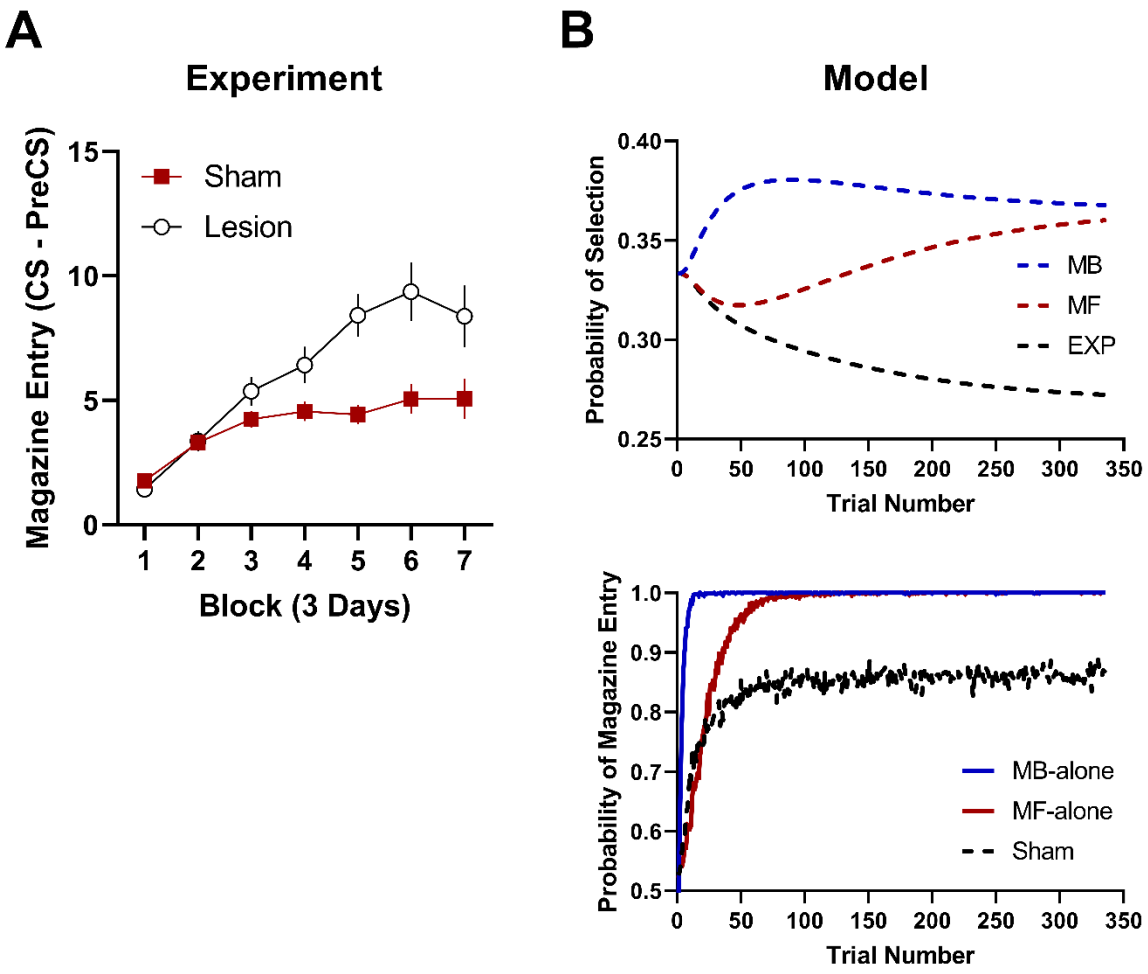

Figure 1. Pre-training lateral OFC lesions enhance simple Pavlovian acquisition. (A) Acquisition to a simple single Pavlovian CS-US relationship. Experimental parameters were a 15s auditory clicker CS immediately followed by the delivery of a grain pellet into a magazine receptacle, a total of 16 CS presentations per session with a variable inter-trial interval averaging 90s. Responding presented as CS-PreCS magazine frequency, i.e. rate of anticipatory approach to the magazine during the CS period minus the immediately preceding PreCS baseline period. Lateral OFC lesions did not significantly affect the rate of acquisition over the first 9 days (Blocks 1-3), but were significantly higher than sham control rats from days 10-21 (Blocks 4-7). Error bars represent ±SEM. Adapted from Figure 1 in Panayi & Killcross (2020) **(B)** Model simulation results. The top part shows the trial-by-trial evolution of the probability of selection of the model-based (MB) system, the model-free (MF) system, and the random exploration (EXP) system, when the arbitration mechanism is spared ('sham' model). The bottom part shows the probability of magazine entry in the sham model compared to those produced by MB-alone or MF-alone variants of the model. Both variants roughly reproduce the experimental results in (A).

4.2 Model simulation results

To account for this unexpected effect of OFC lesions on acquisition we explored a model where OFC function impairment is assumed to be mediated by an impaired arbitration mechanism between Pavlovian model-based (MB) reinforcement learning, Pavlovian model-free (MF) reinforcement learning, and a random exploration (EXP) system. The key hypothesis here is that a pre-training perturbation of the arbitration mechanism results in a single Pavlovian learning system operating during the whole task. In other words, we assume that the OFC lesion makes it unable to adaptively disinhibit particular systems during the task, so that the behaviour is controlled by a single learning process. Because we are agnostic about whether this single learning system would be MF or MB after an OFC lesion, we simulated both alternatives in comparison to normal arbitration. Figure 1B shows the simulation results for the three variants of the model. The top part shows how a normal arbitration mechanism ('sham' model) initially relies on the three systems (MB, MF and EXP), then gives dominance to the MB one for initial learning, while the MF system progressively increase its contribution as its performance slowly improves. Because the proposed arbitration mechanism slowly decreases the contribution of EXP, but never completely gives it up, the resulting performance curve (probability of magazine entry, at the bottom of Figure 1B) increases during about 50 trials and then converges to an asymptote around 0.85. This means that the simulated rats still occasionally explore after learning. In contrast, when the arbitration mechanism is blocked, the model learns with a single system (either MF or MB) which reaches an asymptote at 1 (optimal performance). Interestingly, the MB system alone learns faster than all model variants, because it is not perturbed by any competition with other systems. Strikingly, an MF system alone learns at a non-distinguishable speed than the 'sham' model but then stabilizes at a higher asymptote, similar to the experimental results (Figure 1A).

It is of note that here an MF-alone model is compatible with both the idea that OFC lesion impaired the arbitration mechanism or that it impaired the MB system, as in previous theories (Wilson et al., 2014). Moreover, it is interesting that in such a simple task, the results could also be accounted for by a spared MB system learning the task alone. In the next experiments, we will see that transiently impairing the arbitration mechanism after initial learning results in non-trivial arbitration dynamics that can help capture other experimental data.

5. Experiment 2: Post-training lateral OFC inactivation disrupts simple Pavlovian acquisition

5.1 Experimental results

Given that pre-training OFC lesions enhanced acquisition behaviour, we expected that post-training OFC dysfunction would also enhance acquisition. Surprisingly, post-training OFC inactivation (Figure 2A) and lesions (Panayi & Killcross, 2020)significantly impaired acquisition. Specifically, whereas control animals continued to acquire responding, responding did not change when OFC was inactivated (Session 12-15). Finally, when OFC function is returned (Session 16-17), impaired responding is recovered and no different to the control group. This might suggest that OFC inactivation may have disrupted the behavioural expression but not underlying learning during acquisition in this task. The computational results presented hereafter suggest that this can also be interpreted in terms of a transiently disrupted arbitration mechanism while learning in the MB system was spared.

