The role of prefrontal cortex and basal ganglia in model-based and model-free reinforcement learning

Bruno Andre e Silva Miranda

Sobell Department of Motor Neuroscience and Movement Disorders

University College of London

A dissertation submitted in fulfilment of the requirements for the degree of *Doctor of Philosophy* in University College London

February 2016

Declaration

I, Bruno Andre e Silva Miranda confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Bruno Andre e Silva Miranda
February 2016

Acknowledgements

- **Dr. Steve Kennerley (primary supervisor or the experimentalist)**: for being very enthusiastic about the project from the very beginning until the very end; for his help throughout and optimism about the final results; for his friendship; and finally, for insisting on recording from frontal pole reflecting well his adventurous spirit with experiments

- **Prof. Peter Dayan (secondary supervisor or the theoretician)**: for the constant availability, time spent on the project and words spent on emails; for the adequate enthusiasm with the results, by always suggesting further confirmatory analysis; a role model in the way he thinks and in his work ethic; his ideas of RL, MF-RL and MB-RL are in the origin of this PhD and i hope this data could help somehow in his theories.

- **Nishantha Malalasekera**: for being the PhD mate that everybody needs to have; Nish is a very clever person who helped in various aspects of this project; he is also a good friend

- **Joachim Morrens**: for the work done with the pupil data; for the enthusiasm and hard work in the project; for the nice chats after hours; he is also a good friend

- **Dr. Laurence Hunt**: for suggestions of analysis and for answering several difficult questions

- **Prof. Tim Behrens**: for his intelligence that helped me dealing with some issues with data analysis or data interpretation; for the upgrade examination

- **Lianne McCombe**: for helping on a daily basis, particularly in training and animal welfare; in addition, she was always supportive and kind when help was needed

- **Jonathon Henton and Spencer Neil**: for all the help with the experimental apparatus

- **Dr. Zach Mainen and Dr. Joseph Paton**: for their guidance and role as mentors at the INDP-CNP programme

- **Prof. Roger Lemon and Dr. Alexander Kraskov**: Roger for the incredible stories and for the upgrade examination; Sasha for always having time for answering my questions

- **Deborah Hadley, Kully Sunner, Hayley Mackenzie and Chris Seers**: for all the administrative and computing help

- **Maria Botcharova and Hedi Young**: for making the time spent in the lab more enjoyable with our conversations at lunch

- **Stephan Waldert, Roland Philipp and Ganesh Vignesawan**: for always being keen to have a chat to relax and forget some stressful time in the lab

- **My family and old friends**: to my wife, mostly, who could not have helped more by being always on my side with love, comprehension and motivation; to my parents who educated me in all sense and who will be very proud for this achievement; to my grandmother whose memory (or lack of it) does not let her feel the happiness of the accomplishment; to uncle and aunt, particularly the former for introducing me to academia; to my grandfather, godparents, parents in law, old friends...

A final acknowledgement to Filipe and Gustavo, who has made my life very special since their arrival.

Abstract

Contemporary reinforcement learning (RL) theory suggests that choices can be evaluated either by the model-free (MF) strategy of learning their past worth or the model-based (MB) strategy of predicting their likely consequences based on learning how decision states eventually transition to outcomes. Statistical and computational considerations argue that these strategies should ideally be combined. This thesis aimed to investigate the neural implementation of these two RL strategies and the mechanisms of their interactions.

Two non-human primates performed a two-stage decision task designed to elicit and discriminate the use of both MF and MB-RL, while single-neuron activity was recorded from the prefrontal cortex (frontal pole, FP; anterior cingulate cortex, ACC; dorsolateral prefrontal cortex) and striatum (caudate and putamen). Logistic regression analysis revealed that the structure of the task (of MB importance) and the reward history (of MF and MB importance) significantly influenced choice. A trial-by-trial computational analysis also confirmed that choices were made according to a weighted combination of MF and MB-RL, with the influence of the latter approaching 90%. Furthermore, the valuations of both learning methods also influenced response vigour and pupil response.

Neural correlates of key elements for MF and MB learning were observed across all brain areas, but functional segregation was also in evidence. Neurons in ACC encoded features of both MF and MB, suggesting a possible role in the arbitration between both strategies. Striatal activity was consistent with a role in value updating by encoding reward prediction errors. Finally, novel neurophysiological evidence was found in favour of the role of the FP in counterfactual processing.

In conclusion, this thesis provides insight into the neural implementation of MF and MB-RL computations and their various effects on diverse aspects of behaviour. It supports the parallel operation and integration of the two approaches, while revealing unexpected intricacies.

Table of contents

List of figures

List of tables

Long abstract

Even for some of our most basic daily behaviours, such as taking the underground, our brain engages complicated systems that compete and cooperate to achieve good solutions. When deciding on the best course of action, or route, one can either make use of accumulated past experience, or plan ahead using information from an internal or external map. Most often we end up combining both strategies to a degree that depends on factors such as fatigue, or indeed external challenges such as ”planned” engineering works. Increasing evidence suggests that separate neural circuits learn the value of actions in different ways. Studying these distinct mechanisms of learning as well as how their interactions are implemented could provide foundational explanations for reward-guided choices. Furthermore, the more we understand the mechanisms and constraints of normal learning and decision-making the better we can understand what happens when this goes wrong, as in neurological and psychiatric diseases that exhibit maladaptive decision-making.

This thesis used behaviour, computational modelling and single-neuron physiology in the prefrontal cortex (frontal pole, FP; anterior cingulate cortex, ACC; dorsolateral prefrontal cortex, DLPFC) and dorsal striatum (caudate and putamen), to advance our current understanding of the neural mechanisms about how the two forms of instrumental control – model-free (MF) and model-based (MB) – interact and determine action choice in animals.

In chapter 1, we introduce reinforcement learning (RL) theory as a normative and explanatory theoretical framework for studying how agents interact with the environment, predicting rewards and optimizing future benefits. We use RL-based accounts of behaviour to expose the computations underlying decision-making; these accounts are of particular relevance when there are different sources of information that can inform choice. Here, we concentrate on model-free (MF) and model-based (MB) RL methods, which resemble the two strategies given in the underground example. Model-free RL bases choice on caching previous experience without directly estimating the structure of the environment. By contrast, MB valuation techniques exploit a model of the world for planning or to simulate possible futures. Both methods rely on previous experience but they differ as to how this information is used to infer the long-run future values of choices. It might be natural to

think of competition between MF and MB methods; however, statistical and computational considerations argue that it may be ideal for a learning agent to combine both strategies. How these two forms of learning could work towards a common goal is a topic under active investigation. We discuss various theoretical proposals.

In chapter 2, we provide a neuroscience perspective on the different reward-learning systems, focusing successively on anatomy, neuropsychology and electrophysiology. The basic anatomical connectivity of the prefrontal cortex with its three main subdivisions (orbital, medial and lateral networks) as well as the main organisation of the basal ganglia and its prefrontal connections are detailed. This is important because the anatomical incoming connections of a brain region constrain the type of information it can process, and its outgoing projections dictate the influence this processing can have on other brain regions.

In terms of neuropsychology, from the famously unfortunate case of Phineas Gage – the rail worker who suffered in 1848 a traumatic accident in which an iron rod was propelled through the front of his left brain hemisphere causing orbital and medial prefrontal damage – to the multiple descriptions of patients with similar prefrontal cortex damage, value-based decision making deficits have been consistently emphasised. More recently, the use of more focal lesions in non-human primates has helped to unmask functional subdivisions within the prefrontal cortex, and the chapter highlights the importance for learning of each region. Neuropsychological studies in rodents have also played an important role in teasing apart the neural bases of MF and MB control, based on the relationship between these and the behavioural notions respectively of habitual and goal-directed action selection. They highlighted that dorsolateral striatum has an important role in MF learning, whereas dorsomedial striatum and some areas of the frontal cortex seem to be more critical for a MB system. However, such traditional studies have exploited manipulations such as outcome devaluation that offer limited opportunities to explore continuing trade-offs.

Finally, primate neurophysiological studies have provided complementary insights into the actual neuronal computations preformed, with the reward prediction error signal of dopaminergic cells and the role of striatum in action selection being two particularly good examples. However, as a whole, very few studies have focused on detecting simultaneous neuronal signals of both learning strategies and, only more recently has a class of new tasks been invented for human subjects to specifically examine how the two strategies are combined. We adapt one such task for primate subjects.

In chapter 3, we describe how two subjects were trained to perform a two-stage decision task, similar to the one used in a previous human study. This task is intended to induce trial-by-trial adjustments in choice that combines both MF and MB learning control. We

used a more descriptive method of regression analysis together with a computational trial-by-trial RL-based method, to assess quantitatively the signatures of both learning strategies. With the first approach, outcome history (relevant for both learning strategies) and state-transition knowledge (used in MB computations) had a significant impact on choice, to an extent that decayed exponentially as a function of trials into the past. The computational analysis confirmed that choices were made according to a Hybrid model which used a weighted combination of MF and MB-RL, with the influence of the latter approaching 90% and remaining at this level across weeks of testing. Comparison of the actual data and choices simulated from the best-fitting Hybrid model showed that some significant structure in the behaviour had not been captured, in particular an excessive influence of events that happened on the immediately previous trial. Hence, we built a new combined Hybrid+ RL model which incorporated a credit assignment weighting procedure. Finally, it was also found that both forms of RL influenced the alacrity of responding, in agreement with the speed-accuracy trade-off associated with their computations. In conclusion, the behavioural results presented in this chapter enrich modern views of MB and MF integration.

In chapter 4, we describe analysis of the dynamics of pupil dilation in our two subjects. Changes in pupil diameter have long been reported in cognitive processes and a relationship with learning processes has recently been emphasised. It was found that pupil diameter tracked the expected value of choices at both pre- and post-decision moments, and these choice valuations were best correlated with value estimates derived from a MB system. Further confirmation of task-specific features of MB calculations came from the observations that pupil diameter also independently coded information about the state transition in the task in a way that was modulated by expected values. Finally, when feedback was provided, pupil diameter reflected a reward prediction error signal. Overall, several essential elements of value-based reinforcement learning processes were evident in pupillary response at different key behavioural stages of the task.

In chapter 5, we present neuronal correlates of key elements of RL – such as reward, state-transitions and first-stage choices. These factors were encoded in the activity of single units and across the population of neurons in prefrontal and striatal regions at different time points in the behavioural task introduced in chapter 3. As a region, the ACC most prominently encoded reward and transition information, with both types of information being coded concomitantly at feedback epoch. There were specific relationships between the transition selectivity of FP neuronal activity at the time of transitions and their reward coding later at feedback.

At the time of choices in the task, we considered the distribution of three types of neurons: those selective solely to MF or MB action-values, and those that covaried with both (or the Hybrid action-value). All types of units were found across the recorded regions, but the ACC had a significantly greater proportion of Hybrid action-value neurons than any other region. We found that FP neurons increased their firing rate more when rare transitions happened and the expected chosen value was high. We interpreted this coding as a quantity akin to the foregone expected value or *regret*. On the other hand, caudate and, to a lesser extent putamen, decreased their firing rate for the same situation consistent with a negative prediction error but just for rare trials.

Finally, neurons in both caudate and putamen encoded the second-stage reward prediction error, although we found small qualitative differences between these areas. As a whole, the neural evidence was consistent with the view that MF and MB controllers operated in parallel, but their associated signals were more richly intertwined than had originally been expected.

In chapter 6, a brief summary of the main results of this thesis are outlined, along with an attempt to fit them into a broader understanding of the field of MF- and MB-RL. We also highlight the many questions raised by our study which could fruitfully be addressed in future theoretical and experimental work. This thesis concludes that strong evidence was found in favour of a parallel organisation of MF and MB valuations but more complex and richly intertwined across the prefrontal-striatal circuitry than previously thought. However, through analysis of the different algorithmic elements expected by RL, we also found functional subdivisions performing specific and unique reward-based learning computations.

Chapter 1

Reinforcement learning

This chapter aims to provide a short introduction to reinforcement learning theory, focusing specifically on work related to the theoretical issues of model-free and model-based approaches. It also establishes notation and describes algorithms relevant to the scientific content of the thesis. More comprehensive descriptions can be found elsewhere (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998).

1.1 Reinforcement learning: introduction

Reinforcement learning is a formal framework that studies how to learn in order to maximize reward over time. It requires an agent that observes states describing features of its surroundings, chooses and performs actions that change the state of the world and receives in return either positive (rewards) or negative (punishments) payoffs. (Figure 1.1). The returns define the goal of the agent and the learning aims to choose the best actions to obtain the best long term rewards.

The dynamic interaction between agent and the environment is framed according to a *Markov decision process*. In a Markov decision process, the world is described by a set of environmental states $\mathcal{S}$ evolving in discrete time steps t and conditional on a set of possible actions $\mathcal{A}$ made by the agent at each stage of the decision process. When the environment has a finite horizon, the learning experience is broken into episodes (or trials) and the Markov decision process is called episodic (non-episodic Markov decision processes will not be the focus of this review).

Importantly, the current state $s_t \in \mathcal{S}$ contains all information needed for the agent's decision about the current action $a_t \in \mathcal{A}$ and the choice made gives rise to a next state $s_{t+1} \in \mathcal{S}$ and a potential reward r_{t+1}. This is known as the Markov property, which relies

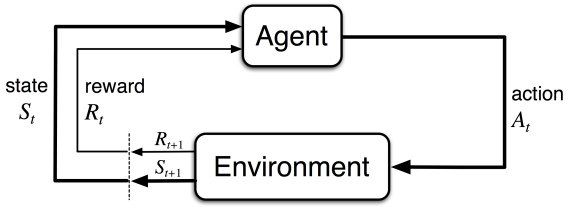


Fig. 1.1 **The reinforcement learning problem.** The agent observes the state of the environment and, according to its policy, takes an action. Depending on the state and the agent's action, the environment evolves to a new state and provides a reward. The goal of the agent is to improve its policy so that it can get more rewards in the long run, implying an optimisation of not only the immediate reward but also of all future returns. From Sutton and Barto (1998).

on the equality between the probability distributions including information about all past events and the most recent event:

$$P(s_{t+1} = s', r_{t+1} = r | s_t, a_t, r_t, s_{t-1}, a_{t-1}, ..., r_1, s_0, a_0) = P(s_{t+1} = s', r_{t+1} = r | s_t, a_t) \tag{1.1}$$

The environment's structure of such state changes is given by a transition function:

$$\boldsymbol{T}(s, a, s') = P(s_{t+1} = s' | s_t = s, a_t = a) \tag{1.2}$$

specifying the probability P that the state changes from s to s' given that the action a was taken. On the other hand, the returns obtained from the world formalise the agent's goal and could be mathematically defined by a reward function:

$$\boldsymbol{R}(s, a, s') = E\left[r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\right] \tag{1.3}$$

which represents the expected value of the upcoming reward given the current state and action, as well as any next state.

A *policy* is a stochastic rule formalised as a function which maps states to probabilities of actions and where $\pi(s, a)$ is the probability that action a is taken in state s. However, an important challenge that the agent faces is to balance immediate rewards with long-term ones that rely on the sequence of actions taken and the future encountered states. To achieve the optimal balance, it is crucial for the agent to evaluate the expected long-term outcome if action a is taken in state s given that policy π is followed thereafter. This is achieved by having a policy's value function assigning the return from a state-action pair as a discounted

sum of future rewards:

$$\boldsymbol{Q}^{\pi}(s,a) = E\left[\sum_{k=0}^{\infty}\gamma^{k}r_{t+k+1}|s_t = s, a_t = a\right] \tag{1.4}$$

where $\gamma \in [0,1]$ is a discount factor which models the fact that future rewards are worth less than immediate ones. The larger γ, the more weight the agent's give to distant rewards, and, typically, the harder the optimization problem. The discounted reward essentially measures the present value of the sum of the returns earned in the future, where the rewards are weighed by a factor γ that decreases exponentially as the time step increases:

$$E\left[\sum_{k=0}^{\infty}\gamma^{k}r_{t+k+1}\right] = r_{t+1} + \gamma r_{t+2} + \gamma^{2}r_{t+3} + ... \tag{1.5}$$

Dynamic programming offers methods to efficiently solve multi-stage decision processes (Bellman, 1957). According to its *Principle of Optimality*, an optimal sequence of decisions *has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy (denoted as π^*) with regard to the state resulting from the first decision*. Therefore, for a given state-action pair and assuming such consistency at successive time points, the sum of the discounted future rewards of all successor states-action pairs could be contracted to the optimal value (denoted as $\boldsymbol{Q}^*$) of the very next state-action pair under policy π:

$$\boldsymbol{Q}^{*}(s,a) = \boldsymbol{Q}^{\pi^{*}}(s,a) = \sum_{s'\in\mathcal{S}}\boldsymbol{T}(s,a,s')[\boldsymbol{R}(s,a,s') + \gamma \operatorname*{argmax}_{a'\in\mathcal{A}}\boldsymbol{Q}^{\pi}(s',a')] \tag{1.6}$$

The above equation, also known as the Bellman equation for a state-action pair, includes the immediate reward and the discounted long-term value averaged over the probability of the subsequent state given by the transition function. Two main methods can be used for solving such optimisation problems: policy iteration and value iteration.

The policy iteration procedure works by iteratively evaluating the policy, and then improving it. The policy evaluation step consists of computing the state-action value function $\boldsymbol{Q}^{\pi}$ starting from state s and following policy π, given by:

$$\boldsymbol{Q}^{\pi}(s,a) = \sum_{s'\in\mathcal{S}}\boldsymbol{T}(s,a,s')[\boldsymbol{R}(s,a,s') + \gamma\boldsymbol{Q}(s',\pi(s'))] \tag{1.7}$$

Once the evaluation process is completed, the $Q^{\pi}(s,a)$ values assign credit to the advantageous actions and can then be used to determine a new policy π' where $\boldsymbol{Q}^{\pi'}(s,\pi'(s)) \geq \boldsymbol{Q}^{\pi}(s,\pi(s))$. When the policy π' is the same as the policy π, then an optimal policy is

achieved. The improvement of the policy occurs concurrently at every state and involves a maximisation over the action space:

$$\pi'_s = \underset{a \in \mathcal{A}}{\operatorname{argmax}} \boldsymbol{Q}^{\pi}(s, a) \tag{1.8}$$

Despite being the optimal solution when there is full knowledge of the available state-action values in stationary environments, it may not be the best approach for dealing with uncertainty of value estimates or in a changing world. In these scenarios, choosing actions that are believed to be suboptimal might actually be good and the correct balance of such exploration-exploitation trade-off is key to the reinforcement learning optimization problem. Other methods, such as the ε-greedy algorithm, exploit the best action most of the time but, with a small probability (ε), an action is selected at random independently of the action-value estimates. This selection technique achieves good performance in several learning situations, but can fail when very low value options exist because the algorithm explores equally among all possible actions. Given these limitations, an alternative and most commonly used method is the *softmax* function where the probability of choosing an action a in state s, is a function of the estimated state-action values $Q^{\pi}(s, a)$ obtained from the policy evaluation procedure:

$$\pi'_{s,a} = \frac{exp(\beta Q^{\pi}(s, a))}{\sum_{a' \in \mathcal{A}} exp(\beta Q^{\pi}(s_t, a'))} \tag{1.9}$$

where $\beta \in [0, \infty]$ is the so called *inverse temperature* parameter that modulates how deterministic are the choices. Low inverse temperatures cause greater choice randomness, whereas high values cause greater difference in selection probability for actions with different value estimates.

Value iteration is another method for finding the optimal policy. In contrast to policy iteration that waits for full value convergence to improve the policy, the value iteration approach truncates evaluation after just one iteration, combining the policy improvement step into its iterations. By directly computing a sequence of state-action value estimates which converge to the optimal value function, this procedure solves the Markov decision problem. It updates the expected value of each state-action pair until the values calculated on two successive steps are close enough and according to the update rule:

$$\boldsymbol{Q}'(s, a) = \sum_{s' \in \mathcal{S}} \boldsymbol{T}(s, a, s')[\boldsymbol{R}(s, a, s') + \gamma \underset{a' \in \mathcal{A}}{\operatorname{argmax}} \boldsymbol{Q}(s', a')] \tag{1.10}$$

Iterating this equation infinitely often, guarantees that optimal values can be obtained.

Once that is achieved, the optimal policy can be easily defined by applying the same principles as defined in policy improvement.

The way both policy and value iteration methods are implemented by reinforcement learning theory is the basis of the differences between model-free and model-based learning strategies. In the model-free reinforcement learning framework, the agent does not have knowledge of the transition and reward functions and it learns by trial-and-error which action is best at each state. On the other hand, a model-based learner finds the best policy by rather using a predictive model of the transition matrices or the expected outcomes its actions produce.

1.2 Model-free reinforcement learning

Model-free reinforcement learning attempts to estimate the long-term value without storing in a model information about the structure of the environment or how the world may respond to a given action. In formal terms, it does not take explicit advantage of the transition function to estimate future values. Alternatively, state-action values are estimated by sampling the Markov decision problem in order to obtain statistical knowledge and to help the agent formalising a choice given the state of the world. Most of these algorithms achieve such objective relying on temporal difference methods. At the heart of the incremental updates of these algorithms is a measure of the difference between estimates of the value function at two successive states, known as the temporal difference prediction error δ:

$$\delta_t = r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \tag{1.11}$$

This error in the estimate drives learning by correcting the prediction through the following update rule:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\delta_t \tag{1.12}$$

where $\alpha \in [0, 1]$ is the learning rate that determines by how much values get updated. This iterative error based rule was inspired by previous mathematical treatments of psychological theories of learning (Bush and Mosteller, 2006), as well as on the famous Rescorla-Wagner model (Rescorla and Wagner, 1972).

Several temporal differences methods were developed to find an optimal policy but three have been particularly used to account for behaviour of animals in psychological experiments: the actor-critic architecture, the Q-learning and the $SARSA$ algorithm. Overall, they

either keep a separate policy independent of the value function (as in the actor-critic learning) or they learn state-action functions generating a look-up table representation of these values (as in the Q-learning and in the $SARSA$).

The actor-critic method is similar to policy iteration, as it improves the policy on the basis of a computed value function (Barto et al., 1983, 1995). It takes advantage of a two-process structure, an actor and a critic (Figure 1.2). The actor is responsible for selecting actions according to a modifiable policy $\pi(s)$, without explicit knowledge of their consequences and based on a set of weighted associations from states to actions, often called action strengths. The critic estimates the value function $V(s)$ and computes a temporal difference prediction error δ used to criticise the actor's decision as well as its own value estimation. Both the action strengths and the value function must be learned based on experience with the environment and as this interaction evolves, the critic's value function becomes progressively more accurate, and the actor's action strengths change so as to yield progressive improvements in behaviour.

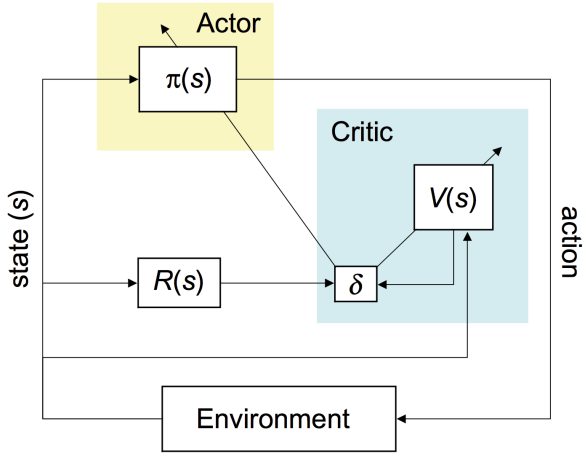


Fig. 1.2 **The actor-critic architecture.** Schematic of the relationship between agent and environment according to the basic actor–critic architecture, where arrows represent direction of computations. $\pi(s)$: policy, determined by action strengths; $R(s)$: reward at state s; δ: temporal- difference reward prediction error; $V(s)$: value function. From Botvinick et al. (2009).

Q-learning is probably the most commonly used model-free temporal difference version of a value iteration algorithm (Watkins, 1989). It starts from an arbitrary initial state-action value function and updates it using observed state transitions and rewards, according to:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_t + \gamma \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q(s_{t+1}, a) - Q(s_t, a_t) \right] \tag{1.13}$$

The prediction error of Q-learning is originated from value estimates that are independent of the policy being followed, an off-policy method, and based on the estimated best next action in the sequence:

$$\delta_t = r_t + \gamma \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q(s_{t+1}, a) - Q(s_t, a_t) \tag{1.14}$$

As in value iteration, the policy can then be defined following the same action-selection methods as described for policy improvement (for example, equations 1.8 or 1.9). Interestingly, the Q-learning algorithm learns the optimal policy even when actions are selected according to a more exploratory or even random policy. In fact, the requirement for correct convergence is that all state-action pairs are tried often enough. In other words, although the exploration-exploitation trade-off needs to be taken into account, the details of the exploration strategy will not affect the convergence of the Q-learning algorithm. However, in some cases such as in situations with large negative rewards, it could be disadvantageous to adopt an off-policy strategy and ignore what the agent actually does.

The SARSA algorithm (Rummery and Niranjan, 1994) is another model-free method that learns state-action functions but, in contrast to Q-learning, it takes into account the temporal difference of the state-action value of the actual action being selected according to the current policy:

$$\delta_t = r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \tag{1.15}$$

Being especially useful in non-stationary environments, a SARSA agent is an on-policy learner because it evaluates the policy that is currently being followed by using state-action-reward-state-action experiences to update the state-action values:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \left[r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \right] \tag{1.16}$$

Because the value function convergence can take a long time and in the meantime potentially suboptimal actions can be performed, SARSA improves the policy at every time step combining a greedy component with exploration (by using, for example, the softmax function) being a type of online model-free policy iteration.

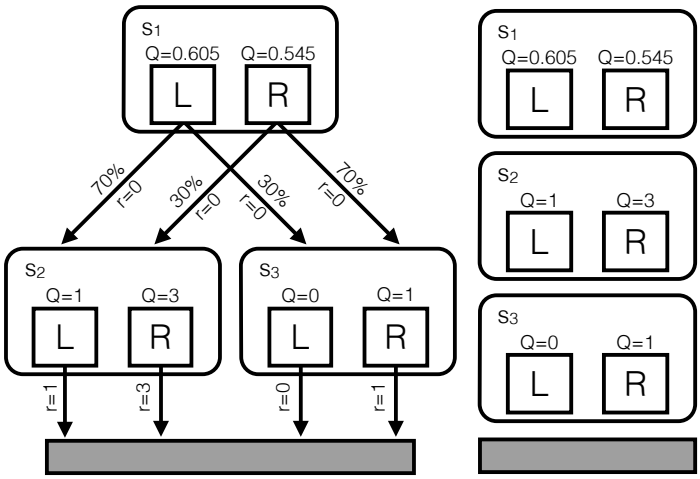
1.3 Model-based reinforcement learning

In contrast to all the above learning methods, which learn directly from experience, model-based reinforcement learning solves the valuation problem by learning the state-transition probabilities defining the structure of the learning task, as well as the returns obtained from the world. Such knowledge of the consequences of actions at states can then be used to guide decisions by estimating the expected rewards in a forward manner or through simulated experience, instead of relying exclusively on real experienced information as it happens in model-free learning. More formally, the agent learns both transition function and reward function that fully describe the Markov decision. This description of the model of the environment, in the form of a probability distribution of all possibilities, is then used to produce or improve a policy. In computational terms, this process of calculating a policy based on a predictive model is often called planning.

In order to illustrate the key principles of model-based learning as well as its main differences with the model-free approaches, an example of an episodic Markov decision process is given in Figure 1.3. In this hypothetical experiment, the goal of the reinforcement learning agent is to maximise its return by making two choices per episode. There are three states (s_1; s_2; and s_3), each with two possible actions (L: left; and R: right). To note that there is also a special absorbing state, corresponding to the end of an episode and allowing a new trial to start. State s_1 is always the starting state and the transition structure of the task follows a probability (P) that depends on the choice made in this initial state: $P(s_2|s_1, L) = 0.7 \wedge P(s_2|s_1, R) = 0.3 \wedge P(s_3|s_1, L) = 0.3 \wedge P(s_3|s_1, R) = 0.7$. There is no reward after the first action in s_I, but the second choice is rewarded according to the reward structure shown in (Figure 1.3). The optimal policy is simply obtained by choosing left in s_1 and then right if transitioned to either s_2 or s_3.

In reinforcement learning the model-based approach to the search for the optimal policy is most often defined over the space of states. Here, the agent computes the action values by searching through the tree of potential future states, summing rewards over state-trajectories and averaging them with respect to the state-transition matrix. Following the example of Figure 1.3 and using Bellman's equation, a model-based learner solves the Markov decision problem by computing the state-action values for s_I according to:

$$Q(s_1, a_i) = P(s_2|s_1, a_i) \underset{a \in \{L,R\}}{\operatorname{argmax}} Q(s_2, a) + P(s_3|s_1, a_i) \underset{a \in \{L,R\}}{\operatorname{argmax}} Q(s_3, a) \tag{1.17}$$



(a) (b)

Fig. 1.3 **Example of a Markov decision task with different model-based (a) and model-free (b) representations.** In this episodic Markov decision problem the agent has to perform two sequential binary decisions. There are three states (s_1; s_2; and s_3), each with two possible actions (L: left; and R: right). A special absorbing state (represented as a grey square) exists, corresponding to the end of an episode and allowing a new trial to start. At the starting state (s_1) each of the choices could lead to either a common (70% probability from left to s_2 or right to s_3) or a rare (30% probability from left to s_3 or right to s_2) second-stage state. This state transition was not associated with any reward (r). In the second-decision (either in s_2 or in s_3), another two-option choice was required and it was reinforced according to different rewards (in s_2, left has reward=1 and right has reward=3; in s_3, left has reward=0 and right has reward=1). The decision tree of model-based representation (a) of the task structure includes the state-transition structure, whereas a model-free controller (b) represents only the expected future value (Q) for each action in each state, without the tree sequence or identity of future consequences.

An important component of the model-based approach is learning the model that is needed to simulate the outcome of taking an action. This is achieved while the agent interacts with the environment but different methods can be used for that aim. A common and simple way for model-learning is to use a tabular maximum likelihood approach, which estimates each transition probability as the ratio of transitions between two states versus the total transitions out of a state (Brafman and Tennenholtz, 2003). In addition, the expected reward for a particular state-action is also computed. In the given example, based on the counts of each experienced transition to either s_2 or s_3 given the choice made in s_1, the agent can make predictions about the actual transition probabilities. The transition matrix estimation will improve as the number of trials increase.

Once the agent has learned the structure of the environment, the policy can be improved given the predictive model. Planning methods can either compute the state-action value functions for the entire state space through policy or value iteration (as described above), or they can focus on the states that the agent is more likely to encounter in the future. In fact, the latter methods can be also used while evaluating the tree of possible future state-action sequences. This on-line search consists of a forward algorithm that recursively estimates the value of each possible state-action pair rooted at the current state. A common way of achieving such goal is by using Monte-Carlo simulations to recursively average the value of each possible state-action pair, where each simulation starts from the current state and extends across the decision tree until the end of an episode. Methods that use such strategy can be more efficient than planning over the entire state space and are called Monte Carlo Tree Search (Browne et al., 2012).

1.4 Interaction between model-free and model-based reinforcement learning

In the field of reinforcement learning, there is an active debate as to the situations in which one of the two model-free or model-based learning strategies is best. This can depend on the learning problem itself as well as on intrinsic properties of the learning agent. Furthermore, there are also complementary proposals that either integrate both approaches or suggest a hierarchical architecture to the reinforcement learning problem. It is therefore important to consider some of the main pros and cons of each learning process.

Model-based methods, as a consequence of their knowledge about the reward and transition function, can take fewer actions to learn state-action values (i.e. they are statistically

efficient) and achieve a better policy than model-free methods. However, this implies that the algorithm can learn quickly enough an accurate structure of the environment and the prospective nature of the model-based approach is computationally demanding if the decision tree is complex (Daw and Dayan, 2014). Another advantage of model-based learners is the way they readily adapt in situations where the reward structure changes rapidly. Given its knowledge of the consequences of actions at states, a policy can be planned without even requiring further experiences in the world. In addition, there is also the opportunity to use simulations to perform targeted exploration and plan a policy that explores states not recently visited or those the agent is uncertain about. Contrasting with this computational complexity, model-free strategies are computationally simple and require much fewer memory resources. However, due to their statistically inefficiency they are more inaccurate as well as less sensitive to changes in goal values. The reason for this is the fact that these approaches learn directly from experience and have to repeat many times the interactions with the real world before the values propagate all the way back to the initial states.

Given these reasons, it seems advantageous for a learning agent to have both strategies and take advantage of the benefits of each one according to the task design or the phase of learning. In fact, some hybrid algorithmic attempts have been made in order to combine model-free and model-based learning systems (Tamar et al., 2012). Notwithstanding, such approach prompts the issue of how interaction occurs and what formally arbitrates between the two when they run in parallel, due to occasional competing control signals. Daw and colleagues (Daw et al., 2005) proposed uncertainties in the value estimates as a way of assigning control, whereas others highlighted a speed-accuracy trade off criteria (Keramati et al., 2011) or the statistical properties of the environmental returns (Simon and Daw, 2011). In all these suggestions, theoretical simulation results were not only in line with the properties of each learning strategy described above, as for example the case of a less uncertain model-based system in initial phases of learning given its statistical efficiency, but they could also provide new insights into the experimental results from animal learning psychology.

The complex interactions between model-free and model-based control is under active investigation as various possibilities exist (Daw et al., 2011; Dayan, 2012a; Huys et al., 2012). Most easily viewed as a competitive process, the interaction between both model-free and model-based systems can also involve cooperation. Reinforcement learning theory also accommodates this idea of mutual assistance in working towards a common goal. The Deep blue chess-playing computer project is a good illustration of an application of such principle, where pruning algorithms used model-free values to substitute for the values of

whole branches when the decision tree gets too big to evaluate directly by a model-based search (Campbell et al., 2002).

Another example is the idea of the Dyna architectures (Sutton, 1990), which combines model-free learning with planning (Figure 1.4). In these algorithms, model-based trains model-free offline by simulating transitions and rewards and replaying experienced state–action pairs. Such simulations can then be used by a model-free learner in control of behaviour to amplify its experiences, reducing the number of environmental interactions needed to update values. If the selection of state-action pairs to be updated is not done randomly (as in Dyna), but instead based on state–action pairs whose values have recently changed, then the method is called prioritized sweeping.

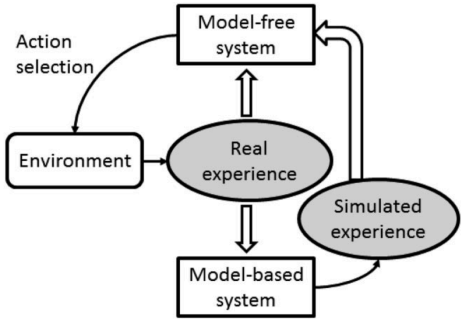
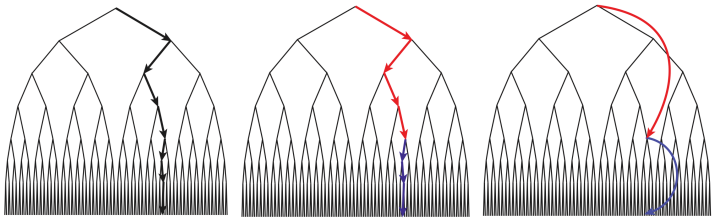


Fig. 1.4 **The Dyna arquitecture.** In Dyna, action selection is controlled by a model-free system that learns values from both real and simulated experiences. These hypothetical experiences are provided by a model-based system that takes advantage of its world's knowledge to replay state–action pairs. This way, both model-free and model-based systems are combined in order to achieve better learning performance. From Botvinick et al. (2009).

Finally, although the experimental work of this thesis involves a reinforcement learning problem with a relatively small state-space, it is relevant to mention some recent computational developments addressing the scaling problem.

Hierarchical reinforcement learning methods (Barto and Mahadevan, 2003; Hengst, 2012) extend standard reinforcement learning algorithms (both model-free and model-based approaches; see Botvinick and Weinstein, 2014) in order to address the problem of the exponential growth of the number of states and learning parameters in more complex scenarios. To achieve such goal the computations require some sort of abstraction, where representation of the learning problem only includes relevant features to behaviour. If the abstraction is applied to states it is called state or structural abstraction (Li et al., 2006), and if it is employed in actions then it is termed temporal abstraction (Sutton et al., 1999). The latter case, for example, involves a behavioural repertoire of sequences of primitive actions

or subroutines that can be considered as one-step options, allowing the system to have an option-specific policy and solve problems with fewer decision steps (Figure 1.5).



(a) (b) (c)

Fig. 1.5 **Temporal abstraction in hierarchical reinforcement learning.** In this Markov decision problem the agent has to perform six sequential binary decisions. Only one of the branches yields reward. Standard reinforcement learning methods (a) use only primitive actions, therefore, six decisions have to be made. Assuming the agent has previously learned the red and blue sequences of actions (b), then the scaling problem is greatly facilitated (c). From Botvinick et al. (2009).

Chapter 2

Reinforcement learning in the prefrontal cortex and basal ganglia

The neural mechanisms of reward-guided decision making and learning processes comprise complex cortico-subcortical circuits (Balleine and O'Doherty, 2009; Ito and Doya, 2011; Kable and Glimcher, 2009; Samejima and Doya, 2007; Sesack and Grace, 2009). Midbrain dopaminergic cells (Bromberg-Martin et al., 2010; Schultz, 2002), striatum (Balleine et al., 2007; Bornstein and Daw, 2011; Ding and Gold, 2013; Hikosaka et al., 2008), the prefrontal cortex (Kennerley and Walton, 2011; Rushworth et al., 2011) and the amygdala (Balleine et al., 2003; Baxter and Murray, 2002; Paton et al., 2006) are the key players in reward valuation and learning. This functional view is also supported by the anatomy (Haber and Knutson, 2009). The contemporary view of the reward circuit is of a spiralling and very interactive midbrain-striatal-prefrontal network (Figure 2.1), which takes into account principles of parallel processing (Alexander et al., 1986) as well as information convergence (Percheron and Filion, 1991).

The next part of this section will provide an overview of the basic anatomy of both the prefrontal cortex and striatum of primates with particular attention paid to its intrinsic and extrinsic connections. A more detailed description will be additionally provided for the brain regions experimentally targeted in this project: the dorsal bank of the anterior cingulate cortex (area 24c), the dorsolateral prefrontal region (area 46), the frontal pole (area 10) as well as the dorsal parts of striatum (caudate and putamen).

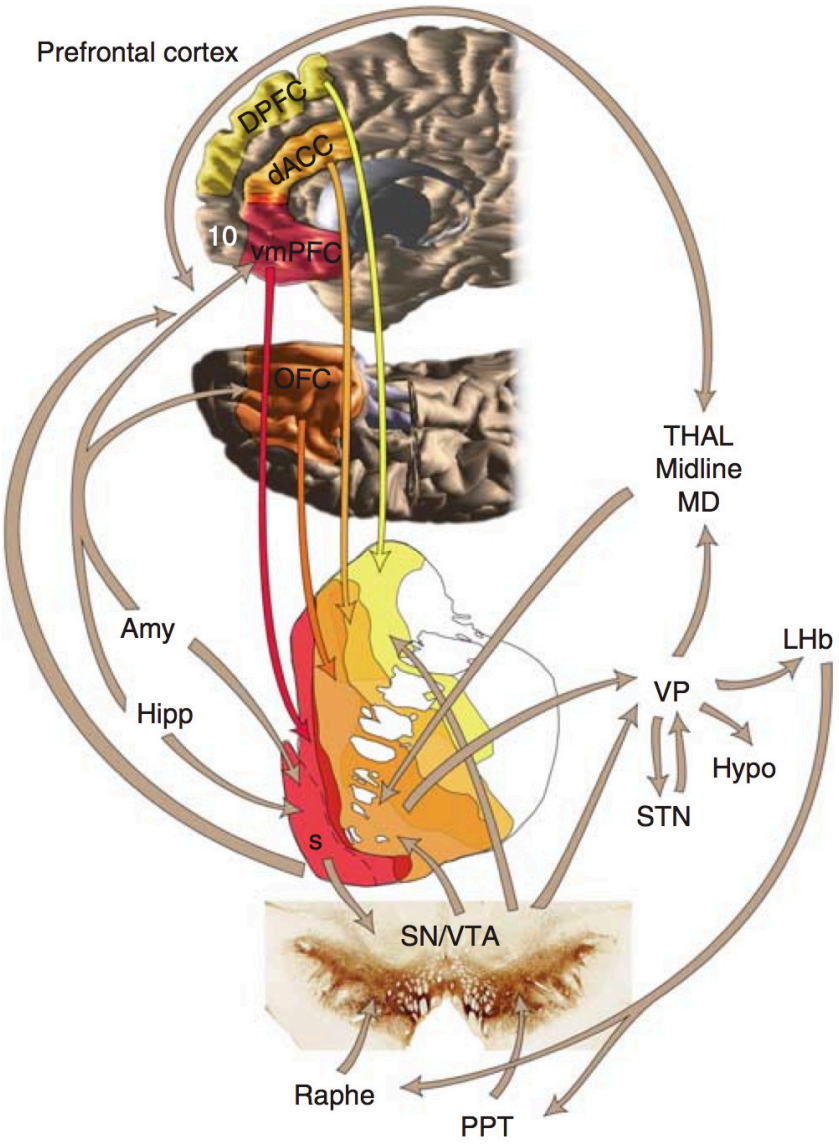


Fig. 2.1 **The prefrontal-basal ganglia reward circuit.** A complex network of connections within ventral and dorsal cortico-subcortical structures support reward learning. The former is mainly composed by the dopaminergic ventral tegmental area (VTA), nucleus accumbens and ventral parts of caudate and putamen, together with orbitofrontal (OFC) and ventromedial prefrontal cortex (vmPFC). The dorsal component includes mostly the substantia nigra (SN) pars compacta, putamen, the dorsolateral prefrontal cortex (DPFC) and motor/premotor regions. Intermediate areas include central parts of striatum as well as the anterior cingulate cortex (dACC). Abbreviations: amygdala (Amy); hippocampus (Hipp); shell of the nucleus accumbens (s); thalamus (THAL); mediodorsal (MD); lateral habenula (LHb); ventral pallidum (VP); subthalamic nucleus (STN); hypothalamus (Hypo); pedunculopontine tegmental nucleus (PPT). Adapted from Haber and Knutson (2009).

2.1 The anatomy of prefrontal cortex and basal ganglia

The prefrontal cortex networks

The prefrontal cortex consists of a heterogeneous region rostral to the motor areas of the frontal cortex in both humans and primates, despite some variations between species (Fuster, 2001; Petrides and Pandya, 2012; Wise, 2008). Studies of the prefrontal neuroanatomical connections have shown a high local connectivity (Averbeck and Seo, 2008). In addition, statistical analyses of these anatomical networks also suggest that every area within the prefrontal cortex is able to access all types of extra-prefrontal cortex information within two connections of its anatomical position, with hippocampal projections being particularly prevalent (Averbeck and Seo, 2008). Such connectivity pattern may explain why the prefrontal cortex has been associated with many high-order cognitive functions. Despite the broad and complex prefrontal networks, the existence of cytoarchitectural heterogeneity and results of many lesion as well as neurophysiological studies suggest that regions within these networks perform specific computations (Petrides and Pandya, 2004). In primates, the patterns of cortico-cortical and cortico-subcortical connections of the prefrontal cortex give rise to a generally accepted subdivision into three distinct anatomical networks: the *orbital*, the *medial* and the *lateral* networks (Murray et al., 2000; Petrides, 2005a; Saleem et al., 2014; Öngür and Price, 2000) (Figure 2.2 and 2.4).

The so-called orbital network consists solely of areas found on the orbital surface of the brain which show a high degree of interconnectivity (Figure 2.2). This network receives a huge sensory input from multiple areas including olfactory and gustatory cortices, visual-related areas in the inferior temporal gyrus as well as somatosensory and visceral areas in the insula and frontoparietal operculum (Öngür and Price, 2000). It also has important reciprocal connections with limbic structures such as the entorhinal and perirhinal cortices as well as the basal and lateral nuclei of the amygdala (Carmichael and Price, 1995a; Öngür and Price, 2000). Further relevant projections exist to distinct portions of the central region of the striatum, particularly to the lateral caudate nucleus and the ventromedial putamen, and to more central parts of the mediodorsal thalamic nucleus. These connections suggest that the orbital network receives and processes information about internal motivational states (Baxter and Murray, 2002). Despite being located in the orbital surface of the prefrontal cortex, the connections of areas 13a and 13b as well as 12o seem to span medial and orbital networks and, therefore, provide connectivity between the two networks (Barbas and Pandya, 1989; Price, 2007).

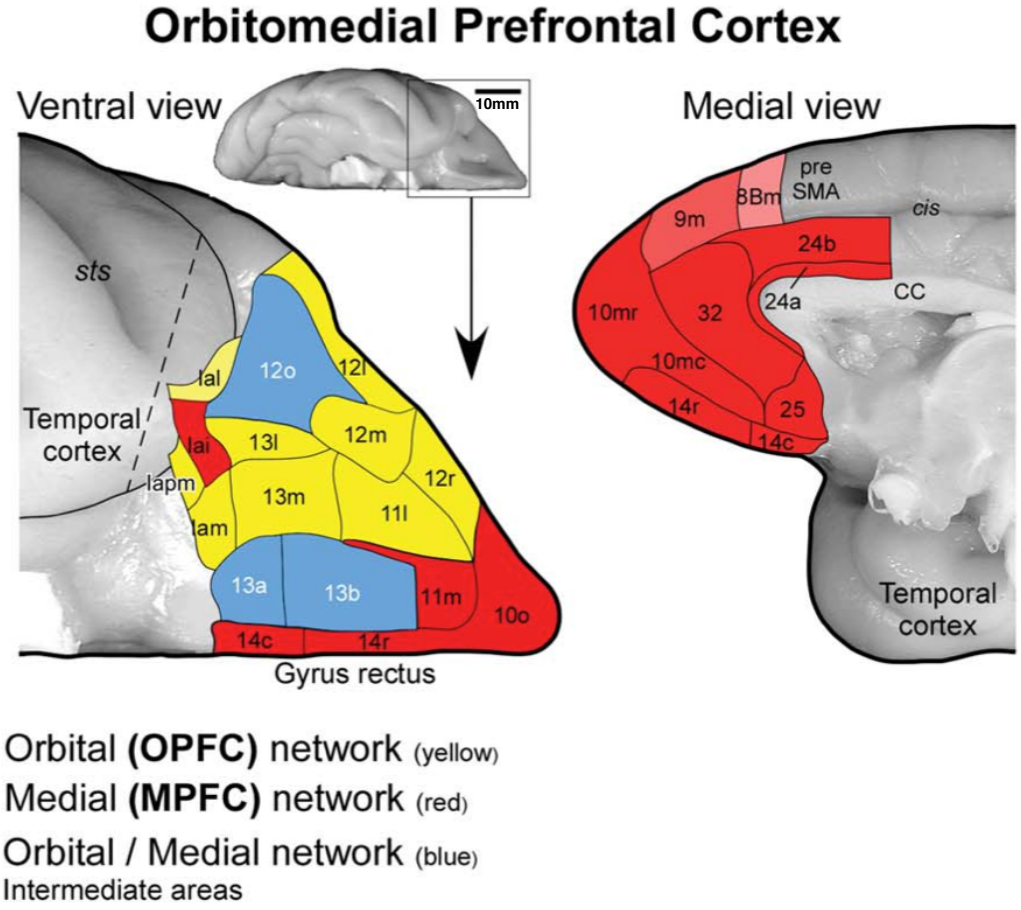


Fig. 2.2 **The medial and orbital prefrontal networks in monkeys.** The medial prefrontal network is shown by red shading and projects mainly to the hypothalamus and periaqueductal gray. The orbital network is yellow shaded and receives several sensory inputs. Blue shaded areas show an intermediate pattern, connecting both networks. Abbreviations: corpus callosum (CC); cingulate sulcus (cis); superior temporal sulcus (sts). Scale bar = 10 mm. Adapted from Saleem et al. (2014).

The medial network consists of brain areas located on both medial and orbital surfaces (Figure 2.2). Contrary to the orbital network, the medial network has few sensory inputs and it seems to be an area that rather projects outputs (Öngür and Price, 2000). It is known to send strong projections to the hypothalamus and periaqueductal grey, which are both associated with visceral and autonomic functions (Keay et al., 1994; Öngür and Price, 2000). Despite sharing similar limbic inputs with the orbital network, the medial network exhibits a host of distinct cortical connections. It has connections throughout the temporal lobe (dorsal temporal pole, rostral superior temporal gyrus and the dorsal bank of the superior temporal sulcus), the cingulate, retrospenial, entorhinal, posterior parahippocampal and dorsomedial prefrontal cortices (Price, 2007). Many of these areas are thought to encode a more complex and invariant source of sensory information than that of the areas connected to the orbital network (Price, 2007). The dorsal anterior cingulate cortex explored in this thesis is part of the medial network and corresponds to area 24c of the anterior cingulate cortex (ACC) region (Figure 2.3). The ACC encompasses areas within the ACC gyrus including areas 24, 25, 32 and area 9 within dorsal ACC, but controversies exist regarding its precise borders, cytoarchitectural nature and subdivisions (Sallet et al., 2011; Vogt et al., 2005). The view adopted in this thesis is that the middle of the dorsal bank of the cingulate sulcus is part of the ACC in non-human primates (Sallet et al., 2011). However, it is important to note that other authors regard this region as part of adjacent medial frontal cortex (area 9) and only consider areas ventral to the dorsal bank of the cingulate sulcus (ie, the ventral bank and below) to be part of ACC (Vogt et al., 2005). The rostral area 24c receives heavy limbic projections, particularly from the amygdala, lateral orbitofrontal cortex, and insula, justifying its role in the emotional influences of voluntary actions (Morecraft and Hoesen, 1998). Furthermore, this region also receives substantial input from medial temporal areas (perirhinal, entorhinal, and the parahippocampal cortices, hippocampal formation and the temporal pole) widely recognized in subserving memory-related functions (Arikuni et al., 1994; Barbas et al., 1999; Carmichael and Price, 1995b; Mishkin et al., 1997; Rosene and Van Hoesen, 1977; Squire et al., 2004; Vogt and Pandya, 1987). On the other hand, its vast efferents towards frontal motor cortices, corticospinal pathways and brainstem nuclei implicate this region in visceromotor, skeletomotor, and endocrine outflow (Vogt et al., 1992). In fact, prefrontal efferents to the ACC are stronger than its reciprocal prefrontal projections, with projections being strongest to the rostral portion of area 24 (Arikuni et al., 1994; Bates and Goldman-Rakic, 1993; Pandya et al., 1981; Vogt and Pandya, 1987). Overall, this connectivity suggests a critical role of ACC in adaptive behaviour by integrating past and present experiences with motivational significance (Sallet et al., 2011).