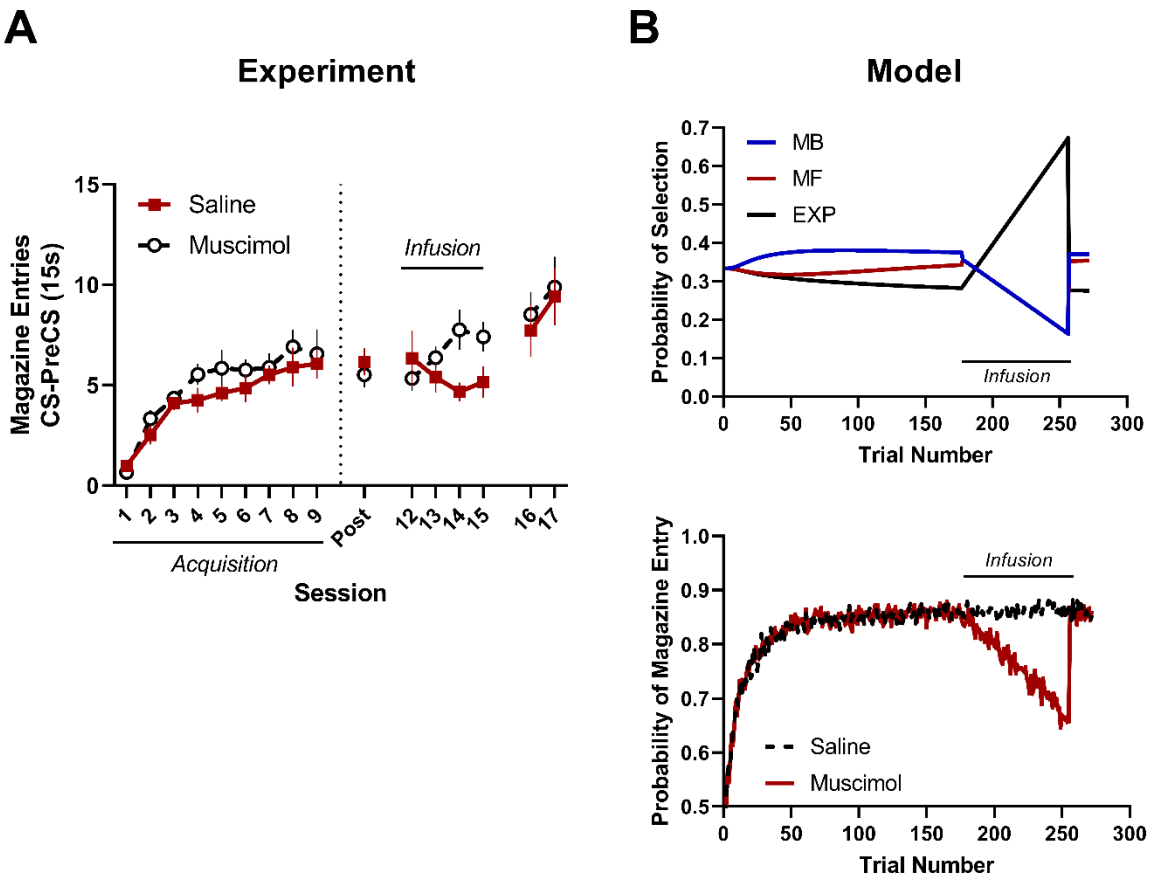

Figure 2. Post-training lateral OFC inactivation disrupts simple Pavlovian acquisition. (A) Acquisition to a simple single Pavlovian CS-US relationship identical to parameters in Figure 1. After 9 days of acquisition (Session 1-9), cannulae targeting lateral OFC were implanted, and animals were given brief re-training (2 days) following post-operative recovery (Post). Next, acquisition was tested following OFC inactivation (Session 12-15) and return of function (no infusions; Session 16-17). Control animals (Saline) continued to acquire responding, whereas OFC inactivation prevented further increases in responding (Session 12-15). The effect of OFC inactivation was no longer detected when OFC function was returned (Session 16-17). Error bars represent ±SEM. Adapted from Figure 2 in Panayi & Killcross (2020) **(B)** Model simulation results. The top part shows the trial-by-trial evolution of the probability of selection of the model-based (MB) system, the model-free (MF) system, and the random exploration (EXP) system, when the arbitration mechanism is spared ('sham' model). The bottom part compares the probability of magazine entry in the sham model compared to a variant where the arbitration mechanism is perturbed during about 100 trial (*i.e.,* inhibition of MB and MF systems' output while increasing random exploration).

5.2 Model simulation results

To account for the disruption of acquisition following OFC dysfunction we explored two variants of the model: one in which the arbitration mechanism and all Pavlovian learning systems are

intact ('sham' model), and one in which the arbitration mechanism is perturbed during about 100 trials and then restored, while all individual Pavlovian learning systems have been preserved. Figure 2B shows the simulation results. Like in Experiment 1, the sham model quickly leaves the control over behaviour most of the time to the MB system, while EXP slowly decreases, and MF slowly improves. When the arbitration mechanism is perturbed, because its operations are required to maintain the right proportions of MB, MF and EXP (in contrast to the pre-training situation of Experiment 1), a single system cannot immediately take over. In contrast, we assume that the infusion of muscimol is here progressively pushing the model to inhibit the currently winning systems' output (MB and MF), thus relying more and more on random exploration (EXP). As a consequence, the number of magazine entries produced by the model decreases, like in the experimental data. When the infusion stops, because the MB and MF systems had been spared, the performance can instantaneously be restored, again like in the experimental data.

One slight difference between the simulation results and the experimental results is that the performance further increases after muscimol infusion (sessions 16-17 in Figure 2A) while it returns to asymptote after 250 trials in Figure 2B. Since we used the same model parameters as in Experiment 1, assuming that the experimental conditions are similar, the performance asymptote should theoretically have been reached between trials 50 and 100, thus around sessions 4-6, so that performance cannot further improve after trial 250. Nevertheless, the model here neglects motivational or attentional factors which may have pushed animals to make more magazine entries during sessions 16-17.

6. Experiment 3: Lateral OFC inactivation disrupts simple Pavlovian extinction

6.1 Experimental results

The observation of impaired Pavlovian acquisition are consistent with reports of impaired extinction in reversal learning tasks (Izquierdo, 2017b), and simple Pavlovian extinction procedures (Lay et al., 2020; Panayi & Killcross, 2014; Zimmermann et al., 2018) following OFC dysfunction. We have previously shown that OFC inactivation disrupts extinction learning over multiple sessions, however within each extinction session OFC inactivation did not prevent extinction behaviour (i.e. decreasing responding). Indeed, within-session extinction appeared more rapidly under OFC inactivation (Figure 3A). This demonstrates clear behavioural flexibility within a session, but an inability to update or consolidate behaviour and/or learning between-sessions. Only after OFC function was returned were OFC inactivated animals able to demonstrate appropriate extinction learning between sessions. Notably, current RL models of OFC function predict that both between- and within-session extinction should be impaired, with performance eventually extinguishing over trials at a much slower rate than controls (Wilson et al., 2014). Therefore, it is important to reconcile these findings within an RL model of OFC function.

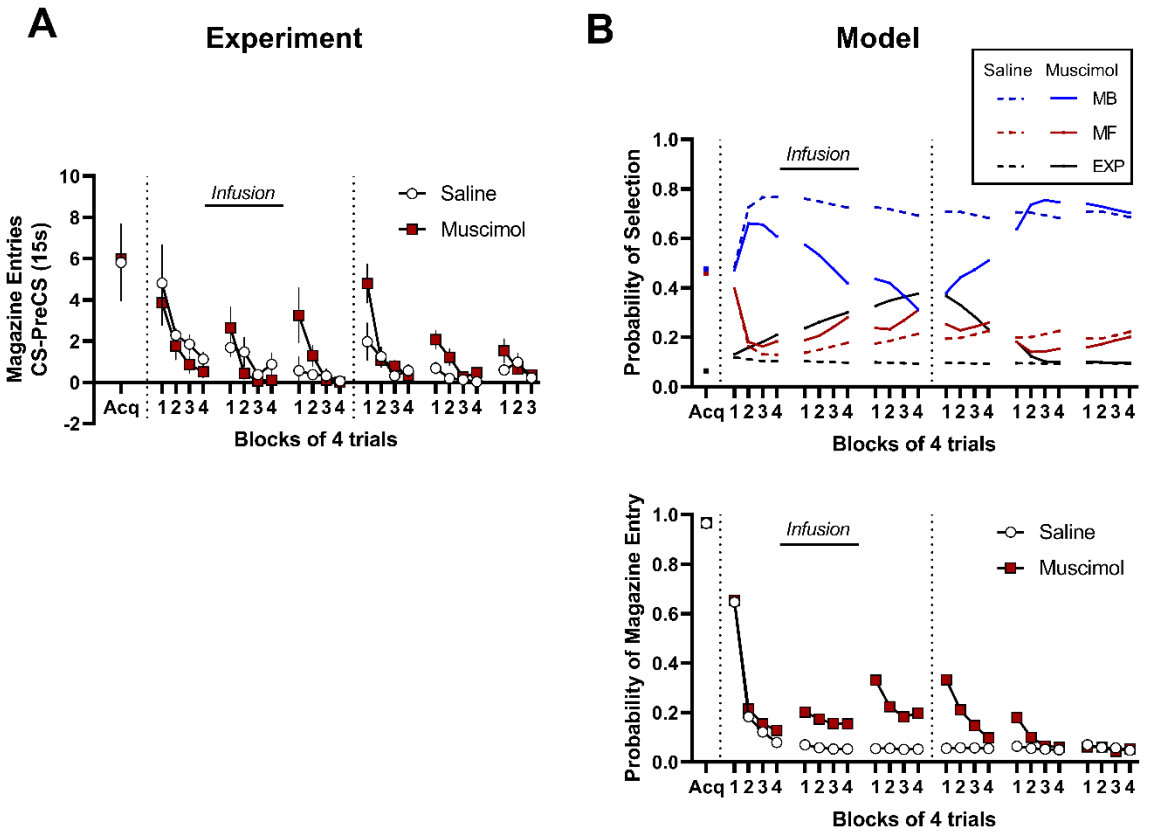

Figure 3. Lateral OFC inactivation disrupts between-session Pavlovian extinction. (A) Acquisition and then extinction of a simple single Pavlovian CS-US relationship. Acquisition parameters were identical to those described in Figure 2; after 9 days of acquisition, cannulae targeting lateral OFC were implanted, and animals were given brief re-training (2 days) following post-operative recovery (Acq). Next, animals received 6 days of extinction with a drug infusion prior to the first 3 days (Infusion; Day 1-3) but not the last 3 days (No Infusion; Day 4-6). OFC inactivation impaired between-session extinction (Days 1-3), but not within-session extinction. Error bars represent ±SEM. Adapted from Figure 2 in Panayi & Killcross, 2014). **(B)** Model simulation results. The top part shows the trial-by-trial evolution of the probability of selection of the model-based (MB) system, the model-free (MF) system, and the random exploration (EXP) system, when the arbitration mechanism is spared ('saline' model; dashed lines), versus when it is perturbed ('muscimol' model; plain lines). The bottom part shows the difference in the within- and between-session dynamics of extinction between the two variants of the model.

6.2 Model simulation results

To account for the disruption of between-session extinction following OFC inactivation we explored again a model where the post-training arbitration status is perturbed, inhibiting the model-based system (which is here largely dominant to account for the fast behavioural extinction during the very first day of infusion) and making behaviour more random. It is important to note that here, if the perturbation of the arbitration had inhibited both MB and MF systems while favouring only random exploration, a sharp degradation of performance

would have been produced during days 2 to 4. This is because a modelled pure random exploration would have had a probability 0.5 to enter the magazine. In contrast, because here the decrease in the probability of selection of the winning MB system was accompanied by an increase of the probabilities of both MF and EXP, the degradation of performance is only mild (Figure 3B), consistent with the experimental results.

To explain the day-to-day (between sessions) inability of rats to consolidate the extinguished behaviour, we consider here that the perturbed arbitration mechanism does not allow the OFC to trigger timely off-line MB inference to compensate for the forgetting of action values (Equation 4) which occurs overnight in the MF system. Nevertheless, because the disturbed arbitration mechanism decreases the contribution of MB, while still relying on it about 40% of the time during task performance, rapid within session extinction is observed during days 3 and 4 in the simulation results (Figure 3B).

Finally, when the infusion stops, the arbitration returns to normal, favouring the MB system again, which promotes full extinction as observed in the experimental results.

7. Experiment 4: Lateral OFC inactivation during acquisition does not impair subsequent associative blocking

7.1 Experimental results

One possible account of impaired simple acquisition following OFC inactivation (Figure 2) is that learning about the CS-US relationship has been disrupted. To test this possibility, we employed a Pavlovian associative blocking procedure, a procedure commonly used to test prediction error learning (Nasser et al., 2017; Steinberg et al., 2013). In a blocking experiment (Figure 4A), first an animal is trained such that a cue (cue A) predicts an outcome (pellet). Next, A is presented in compound with a novel cue (cue B) which also leads to the same pellet outcome. If the animal has learned that cue A sufficiently predicts the pellet outcome already, then very little is learned about cue B i.e. learning about cue A blocks subsequent learning about cue B (Kamin, 1969; Rescorla & Wagner, 1972). However, if learning about cue A is insufficient, then learning about cue B should not be blocked. We predicted that if OFC inactivation is disrupting learning, then OFC inactivation during initial learning about cue A should disrupt the blocking effect.
We found that while OFC inactivation during acquisition of cue A significantly disrupted behaviour (Figure 4B), cue A was still able to effectively block learning to cue B (Figure 4D). This suggests that the impaired acquisition behaviour observed following OFC inactivation did not reflect impaired learning about the CS-US relationship necessary for associative blocking.

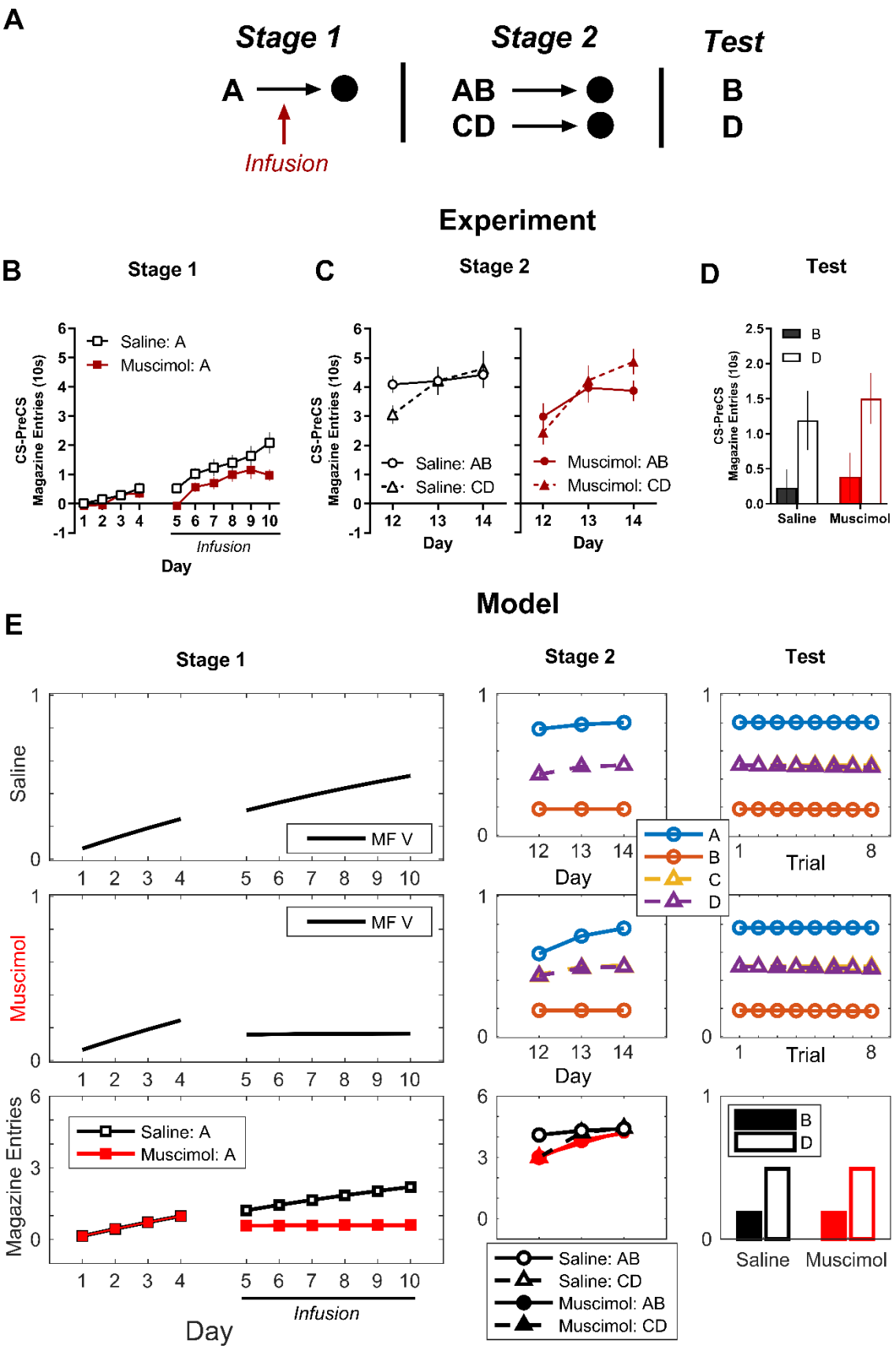
A
Stage 1
Stage 2
Test
A
AB
CD
B
D
Infusion
Experiment
B
Stage 1
C
Stage 2
D
Test
CS-PreCS Magazine Entries (10s)
Saline: A
Muscimol: A
Infusion
Day
Saline: AB
Saline: CD
Muscimol: AB
Muscimol: CD
Saline
Muscimol
E
Model
Stage 1
Stage 2
Test
Saline
Muscimol
MF V
Magazine Entries
Trial
A
B
C
D

Figure 4. Impaired Pavlovian acquisition behaviour following lateral OFC inactivation does not disrupt subsequent associative blocking. The effect of OFC inactivation during acquisition on subsequent learning in a Pavlovian blocking design. **(A)** The design used to achieve blocking of learning to cue B during stage 2 by pre-training cue A in stage 1. OFC infusions of saline or muscimol were performed during stage 1 after the first 4 days of initial acquisition to cue A. Cues A and C were always visual cues, either darkness caused by extinguishing the houselight or flashing panel lights (5Hz). Cues B and D were always auditory cues, either an 80dB white noise or a 5Hz train of clicks. All cues lasted 10s, and reward was always a single food pellet. Cannulae placements depicted in Figure 3-figure supplement 3. **(B)** Pavlovian acquisition to cue A over 10 days, with intact OFC (days 1-4) and following infusion of saline or muscimol to functionally inactivate the OFC (days 5-10). Muscimol infusions significantly suppressed responding to cue A. **(C)** Performance during stage 2 of blocking to cue compounds AB and CD in the saline (left) and muscimol (right) infusion groups. A focused analysis of responding within Day 12 is presented in Figure 3-figure supplement 4. **(D)** Responding during an extinction test to "blocked" cue B and the overshadowing control cue D. Significantly reduced responding to cue B relative to cue D indicates that learning about cue A effectively blocked subsequent learning to cue B in both the muscimol and saline groups. Pavlovian responding quantified by the rate of discriminative responding (CS-PreCS). Error bars depict ± SEM. Adapted from Figure 3 in Panayi & Killcross (2020) **(E)** Model simulation results. The top panel shows the evolution of the value of cues learned by the MF system in the 'saline' control model. The middle panel shows the same simulation for the 'muscimol' model, where the MB compensation for day-to-day forgetting in the MF model is impaired. The bottom panel shows the resulting numbers of magazine entries in both models during the different phases of the task.