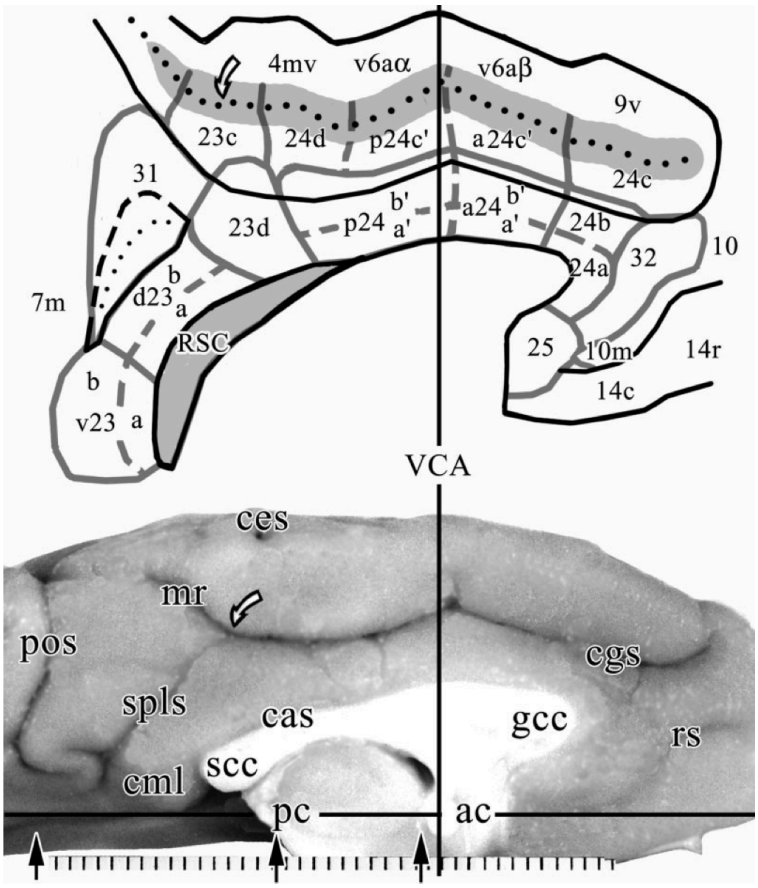


Fig. 2.3 **The anterior cingulate cortex and its subdivisions.** The cingulate sulcus is divided into two parts by a vertical line passing through the anterior commissure (VCA): anterior (ac) includes areas 24c and 24c'; and posterior (pc) composed of the posterior part of area 24c', which contains the rostral cingulate motor area. Abbreviations: retrosplenial cortex (RSC), rostral sulcus (rs), cingulate sulcus (cgs), genu of the corpus callosum (gcc), anterior commissure (ac), central sulcus (ces), marginal ramus (mr), callosal sulcus (cas), splenium of the corpus callosum (scc), splenial sulcus (spls), caudomedial lobule (cml), parieto-occipital sulcus (pos), posterior commissure (pc). Adapted from Vogt et al. (2005).

The lateral network includes several brain areas of the lateral surface of the prefrontal cortex (Figure 2.4), including the dorsolateral prefrontal cortex (area 46d) and the frontal pole (area 10) regions explored in this thesis. Recently, an anatomical subdivision of this network has been proposed based on the unique patterns of cortico-cortical connections: the dorsolateral prefrontal cortex (dorsal region to the principal sulcus and frontal pole), the ventrolateral prefrontal cortex (ventral region to the principal sulcus) and the caudolateral prefrontal cortex (around the arcuate sulcus and parts of the caudal principal sulcus) (Saleem et al., 2014). Brain areas of each subdivision are connected primarily with other areas in the same subnetwork, but some connections also exist across the principal sulcus. The ventrolateral prefrontal cortex is strongly connected with the orbital network and similarly with somatosensory brain regions (Saleem et al., 2014). By contrast, the dorsolateral prefrontal cortex (areas 9d/m and 10, but not area 46d) contains prominent links with regions of the medial network as well as with the posterior cingulate cortex, the superior temporal gyrus, superior temporal sulcus and the posterior parahippocampal cortex (Saleem et al., 2014). Despite some similarities with the medial network, the dorsolateral prefrontal cortex receives much fewer inputs from limbic structures and projects much less to visceral control regions (An et al., 1998; Saleem et al., 2014; Öngür et al., 1998). Unlike most dorsolateral prefrontal cortex, the projections of area 46d (specially its midportion) are unique and target the medial parietal and parieto-occipital areas. As a whole and given its connections, the dorsolateral prefrontal cortex has been more often implicated in maintaining the representation of goals and means to achieve them, important for behaviour monitoring and subsequent choice (Miller and Cohen, 2001; Petrides, 2005b). A more specific involvement in multitask processing has been attributed to the frontal pole of humans, which is detailed later in this chapter (Burgess et al., 2000; Koechlin et al., 1999). It is important to note that in a comparative functional connectivity study it has been suggested that while human medial FP resembles macaque FP, human lateral FP resembles dorsolateral area 46 in the macaque as opposed to macaque FP (Neubert et al., 2014). Nonetheless, one of the potential limitations of the study is the fact that these patterns of connectivity were obtained in different cognitive states (i.e. anesthetized animals versus restive awake humans), making it hard to directly relate findings from both species.

A relevant feature to highlight is the interconnectivity between the three prefrontal regions specifically addressed in the experimental section of thesis – the FP, the ACC and the DLPFC. The dorsolateral prefrontal cortex has reciprocal connections with the dorsal anterior cingulate cortex, representing between twenty and thirty percent of their prefrontal connections (Averbeck and Seo, 2008). Several studies have found reciprocal connections

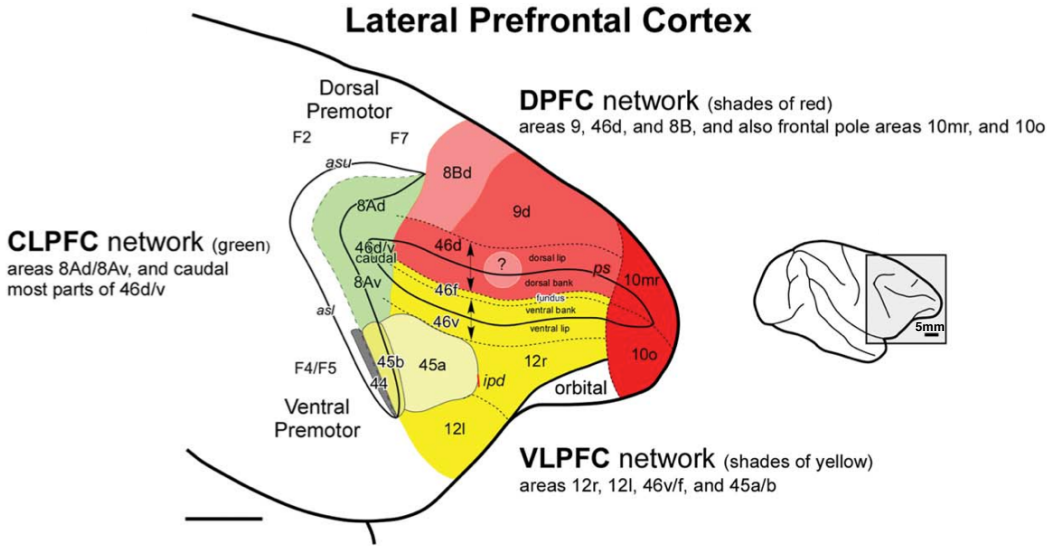


Fig. 2.4 **The lateral prefrontal network and its subdivisions.** The lateral prefrontal cortex is divided into: the ventrolateral prefrontal cortex (VLPFC; shades of yellow), the dorsolateral prefrontal cortex (DPFC; shades of red/orange) and the caudolateral prefrontal cortex (CLPFC; green shading). The region indicated by a question mark (pale red in the mid-portion of 46d) has a pattern of connections different from the patterns of other parts of the dorsolateral prefrontal cortex as it targets the medial parietal and parieto-occipital areas. Abbreviations: arcuate sulcus lower limb (asl); arcuate sulcus upper limb; Adapted from Saleem et al. (2014).

between area 46d and the rostral aspect of area 24c, particularly in the area just rostral to the genu of the corpus callosum (Barbas and Pandya, 1989; Bates and Goldman-Rakic, 1993; Morecraft and van Hoesen, 1993; Vogt and Pandya, 1987). The anterior cingulate cortex projects back to area 46d, but these efferents are not consistently seen across studies and may only be present in the rostral part of area 46d. Projections from area 10 to dorsal anterior cingulate cortex are restricted to its more rostral part, and also only the more rostral parts of the anterior cingulate cortex project back to area 10 (Petrides and Pandya, 2007). A recent study aimed to specifically study the anatomical relationship between the FP, ACC as well as DLPFC, and ended up proposing a circuit of interactions based on their connectivity pattern (Figure 2.5) (Medalla and Barbas, 2010). Interestingly, the authors found that previously described projections from anterior cingulate cortex to dorsolateral prefrontal cortex have a higher prevalence of large boutons than other dorsal prefrontal inputs, reflecting better synaptic efficacy. In addition, the anterior cingulate cortex projections innervate preferentially inhibitory neurons in area 46d and large boutons on spines of excitatory neurons in area 10. These results led the authors to suggest that such specific synaptic findings may be key when explaining the anterior cingulate role in modulating working memory processes and its involvement in multi-task functions (Badre and Wagner, 2004; Ghering and Knight, 2000; Koechlin et al., 2000). However, it is important to note that the anterior cingulate area investigated in this anatomical study corresponded to area 32, although strong similarities with area 24c exist regarding other connections.

The basal ganglia anatomy and its prefrontal connections

The basal ganglia are a group of evolutionary primitive forebrain nuclei in both cerebral hemispheres that receive inputs from nearly all cortical areas and project back, via the thalamus, primarily to the frontal cortex. Even the very first anatomical descriptions prompted speculations about the crucial role of these structures in dealing with *sensory impulses* as well as *the execution of willed action* (Thomas, 1664). In mammals, the basal ganglia are generally divided into dorsal and ventral complexes (Butler and Hodos, 2005). In the primate brain, the dorsal part is composed by what was initially called *corpus striatum*, and includes most parts of the caudate nucleus, putamen and globus pallidus (these last two structures used to be referred as the lentiform nucleus). A later anatomical classification proposed the current nomenclature of striatum (caudate and putamen) and pallidum (dorsal pallidum, composed of the external and internal segments of the globus pallidus, and the ventral pallidum) (Vogt and Vogt, 1941). Importantly, the separation between the caudate

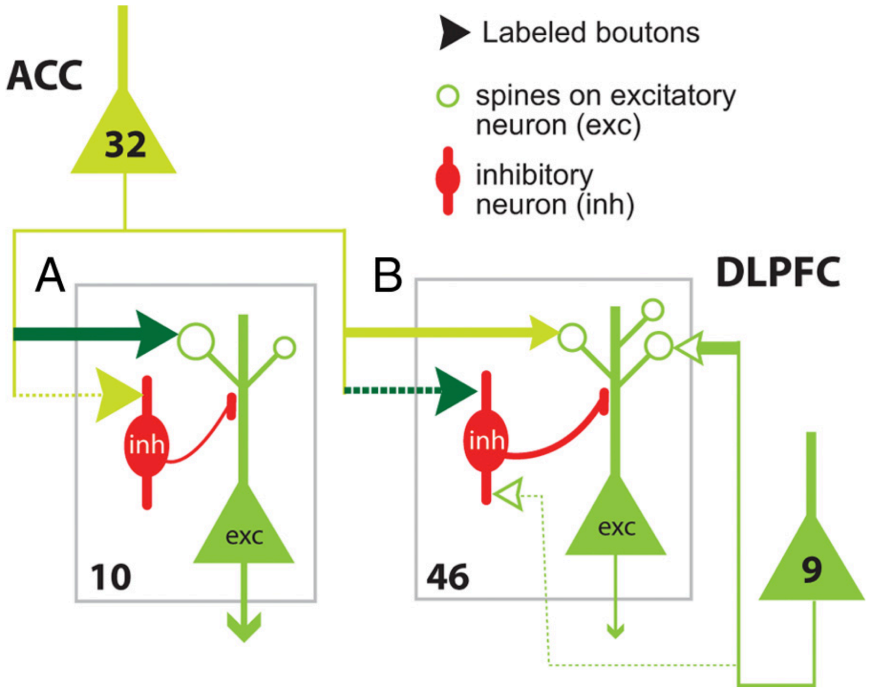


Fig. 2.5 **Connectivity between anterior cingulate cortex, dorsolateral prefrontal cortex and frontal pole.** Anterior cingulate cortex (area 32) projects through large boutons (dark green, larger arrowhead) preferentially to spines of excitatory neurons (light green open circles) of frontal pole (area 10). By contrast, projections to dorsolateral prefrontal cortex (area 46) target dendritic shafts of inhibitory neurons. Overall, these projections have larger boutons than pathways within dorsolateral prefrontal cortex (area 9 to area 46, represented by the light green open arrows). Line thickness represents prevalence of pathways, and size of arrowhead represents size of boutons. Adapted from Medalla and Barbas (2010).

and the putamen is merely a structural one, based solely on the internal capsule separation, not a functional one. On the other hand, the ventral complex comprises the nucleus accumbens, medial and ventral parts of the striatum, as well as striatal parts of the olfactory tubercle (Haber and McFarland, 1999). Due to histochemical and connection differences, the nucleus accumbens is further subdivided into shell and core. In addition to these two main complexes, the mesencephalic dopaminergic system and the subthalamic nucleus are also part of the basal ganglia circuits.

The striatum is the region where most inputs to the basal ganglia converge, receiving projections from all of cerebral cortex, thalamus and brainstem, in particular from midbrain dopaminergic cells. The overall view is that the dorsolateral striatum (in particular the central and caudal aspects of putamen) receives sensorymotor afferents, the inputs to central striatum (more the caudate than putamen) arise from associative cortical areas and the ventromedial striatum receives predominantly limbic projections (Haber and Gdowski, 2004). The present review will focus on both cortical and dopaminergic projections, given that interconnected thalamic and cortical areas project to the same region of the striatum.

The cortico-striatal projections follow a longitudinal topography as well as a medial to lateral one, terminating in a more patchy and interdigitated manner than that found in rodents (Joel and Weiner, 2000). Connections between prefrontal cortex areas and striatum normally respect a degree of topography and depend on the prefrontal network they belong to (Ferry et al., 2000; Haber and Gdowski, 2004; Yeterian and Pandya, 1991). Nevertheless, prefrontal-striatal projections are not fully convergent even within networks (Fig. 2.6).

The targets of prefrontal-striatal projections from medial network areas span both ventral and dorsal parts of striatum (Wise, 2008; Öngür and Price, 2000). Projections from more ventromedial areas (area 25 and area 14c) aim primarily the ventral striatum, terminating in the core and shell of the nucleus accumbens (Ferry et al., 2000; Nakano et al., 1999). Area 32 also shows projections to the core of the nucleus accumbens and rostral ventral putamen, but they are predominant in the medial portion of the head, body and tail of the caudate (Ferry et al., 2000). More anterior areas of the medial network (including areas 10o, 10m and 11m) show projections that are restricted to the medial edge of the caudate (Ferry et al., 2000). Area 24 has interesting patterns of connectivity as it either targets different regions within the ventral striatum or both ventral and dorsal striatum. Areas 24a and 24b target mostly the lateral ventral striatum, but also have some connections to medial ventral striatum (Ferry et al., 2000; Kunishio and Haber, 1994). Medial regions of area 24c show particularly strong connectivity with the core of the nucleus accumbens, while more lateral regions exhibits connections with dorsal striatum (Kunishio and Haber, 1994).

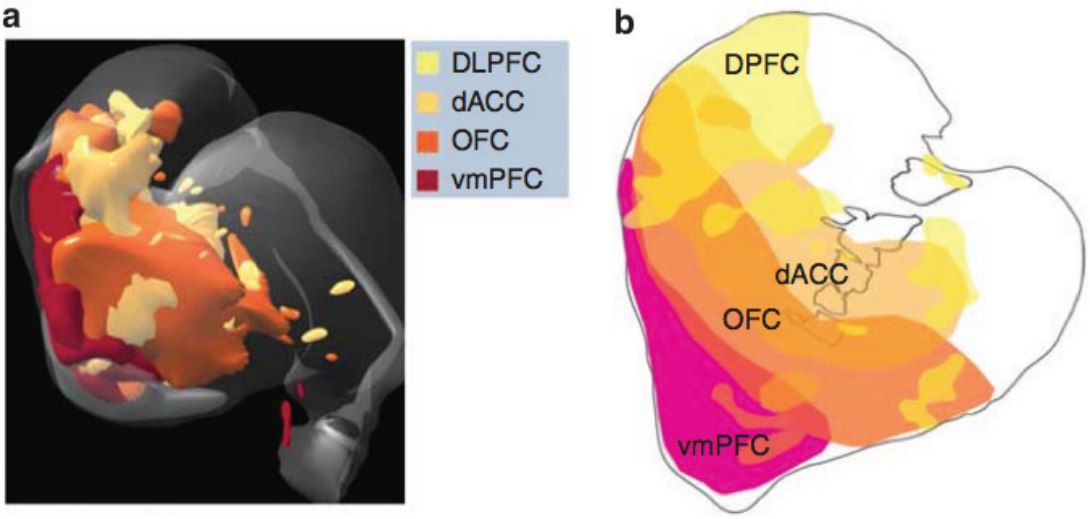


Fig. 2.6 **Convergence of the prefrontal-striatal projections.** A 3D **a)** and 2D **b)** view of a reconstruction of the striatal inputs from prefrontal regions illustrating regions of convergence of inputs. From Haber and Knutson (2009).

Finally, the motor cingulate cortex also sends an overlapping projection to the dorsal region of the striatum (i.e., both caudate and putamen) (Kunishio and Haber, 1994; McFarland and Haber, 2000).

Most areas of the orbital network are connected extensively with ventral and central parts of the head of the caudate nucleus, the dorsal edge of the nucleus accumbens as well as more medial areas of putamen (Ferry et al., 2000; Haber et al., 1995; Öngür and Price, 2000). This connectivity shows a slight degree of overlap with the connectivity pattern of the medial network, although injections into lateral caudate nucleus caused staining almost exclusively within orbital network areas (Ferry et al., 2000).

The dorsolateral prefrontal cortex substantially innervates the striatum but it targets predominantly the more anterior and central region of the caudate nucleus (Arikuni and Kubota, 1986; Selemon and Goldman-Rakic, 1985). These projections are topographically organised and interfaces with several other cortical inputs. Injections made into area 46 and 9 show terminations in the lateral portions of the caudate nucleus and in medial putamen (Calzavara et al., 2007; Yeterian and Pandya, 1991).

Projections from other cortical areas to the striatum have not been studied in such detail as the ones from the prefrontal cortex. The dorsolateral and central parts of posterior putamen are the targets for the vast majority of motor and premotor projections, with the latter ones extending more rostrally than the former ones (Kemp and Powell, 1970). The same region receives somatosensory parietal inputs with a similar somatotopic arrangement

(Künzle, 1977). In regards to saccadic eye movements, the frontal eye fields and the supplementary eye fields areas project to the central and more lateral part of the head and body of the caudate nucleus as well as to ipsilateral putamen (Künzle, 1977).

From a pure anatomical perspective, the dorsolateral striatum is considered the sensorimotor striatum, but it has also been functionally implicated in movement planning, execution and learning. In regards to temporal lobe projections, despite targeting vast areas of the striatum they show relevant preferences (Selemon and Goldman-Rakic, 1985; Van Hoesen et al., 1981; Yeterian and Pandya, 1998; Yeterian and Hoesen, 1978). Superior temporal gyrus inputs overlap with those from dorsal lateral network in the central half of the caudate nucleus. In contrast, inferior temporal areas terminals interdigitate more ventrally with those from the medial and orbital prefrontal networks. Finally, occipital visual fibers terminate in posterior parts of the body of the caudate nucleus (Saint-Cyr et al., 1990).

The dopamine system interacts closely with the striatum. In fact, the striatum receives major dopaminergic projections but it also contributes to the major inputs to the dopamine system. Dopaminergic cells are present in a dorsal tier (the ventral tegmental area, the dorsal group of substantia nigra pars compacta and the retrorubral area) and in a ventral tier (densocellular and ventral group of the substantia nigra pars compacta) of the midbrain. The way dopamine projects to striatum follows a reversed dorsal-ventral topography, with the dorsal tier projecting to the ventral striatum and the ventral tier projecting to the dorsal striatum (Carpenter and Peter, 1972; Haber et al., 2000; Lynd-Balta and Haber, 1994). More specifically, the ventral tegmental area projects mainly to the shell of the nucleus accumbens and the dorsal group of substantia nigra pars compacta to most of the remaining ventral striatum. Regarding the ventral tier, the densocellular group of substantia nigra pars compacta projects primarily to the central striatal area (where most dorsal lateral prefrontal cortex terminals end) and the ventral group of substantia nigra pars compacta primarily to the dorsolateral striatum. It is also important to note that ventral striatum dopaminergic inputs come from a limited midbrain region, whereas projections to the dorsolateral striatum are derived from a much wider area of the midbrain (Haber et al., 2000).

Both ventral tegmental area and substantia nigra pars compacta also send projections to various parts of the prefrontal cortex (Porrino and Goldman-Rakic, 1982). Injections into areas 9 and 46 have shown connections with anterior parts of ventral tegmental area and antero-medial and antero-dorsal substantia nigra pars compacta (Porrino and Goldman-Rakic, 1982). Injections into area 10, 11, 12 and 13 all find labelled neurons throughout the ventral tegmental area but only area 10 appears to connect with substantia nigra pars compacta (Porrino and Goldman-Rakic, 1982). The area most strongly connected to ven-

tral tegmental area appears to be area 24 which connects extensively (Gaspar et al., 1989; Porrino and Goldman-Rakic, 1982). Even within area 24 there appears to be a sharp transition from higher to lower dopamine innervation between areas 24b and area 24c (Williams and Goldman-Rakic, 1998). Dopaminergic neurons have also been described in area 32 on the medial wall of the prefrontal cortex (Raghanti et al., 2008). Finally, these midbrain dopamine projections seem to be topographically organised based upon which part of the prefrontal the projections terminate. In regards to projections from the prefrontal cortex to dopamine neurons, they are not very abundant but they arise from regions such as anterior cingulate, dorsolateral prefrontal and orbital cortices (Frankle et al., 2006).

An overall summary of the prefrontal-basal ganglia connections based on most of the findings discussed above is shown schematically in Figure 2.7.

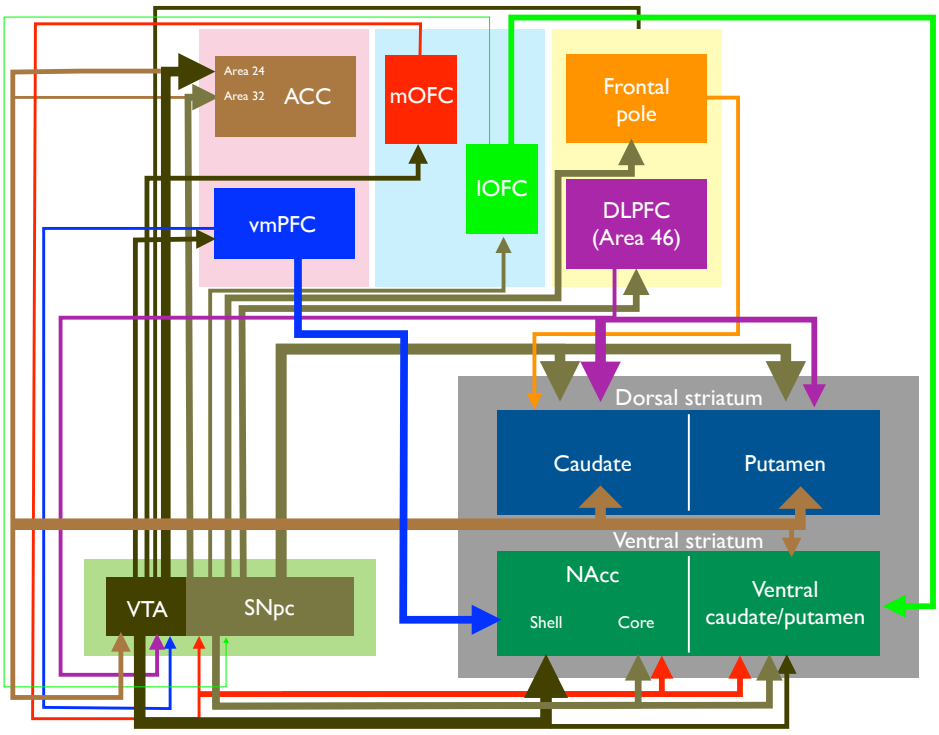


Fig. 2.7 **Schematic representation of prefrontal-basal ganglia connections.** Line thickness denotes strength of projections relative to the respective region. Abbreviations: ACC: anterior cingulate cortex; vmPFC: ventromedial prefrontal cortex (areas 14r/c); mOFC: medial orbitofrontal cortex (areas 13a/b/m/l); lOFC: lateral orbitofrontal cortex (areas 12l/m/o/r); DLPFC: dorsolateral prefrontal cortex; VTA: ventral tegmental area; SNpc: Substantia nigra pars compacta; NAcc: nucleus accumbens.