7.2 Model simulation results

To account for the dissociation between Pavlovian acquisition and associative blocking following OFC inactivation we again simulated a model where we assumed that the arbitration mechanism was perturbed by the muscimol injections, so that the model is unable to compensate for day-to-day forgetting with off-line MB inference (as in Experiment 3). Nevertheless, in contrast to the model simulated in Experiments 1-3 above, here we simulated a simpler version where CS-US relationships with multiple cues are only learnt through the combination of MF and MB learning, without the contribution of a random exploration system.

In this case, because OFC inactivation is assumed to have disabled the MB compensation for forgetting, the CS value learnt by the MF system is decreased, so that the number of magazine entries in response to cue A is reduced (Figure 4E). As a consequence, the 'muscimol' model starts Day 12 with decreased responding to the AB compound, compared to the 'saline' model in which the arbitration mechanism is spared. Nevertheless, the contribution of the MB system makes learning rapid so that the 'muscimol' model does not respond less to AB anymore on Day 14, as in the experimental results (Figure 4C). Importantly, because we assumed an attentional mechanism (Mackintosh, 1975) so that the initially trained cue A catches all the attention of the model at the expense of cue B, the latter is blocked, as in the experimental results (Figure 4D).

Importantly, without such an attentional mechanism the rapid increase in responding to compound AB by the model would lead cue B to acquire at least some value, so that it cannot be fully blocked. Thus, these results highlight the importance of attentional mechanisms in Pavlovian conditioning (Esber & Haselgrove, 2011; Mackintosh, 1975; Pearce & Hall, 1980; Roesch et al., 2010), in addition to the OFC's possible role in Pavlovian system arbitration.

8. Discussion

OFC dysfunction has been successfully modelled as an impairment in MB inferences resulting from disruption of the formation of latent states necessary for a detailed cognitive map of task space (Wilson et al., 2014). While this function effectively accounts for diverse experimental findings relating to the overall function of the OFC, we have recently found that dysfunction specific to the rodent lateral OFC causes a complex pattern of deficits in simple acquisition and extinction learning that is not clearly predicted by these RL theories (Panayi & Killcross, 2014, 2020). Here we propose modifications to these RL models that can account for these findings. Indeed, there is an emerging understanding in the field that our current models of the OFC need to be modified and refined to account for a more nuanced role for the OFC in MB learning (Gardner & Schoenbaum, 2020). Specifically, we suggest that the role of the rodent lateral OFC in the formation and use of MB cognitive maps of task space is as an arbitrator between MB and MF learning systems.

During initial task learning, an organism cannot know whether the task involves complex or simple stimulus-action-outcome contingencies. However, as the organism gains experience, particularly in a simple deterministic environment, an optimal trade-off between more complex (MB) and simple (MF) learning systems is likely to develop i.e. the optimal relative dominance of these systems will depend upon the complexity of task demands. Our simulations suggest that the role of the lateral OFC in initial acquisition can be modelled as an arbitrator of exploration between these systems that normally develops over the course of learning. We also consider a role for the rodent lateral OFC in consolidation via updating of MF values through MB inferences offline between learning events-sessions. These modifications make explicit the implicit understanding that there are numerous psychological processes that underlie even simple learning procedures that are often implicitly acknowledged by researchers.

In our models we suggest that the OFC is indeed critical for MB inferences and the construction of a cognitive map (Bradfield & Hart, 2020; Niv, 2019; Sharpe et al., 2019; Wilson et al., 2014), however at the level of the rodent lateral OFC this is achieved by arbitration between MB and MF system control during learning. Therefore, the effects of lateral OFC dysfunction are predicted to interact with the underlying psychological demands and complexity of a task. For example, in a simple Pavlovian acquisition design, biasing learning systems from the start of training in favour of either MB or MF results in one system dominating learning and behaviour and enhancing Pavlovian approach. However, once arbitration between MB and MF has reached an equilibrium following initial learning, post-training OFC inactivation can significantly disrupt this equilibrium and reinstate behaviours that are no longer appropriate. A role for the lateral OFC in arbitration might also account for the high degree of diversity of task signals

represented within the lateral OFC, a common target of rodent electrophysiological recordings, representing aspects of MB and MF states, actions, and values (Ogawa et al., 2013; Padoa-Schioppa, 2009; Sadacca et al., 2018; Stalnaker et al., 2014; Takahashi et al., 2013; Zhou et al., 2019).

It is also important to highlight that the role of the rodent lateral OFC in the arbitration between learning systems is restricted to Pavlovian (cue-outcome) and not instrumental (action-outcome) learning. The distinction between instrumental and Pavlovian learning systems is often not considered in within RL theories but can be a critical psychological distinction. For example, lateral OFC lesions significantly impair Pavlovian outcome devaluation, but do not appear to disrupt instrumental outcome devaluation (Ostlund & Balleine, 2007; Panayi & Killcross, 2018; Pickens et al., 2005), a finding that cannot be reconciled without considering this distinction. In contrast, rodent medial OFC dysfunction significantly disrupts instrumental but not Pavlovian outcome devaluation procedures (Bradfield et al., 2015; Gardner et al., 2018). Similarly, OFC lesions impair the ability to perform intradimensional shifts (Kim & Ragozzino, 2005; McAlonan & Brown, 2003), as opposed to extradimensional shifts which involve the prelimbic cortex (Joel et al., 1997; Birrell and Brown, 2000; Ragozzino et al., 2003). This highlights that multiple regions of the prefrontal cortex may be required for different types of rule shifts (Haddon & Killcross, 2006; Sharpe & Killcross, 2018). Thus, in addition to the frontopolar cortex and the inferior lateral prefrontal cortex which have been found to play a role in MB-MF arbitration during instrumental tasks in humans (Lee et al., 2014), here we further suggest that the lateral OFC may also play an important role in the MB-MF arbitration when it comes to the Pavlovian domain.

There has been an important focus recently on refining anatomical specificity when exploring medial prefrontal and orbitofrontal cortical function (Barreiros et al., 2021; Coutureau & Killcross, 2003; Killcross & Coutureau, 2003; Laubach et al., 2018). This has been driven by an emerging picture of functional heterogeneity within these cortical subregions. However, while the emerging picture shows functional heterogeneity and dissociations in anatomically adjacent substructures (Bradfield & Hart, 2020; Izquierdo, 2017a; Panayi & Killcross, 2018), there is also a remarkable consistency of overall functional purpose tying these regions together (Roesch & Schoenbaum, 2006; Rudebeck & Murray, 2011a; Wilson et al., 2014). Here we suggest that in parallel to the emerging experimental literature, we must consider refined models of OFC subregion functions that account for these unique heterogeneous results, but also maintain a coherent computational role across the entire OFC.

Competing Interests

The authors declare no competing interests.

References

Arleo, A., & Gerstner, W. (2000). Spatial cognition and neuro-mimetic navigation: a model of hippocampal place cell activity. *Biological Cybernetics*, *83*(3), 287–299.

Barreiros, I. V, Ishii, H., Walton, M. E., & Panayi, M. C. (2021). Defining an orbitofrontal compass: functional and anatomical heterogeneity across anterior-posterior and medial-lateral axes. *Behavioral Neuroscience*, *under review*.

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*, *100*(2), 490–509. https://doi.org/10.1016/j.neuron.2018.10.002

Boakes, R. A. (1977). Performance on learning to associate a stimulus with positive reinforcement. In H. Davis & H. M. B. Hurwitz (Eds.), *Operant-Pavlovian interactions* (pp. 67–97). L. Erlbaum Associates.

Boulougouris, V., Dalley, J. W., & Robbins, T. W. (2007). Effects of orbitofrontal, infralimbic and prelimbic cortical lesions on serial spatial reversal learning in the rat. *Behavioural Brain Research*, *179*(2), 219–228. https://doi.org/10.1016/j.bbr.2007.02.005

Bradfield, L. A., Dezfouli, A., Van Holstein, M., Chieng, B., & Balleine, B. W. (2015). Medial Orbitofrontal Cortex Mediates Outcome Retrieval in Partially Observable Task Situations. *Neuron*, *88*(6), 1268–1280. https://doi.org/10.1016/j.neuron.2015.10.044

Bradfield, L. A., & Hart, G. (2020). Rodent medial and lateral orbitofrontal cortices represent unique components of cognitive maps of task space. *Neuroscience & Biobehavioral Reviews*, *108*, 287–294. https://doi.org/10.1016/J.NEUBIOREV.2019.11.009

Burke, K. A., Franz, T. M., Miller, D. N., & Schoenbaum, G. (2008). The role of the orbitofrontal cortex in the pursuit of happiness and more specific rewards. *Nature*, *454*(7202), 340-U45. https://doi.org/Doi 10.1038/Nature06993

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., & Rosvold, H. E. (1963). Conditioning and extinction of a food-rewarded response after selective ablations of frontal cortex in rhesus monkeys. *Experimental Neurology*, *7*(1), 65–75. https://doi.org/10.1016/0014-4886(63)90094-3

Caluwaerts, K., Staffa, M., N'Guyen, S., Grand, C., Dollé, L., Favre-Félix, A., Girard, B., & Khamassi, M. (2012). A biologically inspired meta-control navigation system for the psikharpax rat robot. *Bioinspiration & Biomimetics*, *7*(2), 25009.

Cazé, R., Khamassi, M., Aubin, L., & Girard, B. (2018). Hippocampal replays under the scrutiny of reinforcement learning models. *Journal of Neurophysiology*, *120*(6), 2877–2896.

Chang, S. E. (2014). Effects of orbitofrontal cortex lesions on autoshaped lever pressing and reversal learning. *Behavioural Brain Research*, *273*, 52–56. https://doi.org/10.1016/j.bbr.2014.07.029

Chudasama, Y., & 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. *Journal of Neuroscience*, *23*(25), 8771–8780. https://doi.org/23/25/8771 [pii]

Cinotti, F., Fresno, V., Aklil, N., Coutureau, E., Girard, B., Marchand, A. R., & Khamassi, M. (2019). Dopamine blockade impairs the exploration-exploitation trade-off in rats. *Scientific*

Reports, *9*(1). https://doi.org/10.1038/s41598-019-43245-z
Cinotti, F., Marchand, A. R., Roesch, M. R., Girard, B., & Khamassi, M. (2019). Impacts of inter-trial interval duration on a computational model of sign-tracking vs. goal-tracking behaviour. *Psychopharmacology*, *236*(8). https://doi.org/10.1007/s00213-019-05323-y
Collins, A. G. E., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*. https://doi.org/10.1038/s41583-020-0355-6
Coutureau, E., & Killcross, A. S. (2003). Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. *Behavioural Brain Research*, *146*(1–2), 167–174. http://www.sciencedirect.com/science/article/pii/S0166432803003498
Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027
Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. https://doi.org/10.1038/nn1560
Daw, N. D., & O'Doherty, J. P. (2014). Multiple systems for value learning. In *Neuroeconomics* (pp. 393–410). Elsevier.
Dayan, P., Niv, Y., Seymour, B., & Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural Networks*, *19*, 1153–1160. https://doi.org/10.1016/j.neunet.2006.03.002
Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From Creatures of Habit to Goal-Directed Learners: Tracking the Developmental Emergence of Model-Based Reinforcement Learning. *Psychological Science*, *27*(6), 848–858. https://doi.org/10.1177/0956797616639301
Delamater, A. R. (2004). Experimental extinction in Pavlovian conditioning: Behavioural and neuroscience perspectives. *Quarterly Journal of Experimental Psychology Section B-Comparative and Physiological Psychology*, *57*(2), 97–132. https://doi.org/Doi 10.1080/02724990344000097
Delamater, A. R. (2007). The role of the orbitofrontal cortex in sensory-specific encoding of associations in Pavlovian and instrumental conditioning. In G. Schoenbaum, J. A. Gottfried, E. A. Murray, & S. J. Ramus (Eds.), *Linking Affect to Action: Critical Contributions of the Orbitofrontal Cortex* (Vol. 1121, pp. 152–173). Blackwell Publishing. https://doi.org/10.1196/annals.1401.030
Delamater, A. R. (2012). On the nature of CS and US representations in Pavlovian learning. *Learn Behav*, *40*(1), 1–23. https://doi.org/10.3758/s13420-011-0036-4
Delamater, A. R., & Oakeshott, S. (2007). Learning about multiple attributes of reward in Pavlovian conditioning. *Annals of the New York Academy of Sciences*. https://doi.org/10.1196/annals.1390.008
Dickinson, A., & Balleine, B. (1995). Motivational Control of Instrumental Action. *Current Directions in Psychological Science*, *4*(5), 162–167. https://doi.org/10.1111/1467-8721.ep11512272
Dollé, L., Chavarriaga, R., Guillot, A., & Khamassi, M. (2018a). Interactions of spatial strategies producing generalization gradient and blocking: A computational approach. *PLoS Computational Biology*, *14*(4), e1006092.
Dollé, L., Chavarriaga, R., Guillot, A., & Khamassi, M. (2018b). Interactions of spatial strategies

producing generalization gradient and blocking: A computational approach. *PLoS Computational Biology*, *14*(4). https://doi.org/10.1371/journal.pcbi.1006092

Dollé, L., Khamassi, M., Girard, B., Guillot, A., & Chavarriaga, R. (2008). Analyzing interactions between navigation strategies using a computational model of action selection. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 5248 LNAI*. https://doi.org/10.1007/978-3-540-87601-4_8

Dollé, L., Sheynikhovich, D., Girard, B., Chavarriaga, R., & Guillot, A. (2010). Path planning versus cue responding: A bio-inspired model of switching between navigation strategies. *Biological Cybernetics*, *103*(4), 299–317. https://doi.org/10.1007/s00422-010-0400-z

Erev, I., & Haruvy, E. (2016). 10. Learning and the Economics of Small Decisions. *The Handbook of Experimental Economics, Volume Two*, 1–136. https://doi.org/10.1515/9781400883172-011

Esber, G. R., & Haselgrove, M. (2011). Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proceedings of the Royal Society B-Biological Sciences*, *278*(1718), 2553–2561. https://doi.org/DOI 10.1098/rspb.2011.0836

Friston, K. J., Lin, M., Frith, C. D., Pezzulo, G., Hobson, J. A., & Ondobaka, S. (2017). Active Inference, Curiosity and Insight. *Neural Computation*, *29*, 2633–2683. https://doi.org/10.1162/NECO

Gallagher, M., McMahan, R. W., & Schoenbaum, G. (1999). Orbitofrontal cortex and representation of incentive value in associative learning. *Journal of Neuroscience*, *19*(15), 6610–6614. http://www.jneurosci.org/cgi/reprint/19/15/6610.pdf

Gardner, M. P. H., Conroy, J. C., Styer, C. V, Huynh, T., Whitaker, L. R., & Schoenbaum, G. (2018). Medial orbitofrontal inactivation does not affect economic choice. *ELife*, *7*. https://doi.org/10.7554/eLife.38963

Gardner, M. P. H., & Schoenbaum, G. (2020). The orbitofrontal cartographer. *PsyArxiv*.

Genzel, L., Schut, E., Schröder, T., Eichler, R., Khamassi, M., Gomez, A., Lobato, I. N., & Battaglia, F. (2019). The object space task shows cumulative memory expression in both mice and rats. *PLoS Biology*, *17*(6). https://doi.org/10.1371/journal.pbio.3000322

Haddon, J. E., & Killcross, S. (2006). Prefrontal cortex lesions disrupt the contextual control of response conflict. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.3243-05.2006

Hall, G. (2002). Associative structures in Pavlovian and instrumental conditioning. In C. R. Gallistel (Ed.), *Steven's handbook of experimental psychology* (Vol. 3, pp. 1–45). John Wiley & Sons.

Hart, E. E., Sharpe, M. J., Gardner, M. P. H., & Schoenbaum, G. (2020). Responding to preconditioned cues is devaluation sensitive and requires orbitofrontal cortex during cue-cue learning. *ELife*. https://doi.org/10.7554/ELIFE.59998

Izquierdo, A. D. (2017a). Functional heterogeneity within rat orbitofrontal cortex in reward learning and decision making. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.1678-17.2017

Izquierdo, A. D. (2017b). Functional Heterogeneity within Rat Orbitofrontal Cortex in Reward Learning and Decision Making. *The Journal of Neuroscience : The Official Journal of the*

Society for Neuroscience, *37*(44), 10529–10540. https://doi.org/10.1523/JNEUROSCI.1678-17.2017

Izquierdo, A. D., & Murray, E. A. (2000). Bilateral orbital prefrontal cortex lesions disrupt reinforcer devaluation effects in rhesus monkeys. *Society for Neuroscience Abstracts*, *26*, 978.