2.2 Lesion studies and neurophysiology of prefrontal and basal ganglia

Dopamine, reward and prediction errors

Dopamine neurons were found to have interesting learning and motivational properties (Ljungberg et al., 1992; Romo and Schultz, 1990; Schultz et al., 1993; Schultz and Romo, 1990). Theoreticians working on learning models related this pattern of response to the one predicted by a temporal difference prediction error signal, which drives value updates in reinforcement learning (Montague et al., 1996, 1995; Schultz et al., 1997). Further studies corroborated predictions derived from the theory (Fiorillo et al., 2003; Tobler et al., 2005). When more quantitative predictions of the model were tested, dopaminergic cells encoded precisely the difference between the current reward and an exponentially weighted average of previous rewards (Bayer and Glimcher, 2005).

The reward prediction error signal of dopamine phasic activity, highlights a potentially substantial role in model-free learning. Dopamine firing in monkeys performing a probabilistic decision task with occasional forced choices reflected both a valuation of the current state as well as the action taken by the animal (Morris et al., 2006). These findings are very much in agreement with a prediction error signal being used in a SARSA-like algorithm (Niv et al., 2006), although in rodents the signal seems to be more in line with Q-learning (Roesch et al., 2007). Despite such role in model-free learning, increased dopamine levels have been associated with an enhancement of model-based over model-free control (Wunderlich et al., 2012b). Concomitantly, depleting dopamine can boost model-free control (de Wit et al., 2012a). A strong possibility for these unexepected findings may be the fact that those effects could in part depend on dopamine's actions on functions implemented in prefrontal cortex. In fact, the disruption of the dopaminergic system has long been associated with executive cognitive deficits (Brozoski et al., 1979; Sawaguchi and Goldman-Rakic, 1991) and its neuromodulatory role has also been included on computational views of working memory (Durstewitz et al., 2000; Vijayraghavan et al., 2007). The details of these effects in structures such as the prefrontal cortex remain largely unknown (Seamans and Yang, 2004). Nevertheless, such functional findings are consistent with the strong anatomical projections from midbrain dopamine cells to striatum and prefrontal cortex, where model-free and model-based computations are mostly undertaken.

The double-faced striatum

If we now turn our attention to the primate neurophysiology of model-free structures, in one of the first studies revealing striatal plasticity during behavioural learning a specific neural pattern emerged as the number of sessions and learning increased (Aosaki et al., 1994b). This striatal activity did not seem to reflect just a direct consequence of a dopamine input, although the dopaminergic teaching signal is necessary for its formation (Aosaki et al., 1994a). In monkeys performing a sequential motor task, different striatal cells also increased their firing rate for either new or well learned sequences (Miyachi et al., 2002). Most importantly, this functional segregation had a clear anatomical match to more anterior and posterior parts of the striatum, respectively (Miyachi et al., 1997). In a very elegant study testing directly model-free estimates, cells in dorsal striatum were found to encode state-action values (Samejima et al., 2005). These values obtained with a Q-learning algorithm fitted to the animal's choices. In addition to these neurons, another study reported caudate cells encoding chosen values following the response (Lau and Glimcher, 2008). Action values appeared earlier than the chosen value. This temporal profile is in agreement with reinforcement learning principles. State-action values are useful in guiding the action selection process as they reflect cached estimates of the available options based on experience. By contrast, chosen values represent the value of a choice that has already passed through the selection process, relevant for an evaluative process such as the dopaminergic prediction error computation. Very similar findings were observed in rodents, but action value neurons were found in fewer number, and present in both dorsal and ventral parts of the striatum (Ito and Doya, 2009). This observation in the ventral striatum is slightly surprising given the fact that activity in this region has been more often associated with expected reward in relation to pavlovian conditioned associations (Cromwell and Schultz, 2003; Day et al., 2006; Shidara et al., 1998). This latter predictive function of ventral striatum can also be viewed as an interest in state value, irrespective of the action choice, often analysed as the summed value of the various options (Cai et al., 2011).

Indirect evidence for potential model-based computations has also been found in striatal neurons. Most of this is because of the involvement of more anterior parts of the caudate in initial phases of learning, which is in line with the more advantageous role of model-based computations at the start of the learning process. This striatal role was brought to attention while simultaneously comparing its neural activity with the prefrontal cortex (Pasupathy and Miller, 2005), where both the rise time and the peak of neuronal activity appropriate for a correct response emerged earlier in caudate than in the dorsolateral prefrontal cortex. Moreover, some of these striatal cells can sustain their activity from trial to trial to hold

information about correct options (Histed et al., 2009). In addition, caudate can play a role in sequence representations (Seo et al., 2012) and is also capable of comparing integrated information such as temporally discounted values (Cai et al., 2011). When microstimulation was applied to the head of the caudate during the feedback-reward period of a correct visuomotor association, the rate of learning increased significantly when compared to non-stimulated blocks, leading animals to improve the acquisition of new associations and reach learning criteria in fewer trials (Williams and Eskandar, 2006).

The prefrontal cortex

The areas of the prefrontal cortex most often involved in reward-guided learning and decision making involve the orbital frontal cortex, the anterior cingulate cortex, the dorsolateral prefrontal cortex and the frontal pole (Kennerley and Walton, 2011; Rushworth et al., 2011).

Orbital network

Lesion studies in the orbitofrontal cortex of monkeys support its involvement in establishing associations between choices and outcomes, particularly in dynamic scenarios (Rudebeck et al., 2008; Rudebeck and Murray, 2008; Walton et al., 2010). Consistent with a proposed anatomical division for prefrontal cortex in a visceromotor medial network (medial wall and ventromedial areas of prefrontal cortex) and a sensory orbital system (posterior, central and lateral areas in the orbital surface) (Carmichael and Price, 1996, 1995c; Öngür and Price, 2000), subtle differences were observed within the orbitofrontal cortex regions(Noonan et al., 2010). Lesions of the ventromedial part hint towards a more value comparison function, whereas monkeys with lateral orbitofrontal lesions showed instead a credit assignment problem due to lack of integration of reward and choice history. Despite not requiring knowledge of a model of the environment, these findings suggest attention to the reward structure. In agreement with this interpretation, recent lesion evidence emphasise the role of the lateral orbitofrontal cortex in motivational updating valuations given that such lesions impair reinforcer devaluation (Rudebeck and Murray, 2011; Rudebeck et al., 2013).

Not many neurophysiology studies addressed this learning function in non-human primates. Neurons in the orbitofrontal cortex modify their responses in order to accommodate qualitative (Hikosaka and Watanabe, 2000; Thorpe et al., 1983) or quantitative (Kobayashi et al., 2010; Tremblay and Schultz, 1999) changes in incentive value. In simple value comparison scenarios where explicit information about the outcomes is available, OFC neurons

encode sensory properties of the reward (such as taste) as well as offer and chosen values (Padoa-Schioppa and Assad, 2006). Their activity also reflects a trial-by-trial evaluation of the value difference between the current and the previous trial (Kennerley et al., 2011). This previous choice information can be used to contextualise the current value, with potential relevance in learning and adaptive behaviour. When directly tested in a reversal learning task with aversive and appetitive outcomes, the orbitofrontal cortex neurons preferentially encoded rewarding stimuli and updated this appetitive value more rapidly (Morrison et al., 2011). Regarding potential differences between a medial and an orbital network, lateral orbitofrontal cortex activity seems related to external or sensory information processing, whereas the medial part is more involved in internal motivational representations (Bouret and Richmond, 2010; Monosov and Hikosaka, 2012).

Medial network

Lesion studies in the anterior cingulate cortex of monkeys revealed its key role in reinforcement learning. Despite not being essential for switching behaviour, lesioned animals show impairment in continuing with the same rewarding option following several reinforcements (Kennerley et al., 2006). Inactivation of the anterior cingulate cortex not only make animals repeat non-valuable choices that were valuable prior to the inactivation (Amiez et al., 2006), but also interferes with action-outcome associations (Shima and Tanji, 1998). Overall, these findings suggest a role of this structure in tracking the reward history of choices, in harmony with representations of outcome volatility found in the anterior cingulate cortex of humans (Behrens et al., 2007). It also seems to be capable of integrating both context and outcome values (Buckley et al., 2009). The above features could reflect the use of a cognitive-like map for model-based learning. If error related activity found in some studies (Debener et al., 2005; Holroyd et al., 2004) is taken into account, it could be also argued that this area is involved in monitoring the model structure of which future choices are based (Rushworth and Behrens, 2008).

Neurophysiology studies do seem to support some of these speculations. Cells in cingulate motor area process reward reductions relevant to an adaptive selection of motor acts (Shima and Tanji, 1998). In foraging decisions, the neuronal activity correlated with the value of leaving the current source of reward and reached its maximum just before the animal decided to leave and try a new option (Hayden et al., 2011a). Further studies also highlighted the integrative role of connecting actions and rewards (Hadland et al., 2003; Hayden and Platt, 2010; Kennerley et al., 2009, 2006; Matsumoto et al., 2003). This learn-

ing feature could also be used when stimuli are relevant for optimal performance (Amiez et al., 2006). Other types of evidence suggest that the reward history modulation found in this area could be a reward prediction error similar to the dopaminergic prediction error signal (Amiez et al., 2005; Kennerley et al., 2011; Matsumoto et al., 2007; Seo and Lee, 2007). However, two important shortcomings prompt prudence on this model-free account. First, the mentioned studies used either simple learning tasks or focused on the action selection process with values learnt from direct experience. Secondly, the reward-related signals observed in the anterior cingulate cortex include complex feedback information (Quilodran et al., 2008), as it encodes multiple decision variable knowledge (Hayden and Platt, 2010; Kennerley and Wallis, 2009a) and outcomes that have been observed but not directly experienced (Hayden et al., 2009). Such elaborated reward signals are more likely to be computed by a model-based region.

Lateral network

Classical studies of dorsolateral prefrontal cortex lesions in monkeys were the first evidence suggesting an involvement of this region in working memory processes (Jacobsen, 1935). Further lesional studies confirmed its relevance in maintaining behaviourally relevant representations as observed by the significant impact in delayed-response performance (Funahashi et al., 1993). In addition to deficits in working memory maintenance, it was also found that impairment of areas 9 and 46 compromise the learning of a memory task in which the monitoring requirements within working memory are minimized by having a rule, where each correct choice is specified by the preceding one (Petrides, 1995). Such role in supporting working memory for abstract rule learning seems to be specific of the dorsal prefrontal cortex region around the principal sulcus and could be important for behaviour control in dynamic environments (Buckley et al., 2009). Electrophysiological evidence also favours the involvement of this region in working memory (Levy and Goldman-Rakic, 2000) and cognitive control (Miller and Cohen, 2001) theories. Dorsolateral prefrontal cortex single neurons are able to keep a persistent response for behaviourally relevant information, such as spatial location (Funahashi et al., 1989) or object identity (Ó Scalaidhe et al., 1999). Furthermore, single-neuron activity in dorsolateral prefrontal cortex is not only capable of encoding rule information (Wallis et al., 2001) but also represents the competition level between two potential matching rules for a choice (Mansouri et al., 2007).

Despite the involvement in specific cognitive tasks, only more recently the neurophysiology findings of dorsolateral prefrontal cortex have been interpreted within the framework

of reinforcement learning theory (Lee and Seo, 2007; Lee et al., 2012). Several studies have found that expected reward modulates the single-neuron activity involved in working memory processes (Leon and Shadlen, 1999; Watanabe, 1996). This encoding of working memory information and reward is stronger in the ventrolateral subdivision than in the dorsolateral prefrontal cortex (Kennerley and Wallis, 2009b). Importantly, such firing rate modulation may indeed contribute to the animal's decision-making process. In a study where the animal was first told about the magnitude of reward and then a stimuli informed the response to be made, the influence of expected value on neuronal activity was only present before the instructive cue was shown and not after the specific response was performed (Amemori and Sawaguchi, 2006). However, most of these value-related signals were observed either in not very complex tasks or in relatively stable situations. In a study with a more dynamic environment where subjects made binary decisions in a matching pennies task, the firing rate of neurons around the caudal principal sulcus encoded the difference in the value functions of both choices as derived by a reinforcement learning model (Barraclough et al., 2004). This activity was influenced by the animal's choice as well as by past outcomes, and it was often maintained across intertrial intervals. Furthermore, recordings in the same region in a task where monkeys had to learn a sequence of choices, showed not only neuronal preference for particular movements within the sequence but also different responses to the same movement depending on the specific sequence (Averbeck et al., 2006). These findings not only highlight an involvement of these signals in a reinforcement learning process but also implicate the dorsolateral prefrontal cortex in the process of action-outcome evaluation. Interestingly, the disruption of dorsolateral prefrontal cortex with transcranial magnetic stimulation in humans impaired MB behaviour in favour of a behaviour driven by MF control (Smittenaar et al., 2013). Furthermore, the role in working memory, rule learning as well as the context-dependent outcome-related activity suggest a potential role of dorsolateral prefrontal cortex in model-based reinforcement learning.

The frontal pole is the most anterior area of the frontal cortex and relatively larger in percentage of human brain volume than in great apes (Semendeferi et al., 2001). Its cytoarchitectural organization promotes heavy cortico-cortical connections with a consequent potential role in coordinating different types of information. These properties have led some authors to place the frontal pole at the pinnacle of a possible rostro-caudal axis processing gradient (Badre and D'Esposito, 2009). Furthermore, the region's development starts relatively late with the highest rates of brain growth occurring around the second half of the first decade of life (Sowell et al., 2004). This is a relevant point given the evidence of habitual control dominance found in younger ages (Klossek et al., 2008). Frontal pole lesions

in humans do not seem to have a strong and widespread effect on classic executive tests. However, rule-break behaviour and problems with planning have been found in complex multi-task scenarios (Burgess et al., 2000). These deficits seem to be more prominent in situations requiring maintenance of goals in a complex environment as well as when the correct way of behaving is underspecified (Burgess et al., 2007). This is consistent with the neuroimaging evidence reporting engagement in branching processes (Koechlin et al., 1999), exploratory decisions (Daw et al., 2006) and monitoring of alternative choices values (Boorman et al., 2011, 2009).

Two very recent studies reported subtle differences compared with controls in the behaviour of monkeys with FP lesions. In one of these studies, subjects performed various tasks and while choosing between new alternatives early stage errors rapidly decreased with learning in control animals. By contrast, FP lesioned animals, showed no such rapid learning but were indistinguishable from controls in later phases of learning. The authors proposed an important role of FP in rapid learning of the relative values of wide-ranging novel alternatives (Boschin et al., 2015). In the other study (Mansouri et al., 2015), FP lesioned animals were compared against controls on the performance of a variant of the Wisconsin Card Sorting Test (WCST) – where subjects are required to respond by matching a sample to one of several test items according to uncued rules that vary dynamically across the session. When compared to controls, damage to FP had an effect on performance but did not impair rule maintenance or rule switching. Instead, enhancing effects on FP lesioned animals were seen as subjects were better able to maintain the relevant rule when intervening distractors (such as free reward and novel tasks between trials of the WCST task) were used. This has lead the authors to hypothesize that the key contribution of FP to cognition is in supporting the exploration and evaluation of the relative value of different alternatives, particularly when novel.

However, the only study to record from single-cells in frontal pole found surprisingly simple neuronal responses (Tsujimoto et al., 2010). In a task where the monkey was cued to repeat or switch the choice of the previous trial, the neurons encoded the response made only when the animal received feedback about the correctness of its choice. The authors considered that the representation of the selected response at the time when the outcome is revealed, reflects an important monitoring role in planning self-generated (i.e., not when experimentally guided but when the subject has to choose based on some memory, rule or stored representation) responses for future choices. Other authors rather emphasised this result as a forward implementation of internal models relevant for the task (Koechlin, 2011) and for model-based reinforcement learning.

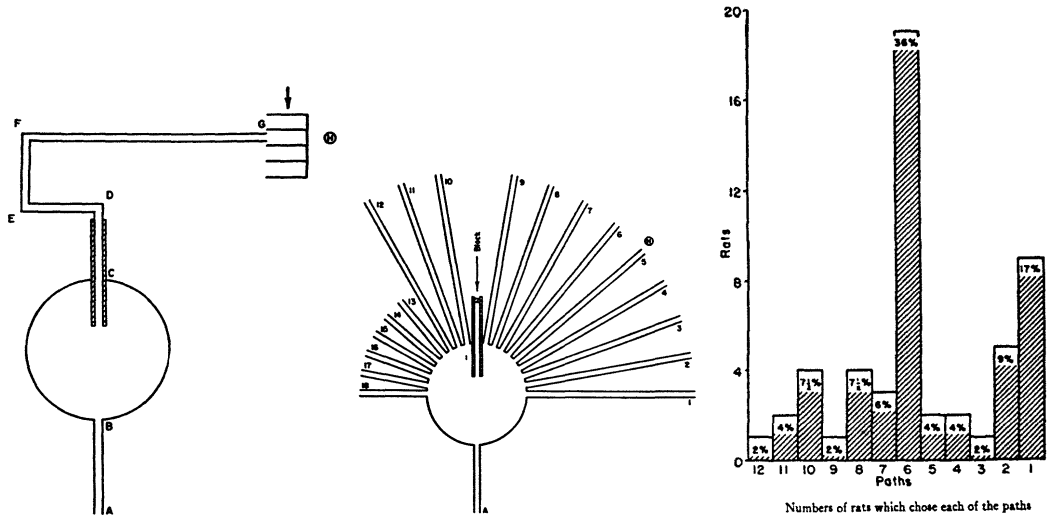
2.3 Animal learning psychology: habitual and goal-directed behaviour

Animals learn associations between their own actions and consequent changes in the environment. This form of acquired behaviour, known as instrumental conditioning, is different from classical conditioning as it allows them to exert control over their surroundings in order to satisfy their needs and desires. One famous psychologist – who rigorously studied this type of learning – was Edward Thorndike, who proposed the influential *Law of Effect*, where actions *closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that when it recurs, they will be more likely to recur*. This stimulus-response theory views rewards during a trial-and-error learning experience acting in a retroactive way, meaning that they increment the strength of the connection between a stimulus and a response that occurred before the reward (Grindley, 1932).

Another prominent animal learning psychologist was Edward Tolman, whose research challenged some stimulus-response ideas. In a spatial orientation experiment (Figure 2.8) (Tolman et al., 1946), rats were first trained in a maze composed of a walled alley leading to a route of some turns before they reached the reinforcement (Figure 2.8a). Then, the experimental apparatus was changed by blocking the walled alley previously leading to the reward and various radiating paths were added (Figure 2.8b). Importantly, this test was performed in extinction. What Tolman and colleagues observed was that the animals had a strong tendency to choose the radiating path appropriately pointing towards where the food used to be (Figure 2.8c).

These and other experimental results led him to bring forward the idea of a *cognitive-like map of the environment* guiding animal behaviour (Tolman, 1948). According to Tolman, stimuli are not processed by a *telephone switchboard* that simply connects them with a response. Instead, they are operated by a *central control room* which uses *intervening variables* (Tolman, 1938) and provides *routes and paths and environmental relationships* that determine animal's responses (Tolman, 1948). Motor hesitations and repetitive looking back and forth at choice points, known as *vicarious trial and error*, was good evidence that such mental processes may exist in animals (Muenzinger, 1938). This school of thought became highly influential and generated several concepts of cognitive psychology, including working memory (Baddeley, 1992).

Later experimental work by Dickinson and Balleine helped in defining an important fragmentation of instrumental behaviour (Dickinson and Balleine, 1994). They divided actions into habitual and goal-directed on the basis of their action-outcome contingencies as



(a) (b) (c)

Fig. 2.8 **Tolman's spatial orientation experiment.** a) Experimental apparatus used in the training phase. b) Modified apparatus used in testing. The reward is located at the end of path number 5. c) Results revealing the number of rats choosing each of the radiating paths.

well as motivational sensitivities. To test the action-outcome criteria they used experimental assays where the instrumental contingency is either altered (Bolles et al., 1980) or degraded (Hammond, 1980). In such paradigms, animals are trained under a high probability of action-outcome association which is then reduced, so that the likelihood of obtaining a reward is similar whether or not the action is performed. A reduction in response rate after the change reveals sensitivity to contingency degradation. On the other hand, outcome devaluation paradigms (Adams and Dickinson, 1981; Balleine and Dickinson, 1991, 1998a; Colwill and Rescorla, 1985; Dickinson, 1996) reveal sensitivity to changes in outcome's motivational value. In these experiments, animals are first trained to perform an action (e.g., pressing a lever) for a desired reward. Then, the incentive value of the outcome is changed by, for example, pairing the outcome with illness or by inducing outcome-specific satiety. Finally, the learned response is tested in extinction (i.e. with no reward), and the behaviour of animals for which the outcome has been devalued is compared with that of animals who have not undergone such devaluation procedure.

Habitual behaviour, according to Dickinson and Balleine, is neither aware of changes in the value of the consequences previously associated with the action nor sensitive to changes in the causal relationship between the action and its consequences. Empirically, a habit is an automatic action that arises from repeated practice and defined by its behaviour inflexibility. It revives Thorndike's law as it specifies responses not controlled by the current value of

the goal, but instead by previously stimulus-response reinforced associations (Dickinson, 1985). When a habit controls behaviour the choice becomes insensitive to motivational manipulations, as it happens in overtrained animals when faced with an outcome devaluation paradigm (Adams, 1982) or in certain circumstances such as chronic stress (Dias-Ferreira et al., 2009). Extended practice also leads to a reduced sensitivity to changes in instrumental contingency (Dickinson, 1998). Good examples of a maladaptive model-free or habitual system include superstitious behaviour (Skinner, 1948) and addiction (Dayan, 2009). Computationally, this stamping in of a reinforced choice from direct experience is analogous to the description of model-free control exposed in the previous chapter.

On the other extreme, a goal-directed action implies not only a representation of the causal relationship between the action and its consequences, but also a sensitivity to changes in the value of the goal or outcome with which the action is associated. Therefore, goal-directed behaviour meets not only the instrumental but also the goal criteria (Dickinson and Balleine, 1994). The awareness of such associative structures relates this definition with Tolman ideas of cognitive maps, at the same time as it also alludes to the learning endeavours of model-based control in estimating both task and reward structures. Moreover, these required storage and manipulation of goal-directed representations show similarities with the requirements for forward search in a decision tree of model-based computations. Although goal-directed behaviour could better suit animal needs, the challenges of maintaining action–outcome relationships are also psychologically acknowledged in the light of a *law of less work* (Hull, 1943), where actions more difficult to make are less likely to occur. In a way, this problem resembles the computational limitations of MB-RL when the decision tree is complex, as discussed in the previous chapter.

How the two forms of instrumental control interact and determine action choice is regarded as a very contemporary psychological issue (Menzel and Fischer, 2011). De Wit and colleagues (de Wit and Dickinson, 2009) argue that instrumental performance at any given stage of learning could be a sum of the goal-directed and habitual components and propose an integrative associative-cybernetic model that mimic the hybrid computational treatments proposed in chapter 1.

2.4 Neural substrates of model-free and model-based reinforcement learning

Two main approaches have been used to study the neurobiology of model-free and model-based learning systems. The psychological concepts of habitual and goal-directed behaviour have been explored in rodents and mostly making use of loss of function experiments, whereas the theoretical reinforcement learning framework have received more attention by researchers performing functional neuroimaging studies in humans.

The involvement of prefrontal-basal ganglia loops in model-free and model-based control has received a lot of attention from these two research communities (Balleine and Dickinson, 1998b; Balleine and O'Doherty, 2009; Bornstein and Daw, 2011; Daw and Doya, 2006; Dayan and Niv, 2008; Dolan and Dayan, 2013; Doll et al., 2012; Ito and Doya, 2011). However, research in non-human primates has rarely tested direct reinforcement learning propositions and not much effort has been made to integrate lesion work and single-neuron findings obtained from non-human primate studies. Finally, no functional evidence of interaction between the two learning systems has been described in non-human primates and more mechanistic details of such implementation are needed.