Izquierdo, A. D., & Murray, E. A. (2004). Combined unilateral lesions of the amygdala and orbital prefrontal cortex impair affective processing in rhesus monkeys. *Journal of Neurophysiology*, *91*(5), 2023–2039. https://doi.org/10.1152/jn.00968.2003

Izquierdo, A. D., & Murray, E. A. (2005). Opposing effects of amygdala and orbital prefrontal cortex lesions on the extinction of instrumental responding in macaque monkeys. *European Journal of Neuroscience*, *22*(9), 2341–2346. https://doi.org/EJN4434 [pii] 10.1111/j.1460-9568.2005.04434.x

Izquierdo, A. D., Suda, R. K., & Murray, E. A. (2004). Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. *Journal of Neuroscience*, *24*, 7540–7548. http://www.jneurosci.org/content/24/34/7540.full.pdf

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor-critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks*, *15*(4–6), 535–547. https://doi.org/10.1016/S0893-6080(02)00047-3

Johnson, A., & Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, *27*(45), 12176–12189. https://doi.org/10.1523/JNEUROSCI.3761-07.2007

Jones, J. L., Esber, G. R., McDannald, M. A., Gruber, A. J., Hernandez, G., Mirenzi, A., & Schoenbaum, G. (2012). Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science*, *338*, 953–956. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3592380/pdf/nihms446679.pdf

Kamin, L. J. (1969). Predictability, surprise, attention and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279–96). Appleton-Century-Crofts.

Kanahara, N., Sekine, Y., Haraguchi, T., Uchida, Y., Hashimoto, K., Shimizu, E., & Iyo, M. (2013). Orbitofrontal cortex abnormality and deficit schizophrenia. *Schizophrenia Research*, *143*(2–3), 246–252. https://doi.org/10.1016/j.schres.2012.11.015

Kato, A., & Morita, K. (2016). Forgetting in Reinforcement Learning Links Sustained Dopamine Signals to Motivation. *PLoS Computational Biology*, *12*(10), 1–41. https://doi.org/10.1371/journal.pcbi.1005145

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology*, *7*(5).

Keramati, M., & Gutkin, B. (2014). Homeostatic reinforcement learning for integrating reward collection and physiological stability. *ELife*, *3*, 1–26. https://doi.org/10.7554/eLife.04811

Khamassi, M. (2020). *Adaptive coordination of multiple learning strategies in brains and robots*. 1–20.

Khamassi, M., & Humphries, M. D. (2012). Integrating cortico-limbic-basal ganglia architectures for learning model-based and model-free navigation strategies. *Frontiers in Behavioral Neuroscience*, *OCTOBER 2012*. https://doi.org/10.3389/fnbeh.2012.00079

Khamassi, M., Lachèze, L., Girard, B., Berthoz, A., & Guillot, A. (2005). Actor-critic models of

reinforcement learning in the basal ganglia: From natural to artificial rats. *Adaptive Behavior*, *13*(2). https://doi.org/10.1177/105971230501300205

Khamassi, M., Mulder, A. B., Tabuchi, E., Douchamps, V., & Wiener, S. I. (2008). Anticipatory reward signals in ventral striatal neurons of behaving rats. *European Journal of Neuroscience*, *28*(9). https://doi.org/10.1111/j.1460-9568.2008.06480.x

Khamassi, M., Quilodran, R., Enel, P., Dominey, P. F., & Procyk, E. (2015). Behavioral regulation and the modulation of information coding in the lateral prefrontal and cingulate cortex. *Cerebral Cortex*, *25*(9). https://doi.org/10.1093/cercor/bhu114

Killcross, A. S., & Blundell, P. (2002). Associative representations of emotionally significant outcomes. In S. C. Moore & M. Oaksford (Eds.), *Emotional Cognition: From brain to behaviour* (Vol. 44, pp. 35–74). John Benjamins Publishing Company.

Killcross, A. S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, *13*(4), 400–408. http://cercor.oxfordjournals.org/cgi/reprint/13/4/400.pdf

Kim, J., & Ragozzino, K. E. (2005). The involvement of the orbitofrontal cortex in learning under changing task contingencies. *Neurobiology of Learning and Memory*, *83*, 125–133. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3206595/pdf/nihms-333784.pdf

Kober, J., Bagnell, A. J., & Peters, J. (2014). Reinforcement Learning in Robotics: A Survey. *Springer Tracts in Advanced Robotics*, *97*, 9–67. https://doi.org/10.1007/978-3-319-03194-1_2

Konidaris, G., & Barto, A. (2006). An adaptive robot motivational system. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4095 LNAI*, 346–356. https://doi.org/10.1007/11840541_29

Krettek, J. E., & Price, J. L. (1977). Cortical Projections of Mediodorsal Nucleus and Adjacent Thalamic Nuclei in Rat. *Journal of Comparative Neurology*, *171*(2), 157–191. https://doi.org/DOI 10.1002/cne.901710204

Laubach, M., Amarante, L. M., Swanson, K., & White, S. R. (2018). What, if anything, is rodent prefrontal cortex? In *eNeuro*. https://doi.org/10.1523/ENEURO.0315-18.2018

Lay, B. P. P., Pitaru, A. A., Boulianne, N., Esber, G. R., & Iordanova, M. D. (2020). Different methods of fear reduction are supported by distinct cortical substrates. *ELife*. https://doi.org/10.7554/eLife.55294

Lee, S., Shimojo, S., & O'Doherty, J. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, *81*(3), 687–699. https://doi.org/10.1016/j.neuron.2013.11.028.Neural

Lesaint, F., Sigaud, O., Clark, J. J., Flagel, S. B., & Khamassi, M. (2015). Experimental predictions drawn from a computational model of sign-trackers and goal-trackers. *Journal of Physiology Paris*, *109*(1–3). https://doi.org/10.1016/j.jphysparis.2014.06.001

Lesaint, F., Sigaud, O., Flagel, S. B., Robinson, T. E., & Khamassi, M. (2014). Modelling Individual Differences in the Form of Pavlovian Conditioned Approach Responses: A Dual Learning Systems Approach with Factored Representations. *PLoS Computational Biology*, *10*(2). https://doi.org/10.1371/journal.pcbi.1003466

Lucantonio, F., Gardner, M. P. H., Mirenzi, A., Newman, L. E., Takahashi, Y. K., & Schoenbaum, G. (2015). Neural Estimates of Imagined Outcomes in Basolateral Amygdala Depend on Orbitofrontal Cortex. *Journal of Neuroscience*, *35*(50), 16521–16530.

https://doi.org/10.1523/JNEUROSCI.3126-15.2015
Machado, C. J., & Bachevalier, J. (2007). The effects of selective amygdala, orbital frontal cortex or hippocampal formation lesions on reward assessment in nonhuman primates. *European Journal of Neuroscience*, *25*(9), 2885–2904. https://doi.org/10.1111/j.1460-9568.2007.05525.x
Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychol Rev*, *82*(4), 279–298. https://doi.org/10.1037/h0076778
Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, *21*(11), 1609–1617.
McAlonan, K., & Brown, V. J. (2003). Orbital prefrontal cortex mediates reversal learning and not attentional set shifting in the rat. *Behavioural Brain Research*, *146*(1–2), 97–103. https://doi.org/S0166432803003437 [pii]
McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y., & Schoenbaum, G. (2011). Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *Journal of Neuroscience*, *31*, 2700–2705. http://www.jneurosci.org/content/31/7/2700.full.pdf
McDannald, M. A., Saddoris, M. P., Gallagher, M., & Holland, P. C. (2005). Lesions of orbitofrontal cortex impair rats' differential outcome expectancy learning but not conditioned stimulus-potentiated feeding. *Journal of Neuroscience*, *25*(18), 4626–4632. https://doi.org/25/18/4626 [pii] 10.1523/JNEUROSCI.5301-04.2005
McEnaney, K. W., & Butter, C. M. (1969). Perseveration of responding and nonresponding in monkeys with orbital frontal ablations. *J Comp Physiol Psychol*, *68*(4), 558–561. http://www.ncbi.nlm.nih.gov/pubmed/4981118
Moorman, D. E., & Aston-Jones, G. (2014). Orbitofrontal Cortical Neurons Encode Expectation-Driven Initiation of Reward-Seeking. *Journal of Neuroscience*, *34*(31), 10234–10246. https://doi.org/Doi 10.1523/Jneurosci.3216-13.2014
Murray, E. A., Moylan, E. J., Saleem, K. S., Basile, B. M., & Turchi, J. (2015). Specialized areas for value updating and goal selection in the primate orbitofrontal cortex. *ELife*, *4*, e11695. https://doi.org/10.7554/eLife.11695
Murray, E. A., O'Doherty, J. P., & Schoenbaum, G. (2007). What we know and do not know about the functions of the orbitofrontal cortex after 20 years of cross-species studies. *Journal of Neuroscience*, *27*(31), 8166–8169. https://doi.org/10.1523/JNEUROSCI.1556-07.2007
Murray, E. A., & Rudebeck, P. H. (2018). Specializations for reward-guided decision-making in the primate ventral prefrontal cortex. *Nature Reviews Neuroscience*, *19*(7), 404–417. https://doi.org/10.1038/s41583-018-0013-4
Nasser, H. M., Calu, D. J., Schoenbaum, G., & Sharpe, M. J. (2017). The Dopamine Prediction Error: Contributions to Associative Models of Reward Learning. *Frontiers in Psychology*, *8*, 244. https://doi.org/10.3389/fpsyg.2017.00244
Niv, Y. (2019). Learning task-state representations. *Nature Neuroscience*, *22*(10), 1544–1553. https://doi.org/10.1038/s41593-019-0470-8
Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-

14.2015
O'Doherty, J., Lee, S., Tadayonnejad, R., Cockburn, J., Iigaya, K., & Charpentier, C. J. (2020). Why and how the brain weights contributions from a mixture of experts. *Arxiv Preprint*, 1–18. https://doi.org/10.31234/osf.io/ns6kq
O'Keefe, J., & Nadel, L. (1978). The Hippocampus as a Cognitive Map. In *Philosophical Studies* (Vol. 27). Clarendon Press: Oxford. https://doi.org/10.5840/philstudies19802725
Ogawa, M., van der Meer, M. A. A., Esber, G. R., Cerri, D. H., Stalnaker, T. A., & Schoenbaum, G. (2013). Risk-responsive orbitofrontal neurons track acquired salience. *Neuron*, *77*(2), 251–258. https://doi.org/10.1016/j.neuron.2012.11.006
Ongur, D., & Price, J. L. (2000). The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans. *Cerebral Cortex*, *10*(3), 206–219. http://www.ncbi.nlm.nih.gov/pubmed/10731217
Ostlund, S. B., & Balleine, B. W. (2007). Orbitofrontal cortex mediates outcome encoding in pavlovian but not instrumental conditioning. *Journal of Neuroscience*, *27*(18), 4819–4825. https://doi.org/Doi 10.1523/Jneurosci.5443-06.2007
Padoa-Schioppa, C. (2009). Range-adapting representation of economic value in the orbitofrontal cortex. *Journal of Neuroscience*, *29*, 14004–14014. http://www.jneurosci.org/content/29/44/14004.full.pdf
Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, *6*. https://doi.org/10.1038/ncomms9096
Panayi, M. C., & Killcross, S. (2014). Orbitofrontal cortex inactivation impairs between- but not within-session Pavlovian extinction: An associative analysis. *Neurobiology of Learning and Memory*, *108*, 78–87. https://doi.org/10.1016/j.nlm.2013.08.002
Panayi, M. C., & Killcross, S. (2018). Functional heterogeneity within the rodent lateral orbitofrontal cortex dissociates outcome devaluation and reversal learning deficits. *ELife*, *7*. https://doi.org/10.7554/eLife.37357.001
Panayi, M. C., & Killcross, S. (2020). Title: The role of the rodent lateral orbitofrontal cortex in simple Pavlovian cue-outcome learning depends on training experience. *BioRxiv*.
Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol Rev*, *87*(6), 532–552. http://www.ncbi.nlm.nih.gov/pubmed/7443916
Pezzulo, G., Rigoli, F., & Chersi, F. (2013). The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. *Frontiers in Psychology*, *4*, 92.
Pickens, C. L., Saddoris, M. P., Gallagher, M., & Holland, P. C. (2005). Orbitofrontal lesions impair use of cue-outcome associations in a devaluation task. *Behav Neurosci*, *119*(1), 317–322. https://doi.org/2005-01705-030 [pii]10.1037/0735-7044.119.1.317
Pickens, C. L., Saddoris, M. P., Setlow, B., Gallagher, M., Holland, P. C., & Schoenbaum, G. (2003). Different Roles for Orbitofrontal Cortex and Basolateral Amygdala in a Reinforcer Devaluation Task. *The Journal of Neuroscience*, *23*(35), 11078–11084. https://doi.org/10.1523/JNEUROSCI.23-35-11078.2003
Price, J. L. (2006). Connections of orbital cortex. In D. H. Zald & S. L. Rauch (Eds.), *The Orbitofrontal Cortex* (pp. 39–55). Oxford University Press.

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. https://doi.org/10.1196/annals.1401.008

Ramirez, D. R., & Savage, L. M. (2007). Differential involvement of the basolateral amygdala, orbitofrontal cortex, and nucleus accumbens core in the acquisition and use of reward expectancies. *Behav Neurosci*, *121*(5), 896–906. https://doi.org/10.1037/0735-7044.121.5.896

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditiong: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokesy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64–99). Appleton Century Crofts.

Roesch, M. R., Calu, D. J., Esber, G. R., & Schoenbaum, G. (2010). All that glitters ... dissociating attention and outcome expectancy from prediction errors signals. *Journal of Neurophysiology*, *104*(2), 587–595. https://doi.org/10.1152/jn.00173.2010

Roesch, M. R., & Schoenbaum, G. (2006). From associations to expectancies: orbitofrontal cortex as a gateway between limbic system and representational memory. In D. H. Zald & A. L. Rauch (Eds.), *The Orbitofrontal Cortex* (pp. 199–235). Oxford University Press.

Rudebeck, P. H., & Murray, E. A. (2011a). Balkanizing the primate orbitofrontal cortex: distinct subregions for comparing and contrasting values. *Critical Contributions of the Orbitofrontal Cortex to Behavior*, *1239*, 1–13. https://doi.org/DOI 10.1111/j.1749-6632.2011.06267.x

Rudebeck, P. H., & Murray, E. A. (2011b). Dissociable effects of subtotal lesions within the macaque orbital prefrontal cortex on reward-guided behavior. *Journal of Neuroscience*, *31*(29), 10569–10578. https://doi.org/10.1523/jneurosci.0091-11.2011

Rudebeck, P. H., & Murray, E. A. (2014). The Orbitofrontal Oracle: Cortical Mechanisms for the Prediction and Evaluation of Specific Behavioral Outcomes. *Neuron*, *84*(6), 1143–1156. https://doi.org/10.1016/j.neuron.2014.10.049

Rudebeck, P. H., Saunders, R. C., Prescott, A. T., Chau, L. S., & Murray, E. A. (2013). Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. *Nature Neuroscience*, *16*, 1140–1145. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3733248/pdf/nihms483559.pdf

Rustichini, A., & Padoa-Schioppa, C. (2015). A neuro-computational model of economic decisions. *Journal of Neurophysiology*, *114*(3), 1382–1398. https://doi.org/10.1152/jn.00184.2015

Sadacca, B. F., Wied, H. M., Lopatina, N., Saini, G. K., Nemirovsky, D., & Schoenbaum, G. (2018). Orbitofrontal neurons signal sensory associations underlying model-based inference in a sensory preconditioning task. *ELife*, *7*, e30373. https://doi.org/10.7554/eLife.30373

Sallet, J., Noonan, M. A. P., Thomas, A., O'Reilly, J. X., Anderson, J., Papageorgiou, G. K., Neubert, F. X., Ahmed, B., Smith, J., Bell, A. H., Buckley, M. J., Roumazeilles, L., Cuell, S., Walton, M. E., Krug, K., Mars, R. B., & Rushworth, M. F. S. (2020). Behavioral flexibility is associated with changes in structure and function distributed across a frontal cortical network in macaques. *PLoS Biology*. https://doi.org/10.1371/journal.pbio.3000605

Schmajuk, N. A., Lam, Y.-W., & Gray, J. A. (1996). Latent inhibition: A neural network approach. *Journal of Experimental Psychology: Animal Behavior Processes*, *22*(3), 321–349. https://doi.org/10.1037//0097-7403.22.3.321

Schoenbaum, G., Chang, C.-Y., Lucantonio, F., & Takahashi, Y. K. (2016). Thinking Outside the Box: Orbitofrontal Cortex, Imagination, and How we can Treat Addiction. *Neuropsychopharmacology*. https://doi.org/10.1038/npp.2016.147

Schoenbaum, G., Nugent, S. L., Saddoris, M. P., & Setlow, B. (2002). Orbitofrontal lesions in rats impair reversal but not acquisition of go, no-go odor discriminations. *Neuroreport*, *13*(6), 885–890. https://doi.org/10.1097/00001756-200205070-00030