Although homologies between rodents and humans can be drawn regarding goal-directed (or MB-RL) and habitual (or MF-RL) (Balleine and O'Doherty, 2009), it is also essential to acknowledge the anatomical and functional differences between the two species. Most studies in rodents have been focused on the basal ganglia computations, an evolutionary conserved structure. Yet primates have more links between motor and associative striatum, as well as less segregation of the dopamine system (Joel and Weiner, 2000). Nevertheless, it is at the level of the prefrontal cortex that this problem assumes more relevance (Fig. 2.9). Clear homologues of primate prefrontal cortex, particularly the granular regions, may not exist in rodents leading some authors to even question its existence (Preuss, 1995; Wise, 2008).

In addition, monkey prefrontal circuits involved in decision making show strong structural and functional neuroimaging similarities in the distribution and connectivity when compared to human brains (Neubert et al., 2014). This is an important issue given the relationship between prefrontal cortex and goal-directed behaviour, together with the fact that model-based theories are at an extremely early stage of research, when compared with model-free ones (Daw and Dayan, 2014; Doll et al., 2012; Doya et al., 2002).

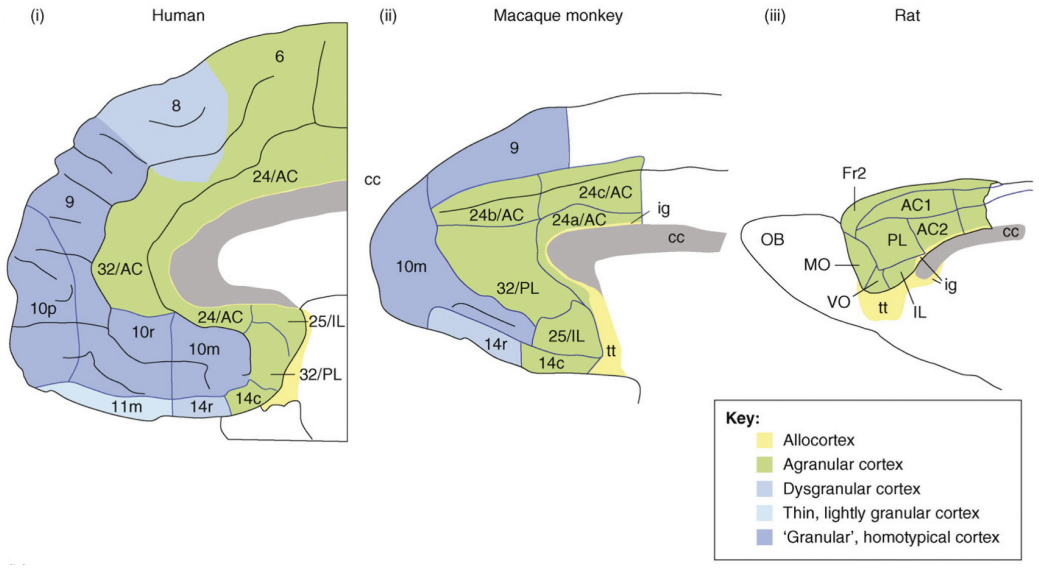


Fig. 2.9 **Medial frontal cortex across species.** A diagrammatic representation of the cytoarchitecture of human (i), macaque (ii) and rat (ii) prefrontal cortices. Abbreviations: AC, anterior cingular area; IL, infralimbic cortex; cc, corpus callosum; Fr2, second frontal area; MO, medial orbital area; PL, prelimbic cortex; VO, ventral orbital area; m, medial; r, rostral; c, caudal; p, posterior. From Wise (2008).

Following a summary of the current knowledge from rodents and human studies, an attempt will be made to integrate non-human primate prefrontal and basal ganglia experiments in model-free and model-based reinforcement learning.

Model-free or habitual neural substrates have been found in both cortical and subcortical brain structures of rodents (Figure 2.10). It is important to remember that in outcome devaluation tests, the goal-directed performance observed with limited training is lost to habitual control if this training gets extended. However, rats with lesions in dorsolateral striatum (posterior caudate and putamen in primates) remain goal-directed after outcome devaluation even after overtraining (Yin et al., 2004). Inactivation of this region also causes habitual performance to be sensitive to changes in the action–outcome contingency (Yin et al., 2006). The infralimbic cortex is a cortical structure involved in habits. Lesions in this region prevent the expression of habitual behaviour after both limited and extended training (Killcross and Coutureau, 2003). Moreover, inactivation of the infralimbic prefrontal cortex in rodents not only reinstates goal-directed behaviour in animals that have previously been habituated (Coutureau and Killcross, 2003), but also impinges on the formation of new habits (Smith et al., 2012). Another region that seems to be involved in habitual behaviour is the amygdala. Disruption of the connection between the anterior portion of the amygdala central nucleus and the dorsolateral striatum in rats, leads to a lack of habitual

responding after outcome devaluation (Lingawi and Balleine, 2012). Finally, dopamine also plays an important role in modulating model-free learning-related plasticity in these cortico-subcortical networks. Given the well-known role of phasic dopamine in reporting reward prediction errors in pavlovian learning (Schultz et al., 1997) and the fact that model-free computations uses similar prediction errors to update values, such involvement is not surprising. Indeed, attenuation of the phasic dopaminergic activity (Wang et al., 2011) as well as impairment of the nigrostriatal (Faure et al., 2005) or mesocortical (Barker et al., 2013; Hitchcott et al., 2007) dopaminergic pathways significantly impairs learning tasks related to habitual behaviour.

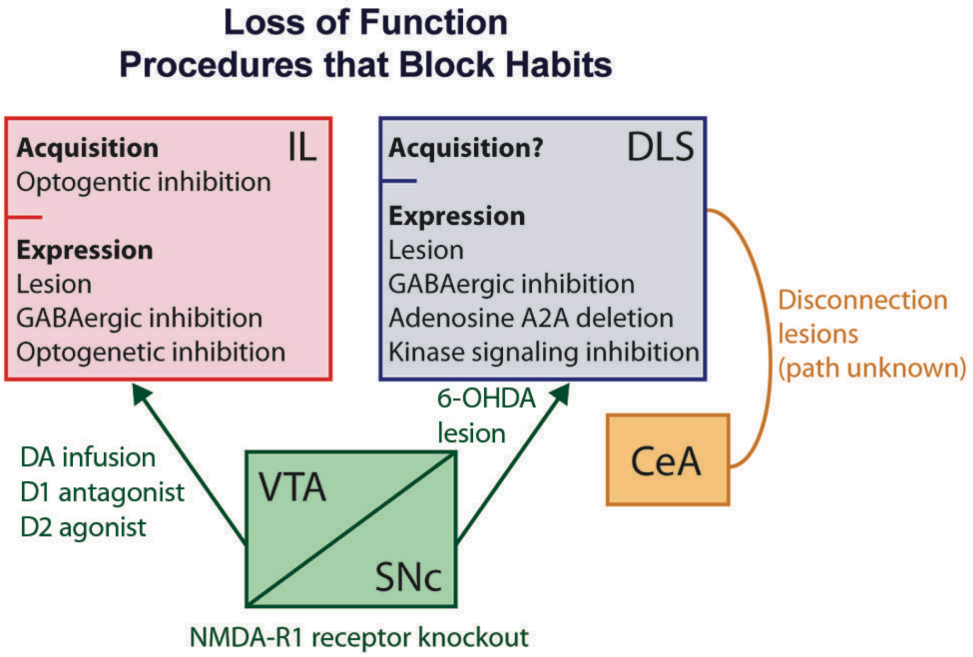


Fig. 2.10 **Neural substrates of habitual behaviour derived from loss of function experiments in rodents.** Anatomical, neuropharmacology and optogenetic manipulations (see text for further details) have all been used to show involvement of the infralimbic cortex (IL) and the dorsolateral striatum (DLS) in habitual or model-free mechanisms. They have also shown an important neuromodulatory role of dopamine (DA) cells in the ventral tegmental area (VTA) and in the substantia nigra pars compacta (SNc). Finally, the central nucleus of the amygdala (CeA) also seems to play a role through an interaction with the DLS, but details behind the real mechanism are less clear. From (Smith and Graybiel, 2014).

In humans, the ventral striatum has been the region where most consistently fMRI BOLD signal correlates with dopaminergic reward prediction errors and is often considered as part of the model-free system (McClure et al., 2003; O'Doherty et al., 2003; Seymour et al., 2004). However, more recent imaging studies have often found the VTA/SNc also

covarying with reward prediction errors (D'Ardenne et al., 2008). However, it is important to note that these studies used pavlovian learning scenarios and, therefore, have not tested model-free substrates as defined in instrumental behaviour. In fact, the entire ventral subcircuit of the prefrontal-subcortical network (including ventral striatum, orbitofrontal cortex and amygdala) is particularly involved in pavlovian learning (Daw and O'Doherty, 2014).

Despite some theoretical work attempting to explain how these pavlovian responses can compete or cooperate with instrumental behaviour (Dayan et al., 2006), the neural basis of these interactions are not clearly elucidated. Nevertheless, other human studies aimed to specifically test instrumental model-free substrates. When a classic outcome devaluation paradigm was applied to humans, the activity of posterior parts of putamen in the training phase increased significantly when the first sessions were compared to the final ones (Tricomi et al., 2009). In another study with planning and overtraining conditions, lateral parts of putamen also encoded values on extensively trained trials at the time of choice (Wunderlich et al., 2012a). Moreover, a shift in neuroimaging activity from anterior caudate to posterior putamen has equally been found as function of motor sequence learning (Jueptner et al., 1997; Lehéricy et al., 2005). Finally, the strength of connections between posterolateral striatum and premotor cortex is significantly correlated with the prevalence of habitual behaviour (de Wit et al., 2012b). All these findings suggest an important role of this region in model-free control of instrumental behaviour, in line with its homology with the dorsolateral striatum of rodents.

The neuroscience of model-based RL is less well explored. Rodents prelimbic cortex and dorsomedial striatum (equivalent to anterior caudate nucleus in primates) are the two main candidates. Lesions in prelimbic cortex interfere with sensitivity to contingency changes between response and a specific reward (Balleine and Dickinson, 1998b). This deficit could actually be a consequence of an inability by these animals to retain learning in working memory (Corbit and Balleine, 2003). Importantly, impairments observed only occur if this region is silenced during training, instead of at the time of testing. In contrast to a role in the expression of goal-directed behaviour, the prelimbic cortex seems to be genuinely involved in goal-directed learning (Ostlund and Balleine, 2005; Tran-Tu-Yen et al., 2009). The dorsomedial striatum receives dense projections from the prelimbic cortex and does appear to be critical for both the learning and the expression of goal-directed behaviour (Yin et al., 2005a,b). Another structure viewed as player of a model-based circuit is the orbitofrontal cortex. In rodents this region uses inference obtained by state transition knowledge to flexibly estimate expected value essential for model-based evaluation (Jones et al., 2012).

From neuroimaging evidence in humans, model-based evaluation has most often been associated with the ventromedial prefrontal cortex. This region encodes and tracks the expected reward of chosen actions (Daw et al., 2006; Gläscher et al., 2009; Tanaka et al., 2004). It does so by taking advantage of a model of the environment (Hampton et al., 2006), being sensitive to task contingencies (Valentin et al., 2007) and combining different sources of information (Behrens et al., 2008). Despite this complexity, it can also include values learned from a model-free learner in MB-RL reasoning (Wunderlich et al., 2012a). On the other hand, lateral prefrontal and intraparietal cortex showed state prediction error signals conforming to an update important for state transition learning (Gläscher et al., 2010). Another structure similarly implicated in a model representation has been the hippocampus, in agreement with its known spatial mapping properties (Bornstein and Daw, 2012). Hippocampal neural activity was found to replay previously experienced routes between trials or during sleep, which could help solving limitations of model-free computations (Johnson and Redish, 2005).

Very few studies addressed the issue of how the two learning systems interact with each other. One of the first studies and an inspiration for this project was the one by Daw et al. (2011) (Figure 2.11). Their design of a sequential decision task with a particular state transition function, was able to reveal concomitant behavioural and neuronal signatures of model-based and model-free learning. Contrary to the expectations, the study found that fMRI BOLD signal in ventral striatum did not reflect a *pure* model-free signal, but was rather explained by a mix of both learning strategies in a proportion that could also explain behaviour. This finding can, nevertheless, be explained by the strong prefrontal projections to this region (Haber and Knutson, 2009) and suggest model-based influences on model-free computations, at least in striatum. Recently, a study (Gershman et al., 2012) has found behavioural evidence supporting a cooperative interaction between the two systems, in agreement with the Dyna algorithm (Sutton, 1990). With a sophisticated multi-phase design investigating retrospective revaluation, they were able to show higher order contingency choices that would imply a model-based system training the model-free system offline. In line with this, some dependency of dopamine expectancy signals on orbitofrontal cortex has been described in rodents (Takahashi et al., 2011). The other alternative of interaction are some model-based algorithms that take in cached model-free quantities in order to reduce their computational demand (Doll et al., 2012; Wunderlich et al., 2012a). In fact, concurrent loading of executive resources with additional information attenuated contributions of model-based behaviour in favour of a more model-free control strategy (Otto et al., 2013).

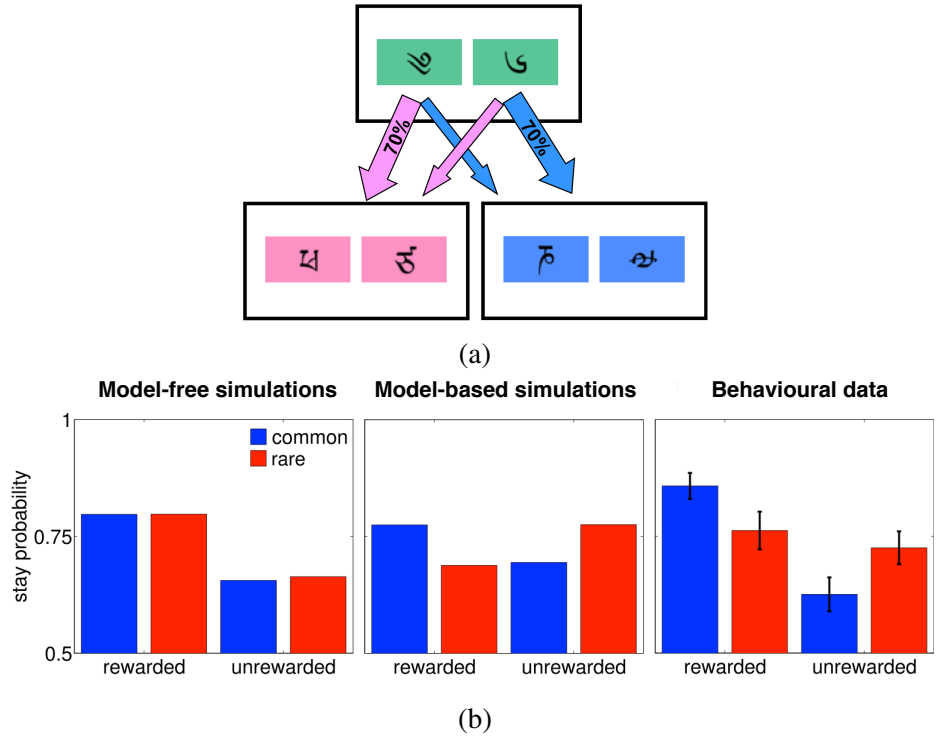


Fig. 2.11 **Behavioural influence of parallel model-free and model-based learning architecture.** (a) The two-stage Markov decision task designed by Daw et al. (2011) aimed to detect simultaneous signals of both model-free and model-based methods as they learn concurrently. The task required human subjects to make a first-stage choice between two stimuli (green background). Each of these first-stage choices was more often (70% of the time) associated with one of the two second-stage states (pink or blue backgrounds). This state transition structure was kept fixed throughout the experiment. In the second-stage, another two-option choice was required and reinforced. The probability of reward in the second-stage changed over time according to an independent random walk for each of the four second-stage stimuli. Participants were instructed to maximize their rewards. (b) The sequential aspect of this task helps detecting computational differences between both learning methods, particularly in terms of prediction patterns by which reward obtained in the second-stage should impact the probability of stay in first-stage choices. As mentioned before, model-free will reinforce any satisfactory action not taking into account the task structure. Therefore, simulation results with a model-free learner show a main effect of reward, regardless of whether the reward was obtained after a common or a rare transition. The model-based system will use the learned state transitions to evaluate actions. As result, the model-based simulations reveal a main effect of the interaction between reward and the transition type. Finally, the behavioural results obtained from the study participants show signs of both strategies.

One the other hand, following the ideas of parallel implementation of both model-free and model-based systems in the brain, a recent study (Lee et al., 2014) tested the theoretical proposal (Daw et al., 2005) of using each system's uncertainty on the value estimates to resolve the competition between the two systems. In fact, the study also uses a two-step decision task but with specific adjustments that allow independent manipulations of values and uncertainties in both model-free and model-based learning systems. Their proposed process of arbitration relies on whether the prediction error signals derived from each learning system were high or low. The model-free learner uses its reward prediction error signal, whereas the model-based system uses a state prediction error that is relevant for the learning of the transition function. A reliability signal is then computed for each learning strategy based on the variance-to-mean ratio of the probability that the respective error signal is zero. Thus, control by the model-based or model-free systems over behaviour is implemented according to the reliability level of each error signal. However, when the reliabilities are the same, and because model-based computations are demanding, model-free is favoured. Neural correlates for this arbitration process were found in the lateral inferior prefrontal cortex as well as in medial frontal pole cortex. Moreover, they also found a signal reflecting the difference in reliability between the model-based and model-free signals in a region of the cingulate cortex.

Chapter 3

Combined model-free and model-based reinforcement learning behaviour

3.1 Abstract

Animals can learn to influence their environment either by exploiting stimulus-response associations that have been productive in the past, or by predicting the likely worth of actions in the future based on their causal relationships with outcomes. These respectively model-free (MF) and model-based (MB) strategies are used to solve the reinforcement learning (RL) problem of interaction with the world to optimise future benefits. Computational constraints and a speed accuracy trade-off imply an advantage in combining both learning strategies, but this has only been demonstrated in humans. We trained rhesus monkeys to perform a two-stage decision task that was designed to elicit and discriminate the use of both MF-RL and MB-RL. A descriptive and logistic analysis of choice behaviour found that the structure of the task (of MB importance) and the reward history (of MF and MB importance) significantly influenced choice as well as response vigour. In addition, when we performed a trial-by-trial computational analysis on our data using different RL algorithms, we found that in the model that best fitted the data, choices were made according to a weighted combination of MF and MB action values (with a weight for MB-RL close to 90%). Generative modelling procedures prompted refinements to commonly used RL models, suggesting additional high-level processing in credit assignment. In conclusion, our data replicate, in non-human primates, results similar to those found in human subjects performing an equivalent decision-making task. These findings support the idea of more integrated views of combined MF and MB strategies in animal learning.

3.2 Introduction

In chapter 1 the theoretical framework of reinforcement learning theory (RL) for studying how agents interact with the environment to optimise future benefits was exposed. Furthermore, we also formally defined both model-free (MF) and model-based (MB) approaches to solve the RL problem. In brief, both methods rely on previous experience but they differ as to how this information is used to infer the values of choices in a sequential decision problem. MB-RL valuation integrates information about reward with knowledge about the state-transition function, which specifies how the state of the world evolves probabilistically given particular actions. Having such a model of the environment allows it to plan or simulate experience, reducing the need to take actions to estimate values, and making it readily adaptive in situations where the transitions or rewards change. However, the prospective nature of this approach is computationally demanding if several decisions are required, because choices are evaluated by searching along paths in an ever-expanding decision tree to calculate the cumulative expected rewards. By contrast, MF-RL is computationally simple and faster. It is blind to the state-transition function and it learns by bootstrapping sampled experience taking changes in expectations as signs of errors in its value predictions. With this approach, MF-RL typically requires more sampling from the world to achieve good performance and is therefore less sensitive to changes in goal values. Either because of limitations in the computational resources or due to a speed-accuracy trade-off, it seems advantageous for a learning agent to have both strategies and take advantage of each according to the required task. In chapter 2 the focus was the analogies to MF and MB-RL approaches that have been drawn with the psychological concepts of Thorndike's law of effect or habits versus Tolman's cognitive maps or goal-directed behaviour, respectively (Daw et al., 2005; Dickinson, 1985; Dickinson and Balleine, 1994; Dolan and Dayan, 2013; Thorndike, 1911; Tolman, 1948). In the same chapter, we reviewed the wealth of experimental evidence suggesting the existence in the brain of complex prefrontal-striatal circuits that may support each of these distinct learning systems in value-guided decision making (Balleine, 2005; Daw and Dayan, 2014; Dolan and Dayan, 2013; Doya et al., 2002). However, very few studies have focused on detecting simultaneous behavioural signatures of both learning strategies, and those have so far focused only on human subjects (Daw et al., 2011; Gershman et al., 2012; Otto et al., 2013; Wunderlich et al., 2012a).

Here, two rhesus monkeys were trained to perform a two-stage Markov decision task (Fig. 3.1) designed to induce trial-by-trial adjustments in choice that combine both MF and MB learning control. By performing a quantitative behavioural analysis with computational

and generative RL modelling, we found that non-human primates solve this task through the use of reward history (of MF and MB importance) as well as information about the state-transition structure (of MB relevance). Furthermore, both strategies also influenced the alacrity of responding, in agreement with the speed-accuracy trade-off associated with their computations. Overall, the following results support modern views that optimal learning may require parallel integration of both MF and MB-RL computations (Balleine and O'Doherty, 2009; Daw and Dayan, 2014; Daw et al., 2005).

3.3 Experimental procedures

Subjects and experimental apparatus

Two male rhesus monkeys (*Macaca mullata*) were used as subjects: subject C was 8 Kg; and subject J was 11 Kg. Daily fluid intake was regulated to maintain motivation on the task. During the experiment, subjects were seated in a primate chair inside a darkened room with their heads fixed and facing a 19-inch computer screen (60Hz video refresh rate) positioned 62 cm from the subject' eyes. Each subject's eye position and pupil dilation was monitored with an infrared eye tracking system having a sampling rate of 240 Hz (ISCAN ETL-200). Both subjects indicated their choice by moving a joystick with a left arm movement towards one of three possible locations (C: left, right and down; J: left, right and up). The reward (C: cranberry juice diluted to one-fourth with water; J: apple juice diluted to one half with water) was provided by a spout positioned in front of the subject's mouth and delivered at a constant flow-rate using a peristaltic pump (Ismatec IPC). We used Monkeylogic software (http://www.monkeylogic.net/) to control the presentation of stimuli and task contingencies, to generate timestamps of behaviourally-relevant events, and to acquire joystick as well as eye data (1000 Hz of analog data acquisition). All visual stimuli used were the same across sessions for both subjects, and were presented at pre-determined degrees of visual angle. The six stimuli were modified to be of equal size and luminance using a custom-made image processing algorithm. Similarly, the background colours used (grey, violet and brown) were luminance adjusted for equality and verified with a luminance meter. Finally, three pictures used as secondary reinforcers were generated as different spatial combinations of the same number of black pixels in a white background, also to assure luminance equality. All experimental procedures were approved by the UK Home Office and were in compliance with the guidelines of the European Community for the care and use of laboratory animals (EUVD, European Union directive 2010/63/EEC).