Schoenbaum, G., Roesch, M. R., Stalnaker, T. A., & Takahashi, Y. K. (2009). A new perspective on the role of the orbitofrontal cortex in adaptive behaviour. *Nature Reviews Neuroscience*, *10*(12), 885–892. https://doi.org/Doi 10.1038/Nrn2753

Schoenbaum, G., Setlow, B., Nugent, S. L., Saddoris, M. P., & Gallagher, M. (2003). Lesions of orbitofrontal cortex and basolateral amygdala complex disrupt acquisition of odor-guided discriminations and reversals. *Learning & Memory*, *10*(2), 129–140. https://doi.org/Doi 10.1101/Lm.55203

Schoenbaum, G., Setlow, B., Saddoris, M. P., & Gallagher, M. (2003). Encoding predicted outcome and acquired value in orbitofrontal cortex during cue sampling depends upon input from basolateral amygdala. *Neuron*, *39*(5), 855–867. http://ac.els-cdn.com/S0896627303004744/1-s2.0-S0896627303004744-main.pdf?_tid=cbae4aea-a4f7-11e4-a27c-00000aacb361&acdnat=1422234670_f56b4b46b535f657f295e85893dff9d3

Schoenbaum, G., & Shaham, Y. (2008). The role of orbitofrontal cortex in drug addiction: a review of preclinical studies. *Biol Psychiatry*, *63*(3), 256–262. https://doi.org/S0006-3223(07)00534-3 [pii]10.1016/j.biopsych.2007.06.003

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

Schultz, W., Stauffer, W. R., & Lak, A. (2017). *The phasic dopamine signal maturing: from reward via behavioural activation to formal economic utility*. https://doi.org/10.1016/j.conb.2017.03.013

Sharpe, M. J., & Killcross, S. (2018). Modulation of attention and action in the medial prefrontal cortex of rats. *Psychological Review*. https://doi.org/10.1037/rev0000118

Sharpe, M. J., Stalnaker, T., Schuck, N. W., Killcross, S., Schoenbaum, G., & Niv, Y. (2019). An Integrated Model of Action Selection: Distinct Modes of Cortical Control of Striatal Decision Making. In *Annual Review of Psychology*. https://doi.org/10.1146/annurev-psych-010418-102824

Stalnaker, T. A., Cooch, N. K., McDannald, M. A., Liu, T. L., Wied, H., Schoenbaum, G., & Tzu-Lan, L. (2014). Orbitofrontal neurons infer the value and identity of predicted outcomes. *Nat Commun*, *5*, 3926. https://doi.org/10.1038/ncomms4926

Stalnaker, T. A., Cooch, N. K., & Schoenbaum, G. (2015). What the orbitofrontal cortex does not do. *Nature Neuroscience*, *18*(5), 620–627. https://doi.org/10.1038/nn.3982

Steinberg, E. E., Keiflin, R., Boivin, J. R., Witten, I. B., Deisseroth, K., & Janak, P. H. (2013). A causal link between prediction errors, dopamine neurons and learning. *Nature Neuroscience*, *16*(7), 966–973. https://doi.org/10.1038/nn.3413

Stringfield, S. J., Palmatier, M. I., Boettiger, C. A., & Robinson, D. L. (2017). Orbitofrontal participation in sign- and goal-tracking conditioned responses: Effects of nicotine. *Neuropharmacology*, *116*, 208–223. https://doi.org/10.1016/j.neuropharm.2016.12.020

Sutton, R. S., & Barto, A. G. (1987). A temporal-difference model of classical conditioning. In

Proceedings of the Ninth Conference of the Cognitive Science Society (pp. 355–378).
Sutton, R. S., & Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.
Takahashi, Y. K., Chang, C. Y., Lucantonio, F., Haney, R. Z., Berg, B. A., Yau, H.-J., Bonci, A., & Schoenbaum, G. (2013). Neural estimates of imagined outcomes in the orbitofrontal cortex drive behavior and learning. *Neuron*, *80*, 507–518. http://ac.els-cdn.com/S0896627313007198/1-s2.0-S0896627313007198-main.pdf?_tid=97ea45dc-cbcc-11e4-9c0e-00000aacb362&acdnat=1426504211_80b43da207445d70382dd3274f8c4f11
Takahashi, Y. K., Roesch, M. R., Stalnaker, T. A., Haney, R. Z., Caiu, D. J., Taylor, A. R., Burke, K. A., Schoenbaum, G., & Calu, D. J. (2009). The Orbitofrontal Cortex and Ventral Tegmental Area Are Necessary for Learning from Unexpected Outcomes. *Neuron*, *62*(2), 269–280. https://doi.org/DOI 10.1016/j.neuron.2009.03.005
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. *Nature Neuroscience*, *14*(12), 1590–1597. https://doi.org/10.1038/nn.2957
Trapold, M. A., & Overmier, J. B. (1972). The second learning process in instrumental learning. In W. F. Prokasy & A. H. Black (Eds.), *Classical Conditioning II: Current Theory and Research* (pp. 427–452). Appleton Century Crofts.
Urcuioli, P. J. (2005). Behavioral and associative effects of differential outcomes in discrimination learning. *Learn Behav*, *33*(1), 1–21. http://www.ncbi.nlm.nih.gov/pubmed/15971490
Van Der Meer, M., Kurth-Nelson, Z., & Redish, A. D. (2012). Information processing in decision-making systems. *Neuroscientist*, *18*(4), 342–359. https://doi.org/10.1177/1073858411435128
van Duuren, E., Lankelma, J., & Pennartz, C. M. (2008). Population coding of reward magnitude in the orbitofrontal cortex of the rat. *Journal of Neuroscience*, *28*(34), 8590–8603. https://doi.org/10.1523/JNEUROSCI.5549-07.2008
van Wingerden, M., Vinck, M., Lankelma, J., & Pennartz, C. M. (2010). Theta-band phase locking of orbitofrontal neurons during reward expectancy. *Journal of Neuroscience*, *30*(20), 7078–7087. https://doi.org/30/20/7078 [pii]10.1523/JNEUROSCI.3860-09.2010
Viejo, G., Khamassi, M., Brovelli, A., & Girard, B. (2015). Modeling choice and reaction time during arbitrary visuomotor learning through the coordination of adaptive working memory and reinforcement learning. *Frontiers in Behavioral Neuroscience*, *9*, 225.
Wagner, A. R., & Brandon, S. E. (1989). Evolution of a Structured ConnectionistModel of Pavlovian Conditioning (AESOP). In S. B. Klein & R. R. Mowrer (Eds.), *Contemporary learning theories: Pavliocian conditioning and the status of tradional learning theories* (pp. 149–189). Lawrence Erlbaum.
Walton, M. E., Behrens, T. E., Buckley, M. J., Rudebeck, P. H., & Rushworth, M. F. (2010). Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. *Neuron*, *65*(6), 927–939. https://doi.org/10.1016/j.neuron.2010.02.027
Wang, T., Bao, X., Clavera, I., Hoang, J., Wen, Y., Langlois, E., Zhang, S., Zhang, G., Abbeel, P., & Ba, J. (2019). *Benchmarking Model-Based Reinforcement Learning*. 1–25.
West, E. A., DesJardin, J. T., Gale, K., & Malkova, L. (2011). Transient Inactivation of

Orbitofrontal Cortex Blocks Reinforcer Devaluation in Macaques. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *31*(42), 15128–15135. https://doi.org/10.1523/JNEUROSCI.3295-11.2011

Westbrook, R. F., & Bouton, M. E. (2010). Latent inhibition and extinction: Their signature phenomena and the role of prediction error. In *Latent Inhibition: Cognition, Neuroscience and Applications to Schizophrenia*. https://doi.org/10.1017/CBO9780511730184.003

Wikenheiser, A. M., & Schoenbaum, G. (2016). Over the river, through the woods: cognitive maps in the hippocampus and orbitofrontal cortex. *Nature Reviews Neuroscience*, *17*(8), 513–523. https://doi.org/10.1038/nrn.2016.56

Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron*, *81*(2), 267–279. https://doi.org/10.1016/j.neuron.2013.11.005

Zhou, J., Gardner, M. P. H., Stalnaker, T. A., Ramus, S. J., Wikenheiser, A. M., Niv, Y., & Schoenbaum, G. (2019). Rat Orbitofrontal Ensemble Activity Contains Multiplexed but Dissociable Representations of Value and Task Structure in an Odor Sequence Task. *Current Biology*. https://doi.org/10.1016/j.cub.2019.01.048

Zimmermann, K. S., Li, C. C., Rainnie, D. G., Ressler, K. J., & Gourley, S. L. (2018). Memory retention involves the ventrolateral orbitofrontal cortex: Comparison with the basolateral amygdala. *Neuropsychopharmacology*, *43*(2), 373–383. https://doi.org/10.1038/npp.2017.139