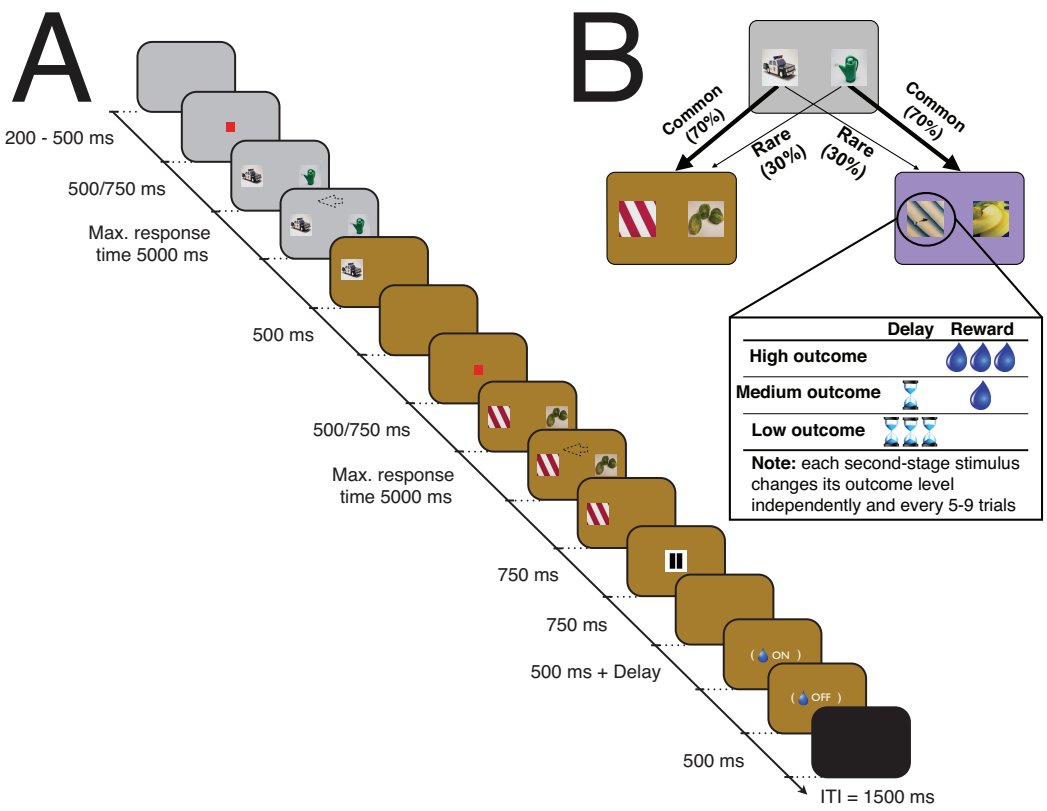


Fig. 3.1 **Two-stage decision task.** (**A**) Timeline of events. Eye fixation was required while a red fixation cue was shown, otherwise subjects could saccade freely and indicate their decision (arrow as an example) with a manual joystick movement. Once the second-stage choice had been made, the nature of the outcome was revealed by a secondary reinforcer cue (here, the pause symbol represents high outcome). Once the latter cue was off the screen, there was a fixed 500 ms delay and the possibility of a further delay (for both medium and low outcomes) before juice was provided (for both high and medium outcomes). The inter-trial interval (ITI) was 1.5 s. (**B**) The state-transition matrix (kept fixed throughout the experiment). Each second-stage stimuli had an independent reward structure (with outcomes being defined by the magnitude of the reward and the delay to its delivery) according to a form of random walk sampled afresh on each session. Task design influenced by Daw et al. (2011).

Task: design and timeline

Subjects performed a two-stage Markov decision task (see Fig. 3.1), similar to the one used in a previous human study (Daw et al., 2011) that was designed to detect simultaneous signatures of MF and MB systems as they concurrently learn. In brief, two decisions had to be made before the subject received a reward. The first-stage state was represented by a grey background and the choice was between two options presented as pictures (the same fixed set of pictures was used throughout the entire task). Each of these first-stage choices could lead to either a common (70% transition probability) or rare (30% transition probability) second-stage state, represented by different background colours (brown and violet). This state-transition structure was kept fixed throughout the experiment. In the second-stage, another two-option choice between pictures was required and it was reinforced according to different levels of reward. Importantly, to encourage learning, each of the four second-stage options had independent reward structures according to a form of random walk (see below) that was sampled afresh on each session. In both decision stages, the choice options (or presented stimuli) were randomized to one of three possible locations. Fifteen percent of the trials were forced (i.e., without allowing a choice as only one option was presented), which could be at either the first or second-stage. Unless stated otherwise, such forced trials were not included in the data analysis. The trial type sequence was randomly generated at the start of the session and was followed even after error trials. Errors could take the form of trials with no choice, no eye fixation, eye fixation break, early joystick response, joystick not centred before choice, or movement towards a location not available; these error trials were all followed by a time-out of 7000 ms. Unless otherwise specified, we excluded error trials from the data analysis (C: $M = 5\%$; J: $M = 8\%$).

The value of a choice option could assume one of three categorical outcome levels, defined according to the amount of juice delivered (determined by the time the juice pump was on) and a specific delay (in addition to a fixed 500ms delay common to all outcome levels) before juice delivery. Therefore, the outcome could be: high (big reward and no delay), medium (small reward and small delay) or low (no reward and big delay). The precise reward amounts for big and small rewards were tailored for each subject to ensure that they received their daily fluid allotment over the course of the experimental sessions. Consequently, the duration for which the reward pump was active (and hence the magnitude of delivered rewards) differed slightly between the two subjects. Furthermore, instead of a fixed reward amount, big and small rewards corresponded to non-overlapping time intervals (C: high reward ranged on average from 682 to 962 ms and medium reward ranged on average from 117 to 390 ms; J: high reward ranged on average from 976 to 1257 ms and

medium reward level ranged on average from 507 to 826 ms) of juice delivery where a small Gaussian drift (mean/standard deviation of 0/200 ms for high reward and 0/100 ms for medium reward) was added. The variance in the reward amount within an outcome level was used not only to promote constant valuation of the outcome value, but also to help the computational model fitting procedure. The additional specific delay periods were fixed throughout the experiment but varied across subjects (C: 750 ms for small delay and 2500 ms for big delay; J: 1500 ms for small delay and 4000 ms for big delay).

Importantly, for each of the second-stage pictures the outcome level remained the same for a minimum number of trials (a uniformly distributed pseudorandom integer between 5 and 9) and then, either stayed in the same level (with one-third probability) or changed randomly to one of the other two possible outcome levels. Three different abstract stimuli were used as secondary reinforcers, providing feedback for each of the three outcome levels. We adopted this strategy so that feedback-related neuronal activity could be analysed in a fixed duration epoch independent of licking movements, and also because of the different delays employed across outcome levels and during reward delivery. Both subjects had prior classical conditioning training with these stimuli, with the above mentioned reward magnitude ranges and delays for each outcome level used in the experiment being respected.

The sequence of events in the behavioural task are shown in Fig. 3.1A. Each trial started with the presentation of a grey background (start epoch). A central square fixation cue 0.4° in width then appeared after a random interval of 200-500ms. After this, subjects were required to keep the joystick in the centre position as well as maintain eye fixation within 3.4° (C) or 2.8° (J) of the fixation cue for a 500ms (C) or 750ms (J) period (fixation epoch). Then, the fixation cue was removed and two pictures (5° in size) appeared at 7° in two of the four available locations (choice epoch). During the task, in the absence of a fixation cue, the animal was free to look around. The maximum time allowed for acquiring eye fixation as well as making a response with the joystick was 5000ms for both choice stages. After a choice was made, the non-selected image was removed and the background color changed according to the second-stage state to which the transition had occurred (transition epoch). The image selected in the first-stage remained on the screen for 500ms before it was removed. Similar fixation and choice epochs were used for the second-stage. Once the choice had been made in the second-stage, the non-selected cue was removed and the selected remained on the screen for 750 ms before the secondary reinforcer stimulus (5° square) appeared at the center of the screen for 750ms (pre-feedback epoch). After the secondary reinforcer stimulus was removed, a fixed 500ms delay period occurred before either the reward delivery (for high outcome) or both small and big additional delays started (for

both medium and low outcomes, respectively). Therefore, a total of 1250ms was the minimum time from the secondary reinforcer presentation to the delivery of any juice (feedback epoch). The inter-trial interval period duration was 1500ms (ITI epoch).

Full behavioural training took six to nine months for each animal (subject J was the first being trained and required more piloting and refinement before the final version of the task; subject C was faster in initial phases of learning but took longer in the last two steps of the protocol) and a similar protocol was applied for both. After acquaintance with basic skills (adaptation to the experiment room, presentation of visual stimuli and joystick movements), a multi-step task-specific training protocol was employed. It consisted of: 1) one-stage choice trials; 2) two-stage choice trials with fixed reward structures; 3) transition training without choice (very brief); 4) two-stage choice trials with transitions but no random walk reward structure; 5) final task without secondary reinforcer; 6) secondary reinforcer classical conditioning; 7) the full version of the task.

Choice behaviour and reaction time analysis

All statistical and computational modelling analyses were conducted using MATLAB® version R2014b (MathWorks). Unless otherwise stated, statistical significance was set at $\alpha = 0.05$. In addition to central tendencies and dispersion measures of the relevant behavioural variables (mean *M*, median *Mdn*, standard deviation *SD*, and standard error of the mean *SEM*), one sample or two-sample t-tests (Hedge's *g* effect size bootstrapped) and one-way ANOVA (η^2 effect size bootstrapped) were used. Whenever multiple comparison tests were required, a Bonferroni correction was applied.

Relevant behavioural variables were defined as: C is first-stage choice (1=car picture or picture A, 0=watering can picture or picture B); R is outcome level (assumed as continuous, with low=1, medium=2, high=3); and T is transition (rare=1, common=0). These variables, when used as predictors in regression analysis, were mean centred, and continuous variables were also scaled by dividing them by twice their standard deviations so that the magnitudes of regression coefficients could be directly compared (Gelman, 2008). Such adjustments in the variables were performed before the computation of the interaction terms.

In order to quantify the various factors predicting first-stage choice (i.e., stimulus A or stimulus B) on the current trial t, a first multiple logistic regression analysis was given by:

$$
\begin{aligned}
log\left[\frac{p(C_t=1)}{p(C_t=0)}\right] = \beta_0 + \beta_{C_{t-1}}C_{t-1} + \beta_{R_{t-1}}R_{t-1} + \beta_{T_{t-1}}T_{t-1} + \beta_{R_{t-1}\times T_{t-1}}R_{t-1}\times T_{t-1} + \\
\beta_{R_{t-1}\times C_{t-1}}R_{t-1}\times C_{t-1} + \beta_{T_{t-1}\times C_{t-1}}T_{t-1}\times C_{t-1} + \\
\beta_{R_{t-1}\times T_{t-1}\times C_{t-1}}R_{t-1}\times T_{t-1}\times C_{t-1}
\end{aligned}
\tag{3.1}
$$

where β corresponds to the estimated regression coefficient for each respective predictor. The β_0 is the constant term. The regression coefficient $\beta_{C_{t-1}}$ modelled a potential independent tendency to stick with the same option from trial to trial (preseveration effect); $\beta_{R_{t-1}}$, $\beta_{T_{t-1}}$ and $\beta_{R_{t-1}\times T_{t-1}}$ measured any potential preference in first-stage picture choice given the previous outcome level, the previous transition and the interaction effect of both, respectively; $\beta_{R_{t-1}\times C_{t-1}}$, $\beta_{T_{t-1}\times C_{t-1}}$ and $\beta_{R_{t-1}\times T_{t-1}\times C_{t-1}}$ were the regression coefficients of interest which quantified the main effect of reward (i.e., the effect of choosing the same first-stage choice given the previous reward), the main effect of transition (i.e., the effect of choosing the same first-stage choice given the previous transition) and the reward $\times$ transition interaction effect (i.e., the effect of choosing the same first-stage choice given the previous reward as well as previous transition) on first-stage choice, respectively.

To evaluate how reward and transition history from more than just the previous trial influenced first-stage choice behaviour, a second logistic regression model was created using the same predictor variables as in equation 3.1 but including information from up to five previous trials:

$$
\begin{aligned}
log\left[\frac{p(C_t=1)}{p(C_t=0)}\right] = \beta_0 + \sum_{i=1}^{5}\Big[\beta_{C_{t-i}}C_{t-i} + \beta_{R_{t-i}}R_{t-i} + \beta_{T_{t-i}}T_{t-i} + \\
\beta_{R_{t-i}\times T_{t-i}}R_{t-i}\times T_{t-i} + \beta_{R_{t-i}\times C_{t-i}}R_{t-i}\times C_{t-i} + \\
\beta_{T_{t-i}\times C_{t-i}}T_{t-i}\times C_{t-i} + \beta_{R_{t-i}\times T_{t-i}\times C_{t-i}}R_{t-i}\times T_{t-i}\times C_{t-i}\Big]
\end{aligned}
\tag{3.2}
$$

Reaction times were defined as the intervals from the moments when the first or second-stage options were presented until the conclusions of the joystick movements towards the specified location (all side locations with the same radius from the centre). For each subject and session, first and second-stage reaction times for each of the three possible side responses were independently log transformed and z-scored. This standardization was mandated by variations in reaction times between side locations as well as biases found in behaviour. Data points greater than three times the *SD*s from the individual means were

removed from regression analysis and also did not contributed to the z-scoring.

To determine the effect of behavioural variables on reaction time at first-stage choice, a multiple linear regression analysis was implemented:

$$RT_t = \gamma_0 + \gamma_{F_t} F_t + \gamma_{R_{t-1}} R_{t-1} + \gamma_{T_{t-1}} T_{t-1} + \gamma_{R_{t-1} \times T_{t-1}} R_{t-1} \times T_{t-1} + \varepsilon_t \tag{3.3}$$

where γ corresponds to the estimated regression coefficient for each respective predictor and ε the residual value. The RT_t is the z-scored log-transformed first-stage reaction time for each trial t. The γ_0 is the constant term and the predictor variable F_t was used to model (linearly-increasing) fatigue by counting the trials in the session. The $\gamma_{R_{t-1}}$, $\gamma_{T_{t-1}}$ and $\gamma_{R_{t-1} \times T_{t-1}}$ were the regression coefficients of interest as they quantify the main effect of reward (i.e., the effect on latency of first-stage response given the previous reward), the main effect of transition (i.e., the effect on latency of first-stage response given the previous transition) and the reward $\times$ transition interaction effect (i.e., the effect on latency of first-stage response given the previous reward as well as previous transition) on first-stage reaction time, respectively.

As for first-stage choice behaviour, the influence on RT of more than just the last trial's variables was also investigated with another linear regression model:

$$RT_t = \gamma_0 + \gamma_{F_t} F_t + \sum_{i=1}^{5} \left[\gamma_{R_{t-i}} R_{t-i} + \gamma_{T_{t-i}} T_{t-i} + \gamma_{R_{t-i} \times T_{t-i}} R_{t-i} \times T_{t-i} \right] + \varepsilon_t \tag{3.4}$$

The time of first-stage eye fixation was defined as the time from the moment when the first-stage central fixation cue was presented until the time when the subject's x and y eye position was within the required fixation radius. The raw data was then log transformed and z-scores were calculated. The regression analysis used the same predictors as described in equation 3.4.

Fixed-effects (fitting the regression models individually to each session) and mixed-effects (assuming regression coefficients to be random effects across sessions) analyses were performed for each subject. Fixed-effects fitting was performed using a generalized linear model regression package (`glmfit` in MATLAB with: a binomial distribution and the logit link function for logistic regressions, a normal distribution and the identity link function for linear regressions), and the statistical importance of each predictor's estimates was assessed by both the p-values obtained from each session as well as their distribution across

sessions (two-tailed one-sample t-test for a mean of 0 and unknown variance). Mixed-effects fitting was achieved with either a non-linear model with a stochastic approximation expectation-maximization method for logistic regression (`nlmefitsa` in MATLAB with importance sampling for approximating the loglikelihood) or a linear model method for the RTs (`filme` in MATLAB). The standard errors for the coefficient estimates as well as their 95% confidence intervals (CI) were reported. For all analyses, linear hypothesis testing on the vector of estimated regression coefficients (performed for each individual session in the fixed-effects, and using the estimated random effects for each predictor) was performed to test either if more than one coefficient or a difference between coefficients was significantly different from zero (`linhyptest` in MATLAB), with F statistic and p-values being reported. As noted, forced first-stage choice trials were excluded from regression fits.

Computational modelling

We fitted choice behaviour in the task in a similar manner to previous human studies (Huys et al., 2011) and assessing three different reinforcement learning approaches: MF learning, MB learning and a hybrid strategy combining the decision values of both (Daw et al., 2011; Gläscher et al., 2010). The task consists of three states (first stage: s_A; second stage: s_B and s_C), each with two actions (a_X and a_Y). Importantly, the action corresponds to the choice of a picture belonging to the respective state; thus when we refer to "action value", we are referring to the value assigned to the stimulus of the chosen action. The main goal is to learn a state-action value function $Q(s, a)$ mapping each state-action pair to its expected future value. On trial t, the first-stage state (always s_A) is denoted by $s_{1,t}$, the second-stage state by $s_{2,t}$, the first and second-stage actions by $a_{1,t}$ and $a_{2,t}$ and the first and second-stage rewards as $r_{1,t}$ (always zero) and $r_{2,t}$. For the model fitting $r_{2,t}$ corresponded to the amount of juice delivered at trial t divided by the maximum amount of juice obtained by the subject within the entire respective session.

In MF-RL the value for the visited state-action pair at each stage i and trial t, $Q(s_{i,t}, a_{i,t})$, is updated based on the difference between predictions at successive states, $\delta_{i,t}$, also known as the reward prediction error. For the first-stage choice, $r_{1,t} = 0$ and $\delta_{1,t}$ is driven by the second-stage value $Q(s_{2,t}, a_{2,t})$. On the other hand, at second-stage there is no further value apart from the immediate reward, $r_{2,t}$, since the following state ($s_{3,t}$) is an absorbing state that corresponds to the end of the current trial and the start of a new one ($Q(s_{3,t}, a_{3,t}) = 0$). Two different MF-RL algorithms were used to fit behaviour: the $SARSA$ variant of temporal difference learning (Rummery and Niranjan, 1994), which has previously been observed in

non-human primates (Morris et al., 2006); and the Q-learning algorithm, as described in rodents (Roesch et al., 2007).

In $SARSA$, the reward prediction error computation takes into account the state-action value of the actual action being selected:

$$\delta_{i,t} = r_{i,t} + Q(s_{i+1,t}, a_{i+1,t}) - Q(s_{i,t}, a_{i,t}) \tag{3.5}$$

By contrast, the Q-learning prediction error is based on the estimated best next action in the sequence, independent of the policy being followed:

$$\delta_{i,t} = r_{i,t} + \max_{a \in \{a_A, a_B\}} Q(s_{i+1,t}, a) - Q(s_{i,t}, a_{i,t}) \tag{3.6}$$

Either of these errors in the estimate drives learning by correcting the respective MF prediction through the following update rule:

$$Q_{MF}(s_{i,t}, a_{i,t}) \leftarrow Q_{MF}(s_{i,t}, a_{i,t}) + \alpha_i \delta_{i,t} \tag{3.7}$$

where α_i is the learning rate at stage i, and can be fit to the observed behaviour. Different learning rates for first-stage (α_1) and second-stage (α_2) were allowed based on previous work (Daw et al., 2011). Given the two-stage design of the task, the model also permits an additional stage-skipping update of first-stage values by having an eligibility trace (Sutton and Barto, 1998), λ parameter, which connects the two stages and allows the reward prediction error at the second-stage to influence first-stage values:

$$Q_{MF}(s_{1,t}, a_{1,t}) \leftarrow Q_{MF}(s_{i,t}, a_{i,t}) + \alpha_1 \lambda \delta_{2,t} \tag{3.8}$$

The parameter λ is also fit to the observed behaviour. Consistent with the episodic structure of the task (with an explicit inter-trial epoch), it is assumed that eligibility does not carry over from trial to trial.

In MB reinforcement learning, the agent not only maps state-action pairs to a probability distribution over the subsequent state but it also learns the immediate reward values for each state. More specifically, it requires knowledge of the probabilities with which each first-stage action leads to each second-stage state, as well as learning the expected outcome associated with each second-stage actions. The MB second-stage state-action values $Q_{MB}(s_{2,t}, a_t)$ are just estimates of the immediate reward $r_{2,t}$. For this reason both MF and MB approaches coincide at the second-stage, and we define $Q_{MB} = Q_{MF}$ at those states. On the other hand, the first-stage action values $Q_{MB}(s_{1,t}, a_{1,t})$ differ and are computed by weighting

their outcomes by the appropriate probabilities:

$$Q_{MB}(s_{A,t}, a_{1,t}) = P(s_{B,t}|s_{1,t}, a_{1,t}) \max_{a \in \{a_A, a_B\}} Q_{MB}(s_{B,t}, a) + P(s_{C,t}|s_{1,t}, a_{1,t}) \max_{a \in \{a_A, a_B\}} Q_{MB}(s_{C,t}, a) \tag{3.9}$$

Various approaches to solve the state-transition probability distribution gave rise to three different MB algorithms, designated here as $Forward_1$, $Forward_2$ and $Forward_3$. In the first model, we assume the agent had explicit knowledge of the correct state-transition probabilities, $P = \{0.3, 0.7\}$. Given the extensive training of both subjects prior to this experiment, it is possible that subjects had explicit knowledge that matched the actual state-transition probabilities. The second model, assumed agents learned to map action-state pairs a_1, s_2 to transition probabilities, $P = \{0.3, 0.7\}$, by counting whether they had more often encountered transitions $a_1 = 1, s_B$ and $a_1 = 2, s_B$ or transitions $a_1 = 1, s_C$ and $a_1 = 2, s_B$ and concluding that the more frequent category corresponds to $p = 0.7$. This latter model corresponds to the one used in the modelling of the original two-step task study (Daw et al., 2011). Finally, in the $Forward_3$ model the agent incrementally learns the transition structure by performing a hypothesis test between $p = \{0.3, 0.7\}$ versus $p = \{0.5, 0.5\}$ with an additional parameter (ζ) modelling the weight given to each of these models.

Finally, a so-called *Hybrid* model assumes that first-stage choices are computed as a weighted sum of the state-action values from MF and MB learning systems:

$$Q_{HYB}(s_{1,t}, a_{1,t}) = (1 - \omega)Q_{MF}(s_{1,t}, a_{1,t}) + \omega Q_{MB}(s_{1,t}, a_{1,t}) \tag{3.10}$$

where ω is a weighting parameter that determines the relative contribution of MB and MF values. When $\omega = 0$ the model reflects pure MF control; when $\omega = 1$, it reflects pure MB control. For convenience the hybrid model was constructed using the best fitting MF ($SARSA$ algorithm) and MB ($Forward_1$) algorithms, given the computational burden of fitting all possible combinations simultaneously.

A careful examination of the data revealed that the original hybrid model required further refinement in order to reproduce more accurately the strong influence of the previous trial on the present one. In this new *Hybrid+* model, the value of the chosen ($a_{1,t}$) or unchosen ($a \neq a_{1,t}$) first-stage action was boosted or suppressed as a function of whether the state-transition ($Trans$) observed at trial t was common or rare and the level of the outcome achieved. Algorithmically, after the previously described Q_{HYB} value update step (Eq. 3.10) an additional boost (or decrease) occurred according to:

$$Q_{HYB}(s_{1,t}, a_{1,t}) \leftarrow \begin{cases} Q_{HYB}(s_{1,t}, a_{1,t}) + L_1, & \text{if } Trans_t = \text{common}, Outcome_t = \text{high} \\ Q_{HYB}(s_{1,t}, a_{1,t}) + L_2, & \text{if } Trans_t = \text{common}, Outcome_t = \text{medium} \\ Q_{HYB}(s_{1,t}, a_{1,t}) + L_3, & \text{if } Trans_t = \text{common}, Outcome_t = \text{low} \end{cases}$$

and

$$Q_{HYB}(s_{1,t}, a \neq a_{1,t}) \leftarrow \begin{cases} Q_{HYB}(s_{1,t}, a \neq a_{1,t}) + L_1, & \text{if } Trans_t = \text{rare}, Outcome_t = \text{high} \\ Q_{HYB}(s_{1,t}, a \neq a_{1,t}) + L_2, & \text{if } Trans_t = \text{rare}, Outcome_t = \text{medium} \\ Q_{HYB}(s_{1,t}, a \neq a_{1,t}) + L_3, & \text{if } Trans_t = \text{rare}, Outcome_t = \text{low} \end{cases}$$

where there are separate parameters L_j for each outcome level which can be positive or negative, expressing support or opposition for that particular outcome level. This extra factor can be seen as a MF implementation of a MB effect (Akam et al., 2015) – MF, since it depends on an effect of the past trial rather than an assessment of a future one; MB, since it includes a one-step version of the interaction to which MB reasoning leads.

For any of the above reinforcement learning strategies, actions were assumed to be stochastic and chosen for each stage according to action probabilities determined by the respective Q-action values:

$$P(a_{i,t} = a | s_{i,t}) = \frac{exp(\beta_i[Q(s_{i,t}, a) + \kappa_i \times rep(a)]}{\sum_{a'} exp(\beta[Q(s_{i,t}, a) + \kappa_i \times rep(a')])} \tag{3.11}$$

where β_i is the inverse temperature parameter (distinct inverse temperatures are allowed for each stage) controlling the determinism of the choices, and so capturing noise and exploration (for $\beta_i = 0$ choices are fully random and for $\beta_i = inf$, choices are fully deterministic in the sense that higher-valued options are always preferred). $rep(a)$ is an indicator variable coding whether the current choice is the same as the one chosen on the previous visit to the same state, with κ_i being a further parameter that captures choice perseverance ($\kappa_i > 0$) or switching ($\kappa_i < 0$) (Lau and Glimcher, 2005), again with the possibility of distinct values for first (κ_1) and second-stage (κ_2) choices.

In the most general form, the conventional *Hybrid* algorithm involved a total of eight free parameters ($\theta = \{\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_1, \kappa_2, \lambda, \omega\}$), nesting pure MB ($\omega = 1$, with arbitrary α_1 and λ) and MF ($\omega = 0$) learning as special cases. The *Hybrid+* algorithm involved three additional parameters L_1, L_2, L_3. We also generated several simpler variants of these models

by allowing $\alpha_1 = \alpha_2, \beta_1 = \beta_2, \kappa_1 = \kappa_2, \kappa_1 = 0, \kappa_2 = 0$ and $\lambda = 0$. All parameters were fixed within a session, but could vary across sessions.

Model fitting procedures

Two forms of log-likelihood maximisation were used to fit the computational models to each subject's choice behaviour and estimate their free parameters. The first approach was a so-called fixed-effects analysis, maximizing the likelihood with respect to the parameters separately for each session. The second approach was a mixed-effects analysis, assuming, for each subject, parameters to be random effects across sessions. This implied maximizing the likelihood with respect to a characterization of empirical priors over the parameters (based on Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ for the vector of parameters $\boldsymbol{h}$; enforcing constraints on the parameters: $0 < \alpha_i < 1$; $\beta_i > 0$; $0 < \lambda < 1$; and $0 < \omega < 1$ by transforming samples from the Gaussian distributions using log and sigmoid transforms). In this scheme, one calculates approximate posterior distributions over the parameters for each session by combining these priors with the likelihoods. The effect of the prior is to regularize and stabilize estimates, particularly when the parameters are not well constrained by the data in particular sessions. The mixed-effects procedures used are identical to those described by Huys et al. (2011), but for completeness are detailed here.

The hyperparameters of the prior distribution $\boldsymbol{\theta}$, which consist of a prior mean $\boldsymbol{\mu}$ and a prior standard deviation $\boldsymbol{\sigma}$, were set to the maximum likelihood estimates (ML) for all N sessions, using empirical Bayes:

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{ML} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, P(\mathcal{A}|\boldsymbol{\theta}) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left(\prod_{i=1}^{N} \int d^N \boldsymbol{h}_i P(\boldsymbol{A}_i|\boldsymbol{h}_i) P(\boldsymbol{h}_i|\boldsymbol{\theta}) \right) \end{aligned} \tag{3.12}$$

where $\mathcal{A} = \{\boldsymbol{A}_i\}_{i=1}^{N}$ comprised all the actions (including first-stage and second-stage) by all the N sessions. For first and second-stage actions taken in all the T trials of the session it was assumed that $P(\boldsymbol{A}_i|\boldsymbol{h}_i) = \prod_{i=1}^{T} P(\boldsymbol{A}_{i,t}|\boldsymbol{h}_i)$.

The above maximization was achieved by Expectation–Maximization (EM) (Dempster et al., 1977). At the k^{th} iteration of the E-step of the algorithm, a Laplacian approximation to the individual posterior distributions of model parameters was used and the maximum a

posteriori estimate $\boldsymbol{m}_i$ of the parameters for each session i was found:

$$P(\boldsymbol{h}|\boldsymbol{A}_i) \approx \mathcal{N}(\boldsymbol{m}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)}) \tag{3.13}$$

$$\boldsymbol{m}_i^{(k)} = \underset{\boldsymbol{h}}{\operatorname{argmax}}\, P(A_i|\boldsymbol{h})P(\boldsymbol{h}|\boldsymbol{\theta}^{(k-1)}) \tag{3.14}$$

where $\mathcal{N}(\boldsymbol{m}_i^{(k)}, \boldsymbol{\Sigma}_i^{(k)})$ denotes a normal distribution over $\boldsymbol{h}$ with mean $\boldsymbol{m}_i^{(k)}$ and standard deviation $\boldsymbol{\Sigma}_i^{(k)}$ derived from the diagonals of the inverse Hessian matrix of the posterior at its maximum $\boldsymbol{m}^{(k)}$. In order to increase the chance of finding a good maximum a posteriori value, the largest value out of 101 separate optimizations was used, one starting from the best value on the previous iteration (or the output of the fixed effects analysis for the first iteration), and 100 more using random starting points. Optimization was performed using MATLAB 's parallel processing toolbox and `fminunc` function.

In the M-step, the hyperparameters $\boldsymbol{\theta}$ were estimated by setting the prior distribution mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ to:

$$\boldsymbol{\mu}^{(k)} = \frac{1}{N}\sum_i \boldsymbol{m}_i^{(k)} \tag{3.15}$$

$$\boldsymbol{\sigma}^{(k)} = \frac{1}{N}\sum_i \left[(\boldsymbol{m}_i^{(k)})^2 + \boldsymbol{\Sigma}_i^{(k)}\right] - (\boldsymbol{\mu}^{(k)})^2 \tag{3.16}$$

To help convergence, the algorithm was initialised with a prior mean and variance that corresponded to the 25% trimmed mean and variance of the parameters initially obtained with the maximum likelihood fixed-effects fit. The E and M steps were then repeated until the changes in the estimates between two E-steps were 0.005, signifying convergence. Once the subject's maximum likelihood prior parameters had converged, a final E-step was performed to determine the maximum a posteriori parameters for each session. All model fitting procedures were verified on surrogate generated data.

Model comparison and validation procedures

We fit the two pure MF ($\omega = 0$), the three pure MB ($\omega = 1$), the *Hybrid* (with ω a free parameter) and the *Hybrid+* models, and then sought to determine the model that was best supported by the behavioural data. To note that rather than fitting all possible algorithmic combinations, the *Hybrid* model variants tested just combined the best MF, the $SARSA$, and the best MB, the $forward_1$ algorithms. Regarding the *Hybrid+* model fit, it used the already

best fitting *Hybrid* model variant for each subject but with all free parameters, including the extra three parameters, estimated afresh.

For the fixed-effects analyses, the Bayesian information criterion, *BIC*, (Schwarz, 1978) based on the negative log-likelihood, was determined for each algorithm tested. Since the *Hybrid* algorithm nested MF and MB algorithms, likelihood-ratio tests (*LRT s*) were also used to compare it against the other learning approaches.

For the mixed-effects analyses, model comparison was achieved by computing, for each model $\mathcal{M}$ and given all the observed first and second-stage choices $\mathcal{A}$, the posterior log likelihood $logP(\mathcal{M}|\mathcal{A})$. Because each of the models tested are equally likely a priori, the model log likelihood $logP(\mathcal{A}|\mathcal{M})$ is the measure to examine. To approximate this quantity at the subject-level and at the individual session-level, we followed a similar approach as the one described in Huys et al. (2011). The approximation at the subject-level was obtained via Kass and Raftery (1995):

$$\begin{aligned} logP(\mathcal{A}|\mathcal{M}) &= \int d\boldsymbol{\theta} P(\mathcal{A}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathcal{M}) \\ &\approx -\frac{1}{2} BIC_{int} = logP(\mathcal{A}|\hat{\boldsymbol{\theta}}^{ML}) - \frac{1}{2}|\mathcal{M}| log(|\mathcal{A}|) \end{aligned} \tag{3.17}$$

where $|\mathcal{A}|$ is the total number of trials performed by the subject in all sessions, and $|\mathcal{M}|$ is the number of prior parameters fitted (mean and variance for each parameter). The subscript "int" to the *BIC* was added because the log likelihood $logP(\mathcal{A}|\hat{\boldsymbol{\theta}}^{ML})$ is not the sum of individual likelihoods, but the sum of integrals over the individual session's parameters approximated via sampling:

$$\begin{aligned} logP(\mathcal{A}|\hat{\boldsymbol{\theta}}^{ML}) &= \sum_i log \int d\boldsymbol{h} P(\boldsymbol{A}_i|\boldsymbol{h}) P(\boldsymbol{h}|\hat{\boldsymbol{\theta}}^{ML}) \\ &\approx \sum_i log \frac{1}{K} \sum_{k=1}^{K} P(\boldsymbol{A}_i|\boldsymbol{h}^k) \end{aligned} \tag{3.18}$$

where $K = 1000$ indicates the number of samples drawn from the empirical prior distribution $\boldsymbol{h}^k \sim P(\boldsymbol{h}|\hat{\boldsymbol{\theta}}^{ML})$. This ensures comparison of not how well a particular model fits the data when its parameters are optimised, but rather how well it fits on average under the random effects empirical prior over the parameters.

In addition to comparing the BIC_{int}, which is akin to a likelihood ratio test, the mixed-effects model comparison also included the exceedance probability (Stephan et al., 2009)

of each model being more likely than any of the other models tested. The computation of this latter measure involved the model likelihood obtained in the fitting of the maximum a posteriori estimates for each session of a given subject and the calculations were performed using the `spm_BMS` function contained in SPM8 (http://www.fil.ion.ucl.ac.uk/spm/). To assess how well each model performed on subject's data, the overall predictive probability for all choices in all trials and sessions, $P(\mathcal{A}|\hat{\boldsymbol{\theta}}^{ML}) = \sqrt[TN]{P(\mathcal{A}|\hat{\boldsymbol{\theta}}^{ML})}$, was calculated and tested, according to a binomial test, whether this was greater than chance.

We further tested the best-fitting models and their respective distribution of parameter priors, $\mathcal{N}(\hat{\boldsymbol{\mu}}^{ML}, \hat{\boldsymbol{\sigma}}^{ML})$, by using them to simulate choice data for each subject (100 simulation runs for each session) on the task respecting the exact same reward structures as present in the behavioural data. We then performed the same descriptive and logistic analysis as used for choice behaviour. Finally, Pearson's linear correlation coefficients were used to assess the relation between the behavioural computational modelling estimates (observed or simulated) and other variables, such as the logistic regression coefficients.

3.4 Results

Both subjects engaged well with the task: subject C performed 15585 trials over 30 sessions ($M = 520$; $SD = 66$); and subject J performed 14664 trials over 27 sessions ($M = 543$; $SD = 101$). There was no significant overall preference for any first-stage choice picture across sessions (C: $t(29) = -0.88$, $p = 0.387$, $g = -0.16$; J: $t(26) = 0.06$, $p = 0.950$, $g = 0.01$). Given the three physically possible actions, potential side biases in first-stage choice were analysed and there were small but significant differences in both subjects (C: $F(2, 87) = 53.42$, $p < 0.001$, $\eta^2 = 0.55$, with relative preferences of left < right < down surviving tests for multiple comparisons; J: $F(2, 78) = 37.21$, $p < 0.001$, $\eta^2 = 0.49$, with relative preferences of right < up < left surviving tests for multiple comparisons).

We first assessed MF and MB-RL without using the trial-by-trial computational modelling, taking advantage of the task design and the different expected patterns by which reward as well as transition type (common or rare) impact future first-stage choices. MF-RL does not exploit information about the task's structure, and so it predicts there should be no difference in the probability of repeating a first-stage choice dependent on the outcome following a common versus a rare trial (simulations in Fig. 3.2A). By contrast, MB-RL predicts that such a difference will exist (simulations in Fig. 3.2B). Additionally, we considered a hybrid approach which assumes that first-stage choices are computed as a weighted sum of the state-action values from MF and MB learning systems and predicts an intermedi-

ate choice behaviour pattern (simulations in Fig. 3.2C; to note that these results are much closer to the MB-RL simulations because our simulations used the parameters best fit to the subjects' data and the MB weight estimated was close to 90%).

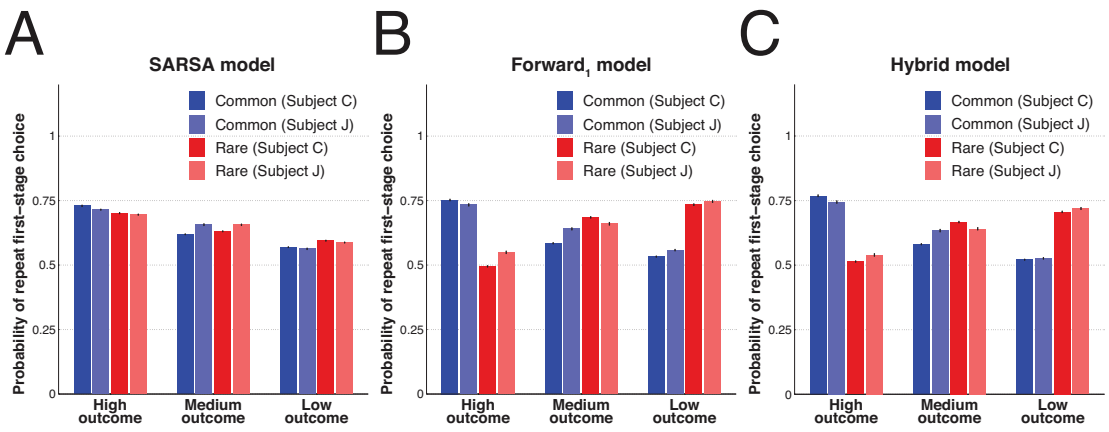


Fig. 3.2 **Comparison of the impact of both reward and transition information on first-stage simulated behaviour from each learning strategy.** Likelihood of repeating first-stage simulated choice as a function of outcome level and transition type (common transition in blue and rare transition in red) for the best pure model-free $SARSA$ algorithm (**A**), the best pure model-based $Forward_1$ algorithm (**B**) and the best *Hybrid* model (**C**). Values were averaged across all sessions, and across 100 simulation runs for each session of the respective subject (respecting the exact same reward structure and using the parameters best fit to the subjects' data within each class of algorithm). Error bars depict standard errors of the mean.

As presented in Fig. 3.3, both experimental subjects were much more likely to repeat the same first-stage choice if a high outcome was achieved through a common transition than when the same reward was obtained following a rare path. The opposite pattern was seen following the worse outcome. This behaviour implies knowledge of the state-transition structure of the task and, therefore, could be seen as a signature of MB-RL. Interestingly, the medium outcome in both subjects elicited a response profile more comparable to the worst (negative) outcome scenario. This suggests that this outcome did not quite have a neutral subjective value, likely because the utility of the reward magnitude was insufficient to overcome the opportunity cost of the imposed delay to juice delivery.

To better quantify the contributions in behaviour of both learning systems, we performed logistic regression analyses on the first-stage choices (i.e. aiming to predict the chosen picture at first-stage). In this case, the predictions from the theory are that a pure MF learner will have the main effect of reward (i.e. the interaction between previous outcome level, R, and previous first-stage choice, C) as the dominant predictor influencing future

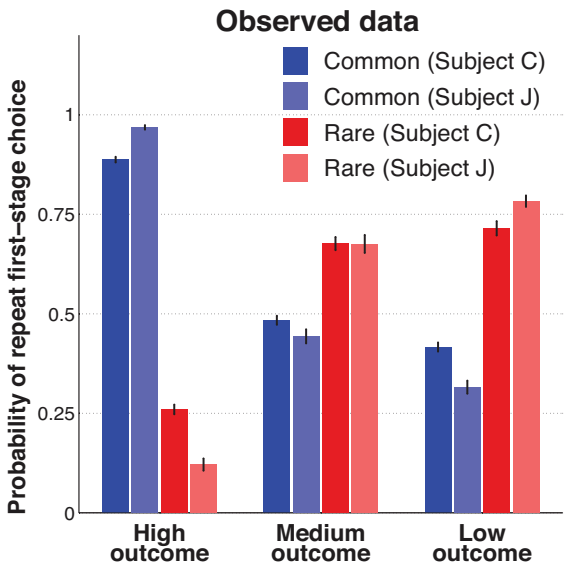


Fig. 3.3 **The impact of both reward and transition information on observed first-stage choice behaviour.** The probability of repeating the same first-stage choice, averaged across sessions for each subject, as a function of outcome level and transition type (common transition in blue; rare transition in red) of the previous trial. Error bars depict standard errors of the mean.

first-stage choices (MF simulated first-stage choice in Fig. 3.4A). On the other hand, MB behaviour will reveal a rather prominent reward $\times$ transition effect (i.e. the interaction between previous outcome level, R, previous transition, T, and previous first-stage choice, C; MB simulated first-stage choice in Fig. 3.4B). However, choices derived from an agent with concurrent use of both learning systems (simulations in Fig. 3.4C) will show not only a significant main effect of reward (which is of MF importance) but also the influence of the reward $\times$ transition effect (which is of MB importance).

We assessed how our subjects weighed the previous trial's information in their first-stage decision (Equation 3.1; see Fig. 3.5 and Table 3.1). We found a significant main effect of reward in favour of first-stage choice repetition following higher outcome levels (C: fixed-effect estimates M = 1.79, SEM = 0.09, $t(29)$ = 20.75, $p < 0.001$, g = 3.79, random-effect estimate = 1.61, 95% CI [1.44,1.77]; J: fixed-effect estimates M = 3.16, SEM = 0.18, $t(26)$ = 16.47, $p < 0.001$, g = 3.17, random-effect estimate = 2.60, 95% CI [2.20,3.00]; see $R \times C$ predictor in Fig. 3.5). In addition, a significant reward $\times$ transition effect was also present, and reflected the adaptive switch in first-stage choice following a high outcome obtained through a rare transition (C: fixed-effect estimates M = -7.86, SEM = 0.39, $t(29) = -19.91$, $p < 0.001$, $g = -3.63$, random-effect estimate = -7.66,

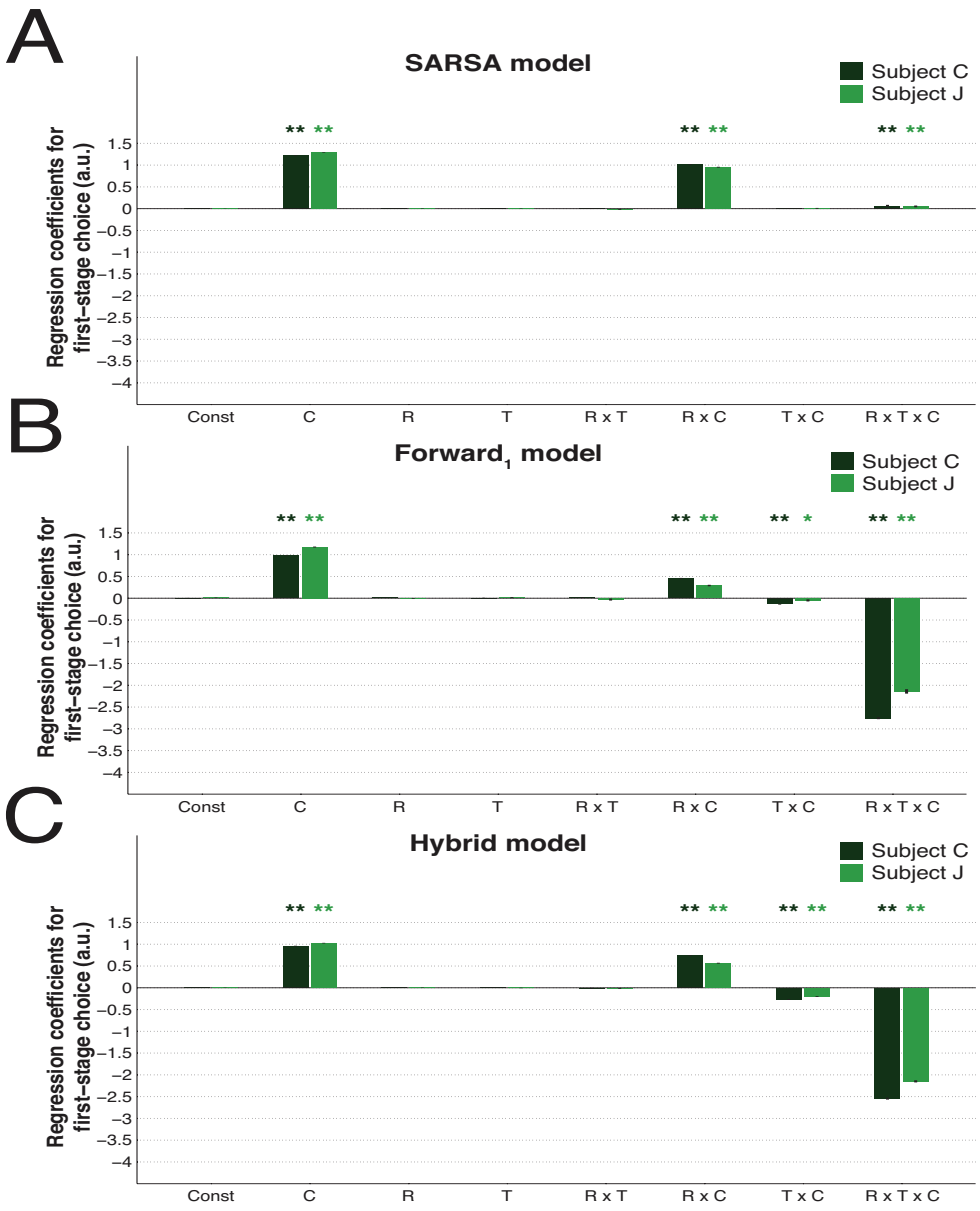


Fig. 3.4 **Logistic regression on simulated first-stage chosen picture from each learning strategy, using the results from the previous trial's predictor variables.** The variables used as predictors of the dependent variable first-stage choice (1=car picture, 0=watering can picture) were: C is previous first-stage choice (1=car picture, 0=watering can picture); R is previous outcome level (assumed as continuous and with low=1, medium=2, high=3); and T is previous transition (rare=1, common=0). Const is the constant term. Predictors were mean centred and continuous variables were also scaled by dividing them by two standard deviations (adjustments made before the computation of the interaction terms). Results for simulated choice behaviour (100 simulations per session for each subject and respecting the exact same reward structure) generated using the best-fitted mixed-effects parameters of the pure model-free $SARSA$ algorithm (**A**), pure model-based $Forward_1$ algorithm (**B**) and *Hybrid* model (**C**). Bar and error bar values correspond, respectively, to the mean and SE of the fixed-effects coefficients. ** for $\alpha = 0.01$ and * for $\alpha = 0.05$ in two-tailed one sample t-test with null-hypothesis mean equal to zero for the fixed-effects coefficients.

95% CI [-8.40,-6.93]; J: fixed-effect estimates M = -12.68, SEM = 0.09, $t(26) = -20.21$, $p < 0.001$, $g = -3.89$, random-effect estimate = -15.98, 95% CI [-18.36,-13.60]; see $R \times T \times C$ predictor in Fig. 3.5). Moreover, these two logistic predictors were not only both significantly different from zero (C: all fixed-effects F-tests with $p < 0.001$, random-effects $F(2) = 386.07$, $p < 0.001$; J: all fixed-effects F-tests with $p < 0.001$, random-effects $F(2) = 173.68$, $p < 0.001$), but the reward $\times$ transition effect was significantly greater than the main effect of reward (C: all fixed-effects F-tests with $p < 0.001$, random-effects $F(2) = 577.68$, $p < 0.001$; J: all fixed-effects F-tests with $p < 0.001$, random-effects $F(2) = 231.14$, $p < 0.001$).

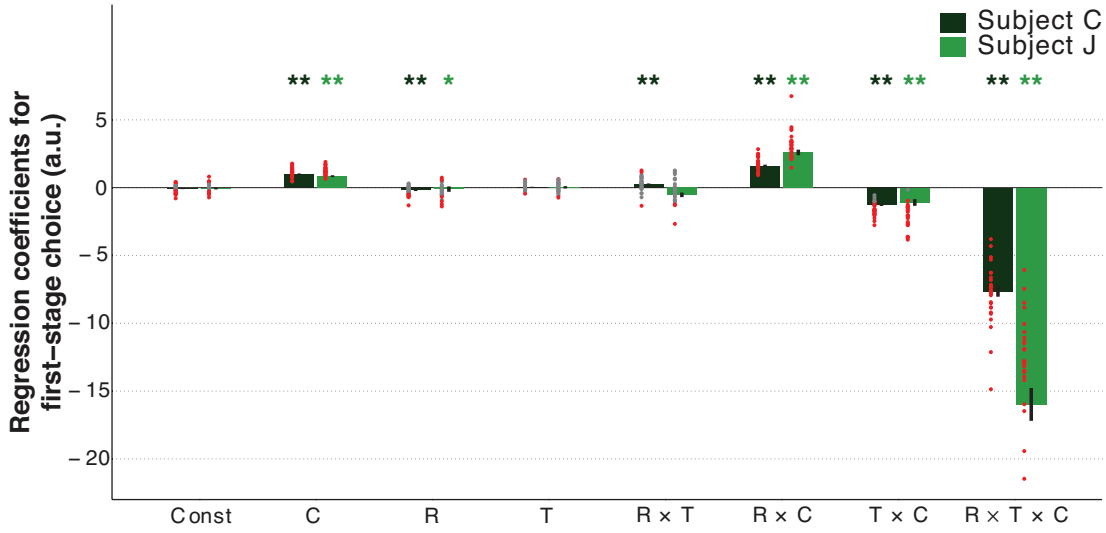


Fig. 3.5 **Logistic regression on observed first-stage chosen picture using predictor variables from the previous trial.** The variables from the previous trial t used as predictors of the dependent variable first-stage choice (1=car picture, 0=watering can picture) were: C is previous first-stage choice (1=car picture, 0=watering can picture); R is previous outcome level (assumed as continuous and with low=1, medium=2, high=3); and T is previous transition (rare=1, common=0). Const is the constant term. Predictors were mean centred and continuous variables were also scaled by dividing them by two standard deviations (adjustments made before the computation of the interaction terms). Each dot represents the fixed-effects regression estimate for a given session (coloured red when $p < 0.05$ and grey otherwise). Bar and error bar values correspond, respectively, to the mixed-effect regression estimate and its standard error. ** for $\alpha = 0.01$ and * for $\alpha = 0.05$ in two-tailed one sample t-test with null-hypothesis mean equal to zero for the fixed-effects estimates.

Table 3.1 **Fixed and mixed-effects logistic regression results for the previous trial predictors of observed first-stage chosen picture**

	Fixed-effects analysis		Mixed-effects analysis	
	Subject C	Subject J	Subject C	Subject J
Const	-0.10 (0.05)	-0.06 (0.07)	-0.06 (0.05)	-0.05 (0.08)
C_{t-1}	1.13 (0.06)**	1.20 (0.06)**	0.98 (0.06)**	0.85 (0.07)**
R_{t-1}	-0.26 (0.06)**	-0.28 (0.09)**	-0.19 (0.06)**	-0.11 (0.20)
T_{t-1}	0.07 (0.05)	0.03 (0.07)	0.02 (0.04)	0.03 (0.09)
$R_{t-1} \times T_{t-1}$	0.29 (0.10)**	-0.20 (0.17)	0.22 (0.09)*	-0.53 (0.17)**
$R_{t-1} \times C_{t-1}$	*1.79 (0.08)***	*3.07 (0.19)***	*1.61 (0.08)***	*2.60 (0.20)***
$T_{t-1} \times C_{t-1}$	-1.54 (0.09)**	-2.10 (0.18)**	-1.27 (0.08)**	-1.09 (0.25)**
$R_{t-1} \times T_{t-1} \times C_{t-1}$	*-7.86 (0.39)***	*-12.19 (0.68)***	*-7.66 (0.38)***	*-15.98 (1.21)***

The variables from the previous trial t used as predictors of the dependent variable first-stage choice (1=car picture, 0=watering can picture) were: C is first-stage choice (1=car picture, 0=watering can picture); R is outcome level (assumed as continuous and with low=1, medium=2, high=3); and T is transition (rare=1, common=0). Const is the constant term. Predictors were mean centred and continuous variables were also scaled by dividing them by two *SD* (adjustments made before the computation of the interaction terms). In italic are the predictors of interest, such as the reward main effect ($R_{t-1} \times C_{t-1}$) and the reward $\times$ transition effect ($R_{t-1} \times T_{t-1} \times C_{t-1}$). Values of fixed-effects results are the mean and in between brackets the *SEM* of the regression coefficients across sessions; mixed-effects results are the estimated regression coefficients and in between brackets their *SE*. ** for $\alpha = 0.01$ and * for $\alpha = 0.05$ in either two-tailed one sample t-test with null-hypothesis mean equal to zero for the fixed-effects results or confidence interval estimation for the mixed-effects results.

It is important to note that both subjects presented a small but significant main effect of transition (i.e. the interaction term between the previous transition, T, and the previous first-stage choice, C) on first-stage choice (C: fixed-effect estimates M = -1.54, SEM = 0.09, $t(29) = -16.20$, $p < 0.001$, $g = -2.96$, random-effect estimate = -1.27, 95% CI [-1.42,-1.11]; J: fixed-effect estimates M = -2.16, SEM = 0.18, $t(26) = -11.92$, $p < 0.001$, $g = -2.30$, random-effect estimate = -1.09, 95% CI [-1.58,-0.61]; see $T \times C$ predictor in Fig. 3.5), a result that we had not expected. However, a similar transition main effect was present in the analysis of simulated data derived from the best RL model (see $T \times C$ predictor in Fig. 3.4C). Hence, specific correlations within the task design and the reward structure are likely the cause of this effect. Finally, both subjects tended to perseverate on the same choice from trial to trial irrespective of any other variable (C: fixed-effect estimates M = 1.13, SEM = 0.06, $t(29)$ = 17.87, $p < 0.001$, $g = -3.26$, random-effect estimate = 0.98, 95% CI [0.86,1.11]; J: fixed-effect mean estimates M = 1.25, SEM = 0.06, $t(26)$ = 21.90, $p < 0.001$, g = 4.21, random-effect estimate = 0.85, 95% CI [0.71,0.98], see C predictor in Fig. 3.5). A similar choice stickiness has been previously described in non-human primates, as well as in humans performing a similar two-step task (Daw et al., 2011; Lau and Glimcher, 2005).

Finally, according to both MF and MB-RL, events in recent trials are expected to exert a greater influence than those in more distant trials – indeed, the influence typically decays exponentially. In order to test this hypothesis, a further logistic regression analysis on first-stage choice was performed taking into account more than just the last trial's reward and transition information (Equation 3.1; Figs. 3.6 and 3.7; and Table 3.2). We found that first-stage choice contribution from both reward history (see $R \times C$ predictors in Fig. 3.7) and combined reward and transition information (see $R \times T \times C$ predictors in Fig. 3.7) obtained up to five trials back reduced exponentially with trials into the past (decay constants of reward main effect / reward $\times$ transition interaction for C: -0.78, 95% CI [-0.98,-0.56] / -0.94, 95% CI [-1.1,-0.78] and for J: of -1.62, 95% CI [-2.08,-1.17] / -1.50, 95% CI [-1.77,-1.22]). Overall, the results of our behavioural analysis are consistent with a first-stage choice behaviour where both MF and MB-RL strategies coexist but, despite this hybrid learning feature, MB control significantly dominated the behaviour of both subjects (compare coefficient weights of the predictors $R \times C$ versus $R \times T \times C$ in Fig. 3.5 and 3.7).

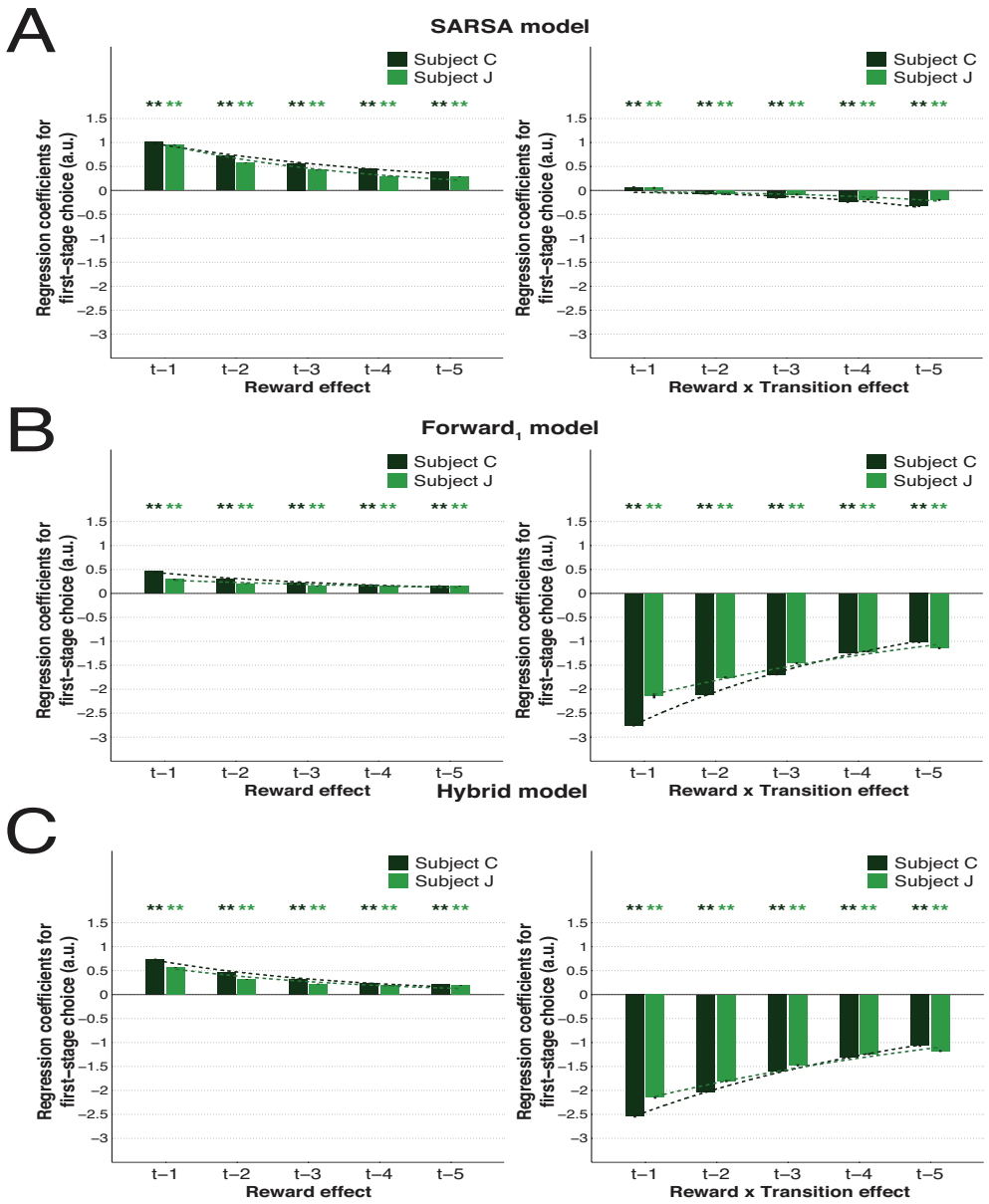


Fig. 3.6 **The impact of both reward and transition information from the five previous trials on simulated first-stage chosen picture according to each learning strategy.** Multiple logistic regression results on simulated first-stage chosen picture data (100 simulations per session for each subject and respecting the exact same reward structure) generated using the best-fitted mixed-effects parameters of the pure model-free $SARSA$ algorithm (**A**), pure model-based $Forward_1$ algorithm (**B**) and *Hybrid* model (**C**) for the main effect of reward (left column) and reward $\times$ transition interaction term (right column) from the five previous trials. Bar and error bar values correspond, respectively, to the mean and SE of the fixed-effects coefficients. Dashed lines illustrate the exponential best fit on the mean fixed-effects coefficients of each trial into the past. ** for $\alpha = 0.01$ and * for $\alpha = 0.05$ in two-tailed one sample t-test with null-hypothesis mean equal to zero for the fixed-effects estimates.

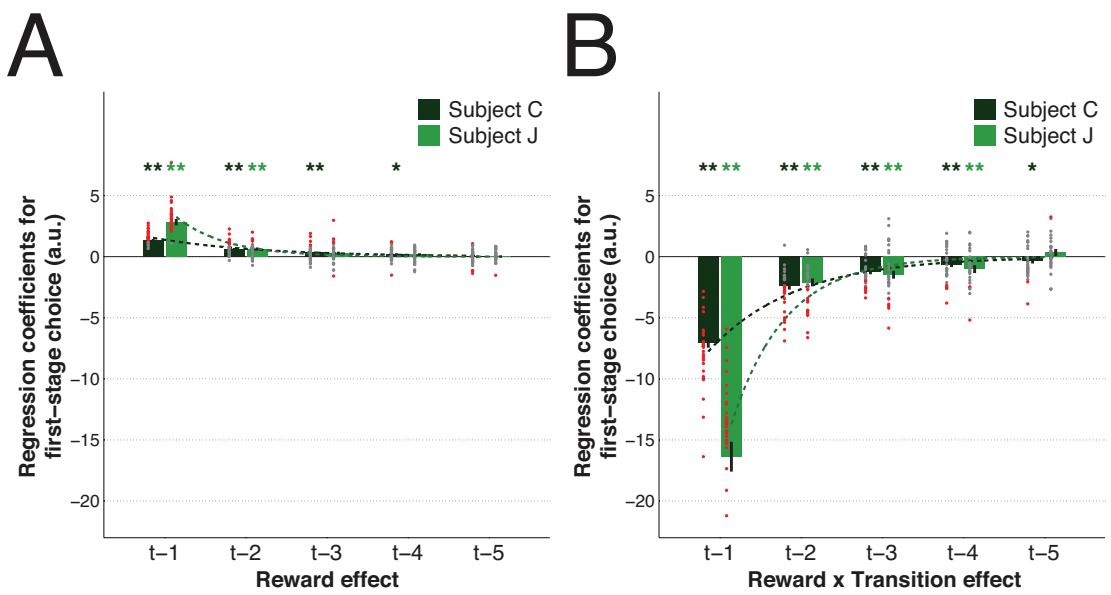


Fig. 3.7 **Logistic regression on observed first-stage chosen picture using predictor variables from the five previous trials.** For the given trial t, the variables used as predictors of the dependent variable first-stage choice (1=car picture, 0=watering can picture) were: C is first-stage choice (1=car picture, 0=watering can picture); R is outcome level (assumed as continuous and with low=1, medium=2, high=3); and T is transition (rare=1, common=0). Const is the constant term. Predictors were mean centred and continuous variables were also scaled by dividing them by two standard deviations (adjustments made before the computation of the interaction terms). Each dot represents the fixed-effects regression estimate for a given session (coloured red when $p < 0.05$ and grey otherwise). Bar and error bar values correspond, respectively, to the mixed-effect regression estimate and its standard error. ** for $\alpha = 0.01$ and * for $\alpha = 0.05$ in two-tailed one sample t-test with null-hypothesis mean equal to zero for the fixed-effects estimates.

Table 3.2 **Fixed and mixed-effects logistic regression results for predictors of first-stage chosen picture up to five trials back**

	Fixed-effects analysis		Mixed-effects analysis	
justifyn Predictors	Subject C	Subject J	Subject C	Subject J
Const	-0.11 (0.06)	-0.08 (0.08)	-0.06 (0.05)	-0.07 (0.08)
C_{t-1}	0.71 (0.06)	0.91 (0.07)**	0.56 (0.05)**	0.57 (0.07)**
R_{t-1}	-0.35 (0.06)**	-0.32 (0.11)**	-0.22 (0.06)**	-0.12 (0.22)
T_{t-1}	0.07 (0.05)	0.07 (0.08)	0.01 (0.05)	0.05 (0.09)
$R_{t-1} \times T_{t-1}$	0.27 (0.12)*	-0.18 (0.18)	0.21 (0.11)	-0.45 (0.18)*
$R_{t-1} \times C_{t-1}$	*1.54 (0.10)***	*3.26 (0.22)***	*1.36 (0.09)***	*2.82 (0.26)***
$T_{t-1} \times C_{t-1}$	-1.92 (0.11)**	-2.53 (0.21)**	-1.43 (0.08)**	-1.39 (0.29)**
$R_{t-1} \times T_{t-1} \times C_{t-1}$	*-7.85 (0.50)***	*-13.76 (0.70)***	*-7.06 (0.39)***	*-16.37 (1.22)***
C_{t-2}	0.37 (0.05)**	0.17 (0.08)	0.32 (0.04)**	0.11 (0.07)
R_{t-2}	-0.05 (0.05)	0.03 (0.08)	-0.03 (0.05)	0.03 (0.08)
T_{t-2}	0.11 (0.06)	0.07 (0.07)	0.06 (0.04)	0.08 (0.05)
$R_{t-2} \times T_{t-2}$	0.28 (0.11)*	0.17 (0.14)	0.14 (0.09)	0.14 (0.14)
$R_{t-2} \times C_{t-2}$	*0.70 (0.10)***	*0.57 (0.12)***	*0.62 (0.08)***	*0.63 (0.13)***
$T_{t-2} \times C_{t-2}$	-0.72 (0.13)**	-0.13 (0.15)	-0.67 (0.09)**	-0.09 (0.16)
$R_{t-2} \times T_{t-2} \times C_{t-2}$	*-2.75 (0.31)***	*-2.68 (0.33)***	*-2.43 (0.26)***	*-2.11 (0.32)***
C_{t-3}	0.17 (0.06)**	0.07 (0.09)	0.17 (0.05)**	0.06 (0.08)
R_{t-3}	0.12 (0.06)	-0.02 (0.08)	0.07 (0.04)	-0.03 (0.08)
T_{t-3}	0.15 (0.06)*	0.07 (0.06)	0.12 (0.04)**	0.09 (0.05)
$R_{t-3} \times T_{t-3}$	0.09 (0.11)	-0.19 (0.15)	0.11 (0.10)	-0.14 (0.15)
$R_{t-3} \times C_{t-3}$	*0.31 (0.11)***	*0.30 (0.17)*	*0.33 (0.08)***	*0.28 (0.13)**
$T_{t-3} \times C_{t-3}$	-0.26 (0.11)*	0.19 (0.15)	-0.21 (0.09)*	0.25 (0.11)*
$R_{t-3} \times T_{t-3} \times C_{t-3}$	*-1.33 (0.22)***	*-1.31 (0.40)***	*-1.26 (0.19)***	*-1.48 (0.32)***
C_{t-4}	0.05 (0.06)	-0.07 (0.07)	0.04 (0.04)	-0.05 (0.06)
R_{t-4}	0.03 (0.06)	-0.04 (0.06)	-0.19 (0.06)	-0.04 (0.05)
T_{t-4}	0.02 (0.06)	0.04 (0.06)	0.02 (0.06)	0.03 (0.05)
$R_{t-4} \times T_{t-4}$	0.04 (0.10)	-0.11 (0.15)	0.06 (0.09)	-0.17 (0.17)
$R_{t-4} \times C_{t-4}$	*0.23 (0.10)**	*0.07 (0.11)*	*0.20 (0.08)**	*0.17 (0.11)*
$T_{t-4} \times C_{t-4}$	0.06 (0.12)	0.23 (0.14)	0.06 (0.09)	0.31 (0.15)*
$R_{t-4} \times T_{t-4} \times C_{t-4}$	*-0.70 (0.25)***	*-0.89 (0.29)***	*-0.66 (0.19)***	*-1.02 (0.32)***
C_{t-5}	0.06 (0.05)	0.15 (0.05)*	0.06 (0.04)	0.11 (0.05)*
R_{t-5}	0.10 (0.05)	-0.02 (0.05)	0.11 (0.04)**	-0.04 (0.06)
T_{t-5}	0.07 (0.05)	0.01 (0.05)	0.02 (0.04)	0.02 (0.05)
$R_{t-5} \times T_{t-5}$	0.05 (0.10)	-0.05 (0.16)	0.08 (0.09)	0.12 (0.15)
$R_{t-5} \times C_{t-5}$	*0.01 (0.12)*	*0.10 (0.10)*	*-0.01 (0.11)*	*0.01 (0.10)*
$T_{t-5} \times C_{t-5}$	-0.02 (0.15)	0.22 (0.11)	-0.04 (0.12)	0.17 (0.11)
$R_{t-5} \times T_{t-5} \times C_{t-5}$	*-0.50 (0.24)**	*0.22 (0.26)*	*-0.35 (0.22)*	*-0.34 (0.29)*

Results are mean (SE) of the regression coefficients across sessions. For the given trial *t*, the variables used as predictors of the dependent variable first-stage choice (1=car picture, 0=watering can picture) were: C is first-stage choice (1=car picture, 0=watering can picture); R is outcome level (assumed as continuous and with low=1, medium=2, high=3); and T is transition (rare=1, common=0). Const is the constant term. Predictors were mean centred and continuous variables were also scaled by dividing them by two standard deviations (adjustments made before the computation of the interaction terms). In italic are the predictors of interest. ** for $\alpha = 0.01$ and * for $\alpha = 0.05$ in either two-tailed one sample t-test with null-hypothesis mean equal to zero for the fixed-effects results or confidence interval estimation for the mixed-effects results.

Computational modelling results

A variety of full MF as well as full MB-RL algorithms were first fitted to each subject's trial-by-trial choices using both fixed-effects (individual fits for each session) and mixed-effects (taking parameters of each subject to be random effects across sessions) fitting procedures. The complexity-adjusted likelihoods of the models were then compared to determine which best fit behaviour (Tables 3.3, 3.4 and 3.5). In line with previous studies (Daw et al., 2011; Gläscher et al., 2010), we also considered a *Hybrid* model in which the best MF and MB algorithms operated in parallel, and with their decision values being combined to determine the choice probabilities (Table 3.6). The relative weight of MB versus MF control was considered as a further free parameter, ω (high ω indicating a bias towards MB-RL choice) on each session. All procedures not only aimed to validate the regression findings showing contributions of both systems to behavioural choice, but also attempted to find the optimal combination of parameters within each RL strategy.

Table 3.7 summarises the model comparison measures for the best fitting algorithms within each learning strategy and for each subject. The results indicate that choice behaviour from both subjects was best explained by a combined MF and MB strategy, corroborating the dual-control findings obtained with the initial behavioural analysis. The fitting to the data of the best *Hybrid* model (among its different possible variants as shown by Table 3.6) was significantly better than chance (predictive probabilities significantly exceeded chance level for both subjects, $p < 0.05$), and had lower *BIC* scores as well as better exceedance probability when compared to pure MF and pure MB models (being only beaten by the *Hybrid+* model, which we discuss below). This winning *Hybrid* model combined the $SARSA$ algorithm (the best approach within pure MF approaches as shown by Tables 3.3 and 3.4) without an eligibility trace parameter, and the $Forward_1$ algorithm (the best pure MB approach as in Table 3.5) where the state-transition probabilities are assumed to be known from the beginning of the task. Furthermore, this model also incorporated the previously mentioned choice perserverance tendency in both first- and second-stage choices (C: κ parameter value is the same for first- and second-stage choices; J: two separate κ parameter values for each stage choice) and a common learning rate at both decision stages.

Table 3.8 displays the empirical prior distributions over the parameters for each of the best algorithms. The mixed-effects analysis led to a significantly better fit than the fixed-effects analysis (with the exceptions of pure MF and pure MB algorithms in subject J; all remaining t-tests on BIC versus BIC_{int} with $p < 0.05$), and indeed the fixed-effects parameter estimates mostly conformed to a normal distribution with relatively similar values to the mixed-effects results (on probability plots inspection). With regard to the balance between

Table 3.3 **Model comparison results for the model-free *SARSA* models**

SARSA	Fixed-effects *BIC* sum		Mixed-effects BIC_{int}	
parameters	Subject C	Subject J	Subject C	Subject J
α, β, κ_1	35873	34670	35717	34538
α, β, κ_2	36581	35191	36444	35170
α, β, κ	35814	34133	35679	34043
α, β, λ	36330	35157	36149	35171
$\alpha_1, \alpha_2, \beta$	36511	35671	36494	35700
α, β_1, β_2	36532	35682	36393	35596
$\alpha_1, \alpha_2, \beta_1, \beta_2$	36615	35770	36464	35775
$\alpha_1, \alpha_2, \beta, \kappa_1$	35822	34759	35689	34588
$\alpha_1, \alpha_2, \beta, \kappa_2$	36584	35295	36430	35274
$\alpha_1, \alpha_2, \beta, \kappa$	35821	34252	35687	34087
$\alpha_1, \alpha_2, \beta, \lambda$	36335	35267	36116	35231
$\alpha, \beta_1, \beta_2, \kappa_1$	35785	34742	35455	34458
$\alpha, \beta_1, \beta_2, \kappa_2$	36619	35306	36348	35175
$\alpha, \beta_1, \beta_2, \kappa$	35934	34267	35637	34022
$\alpha, \beta_1, \beta_2, \lambda$	36447	35260	36163	35176
$\alpha, \beta, \kappa_1, \kappa_2$	35908	34256	35624	34069
$\alpha, \beta, \kappa_1, \lambda$	35723	34415	35424	34185
$\alpha, \beta, \kappa_2, \lambda$	36403	34825	36095	34791
$\alpha, \beta, \kappa, \lambda$	35697†	33931†	35420	33750†
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_1$	35895	34859	35494	34509
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_2$	37797	35690	36765	35698
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa$	35949	34388	35723	34088
$\alpha_1, \alpha_2, \beta_1, \beta_2, \lambda$	36478	35360	36142	35272
$\alpha, \beta_1, \beta_2, \kappa_1, \kappa_2$	35873	34370	35408	34049
$\alpha, \beta_1, \beta_2, \kappa_1, \lambda$	35794	34553	35319	34183
$\alpha, \beta_1, \beta_2, \kappa_2, \lambda$	36531	34888	36117	34774
$\alpha, \beta_1, \beta_2, \kappa, \lambda$	35856	34058	35385	33754
$\alpha_1, \alpha_2, \beta, \kappa_1, \kappa_2$	35891	34373	36074	34133
$\alpha_1, \alpha_2, \beta, \kappa_1, \lambda$	35707	34539	35356	34240
$\alpha_1, \alpha_2, \beta, \kappa_2, \lambda$	36426	34936	36074	34856
$\alpha_1, \alpha_2, \beta, \kappa, \lambda$	35725	34061	35382	33788
$\alpha, \beta, \kappa_1, \kappa_2, \lambda$	35784	34055	35348	33780
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_1, \kappa_2$	35983	34494	35439	34087
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_1, \lambda$	35858	34677	35311	34233
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_2, \lambda$	36566	34994	36090	34865
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa, \lambda$	35858	34188	35331	33796
$\alpha, \beta_1, \beta_2, \kappa_1, \kappa_2, \lambda$	35880	34186	35266	33766
$\alpha_1, \alpha_2, \beta, \kappa_1, \kappa_2, \lambda$	35791	34186	35287	33824
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_1, \kappa_2, \lambda$	35946	34312	35260†	33825

† Best fitting *SARSA* model-free learning model for the respective analysis type. Abbreviations: Bayesian Information Criteria (*BIC*); integrated BIC (BIC_{int}); learning rate for first-stage (α_1) and second-stage (α_2) choice; α is when $\alpha_1 = \alpha_2$; inverse temperature for first-stage (β_1) and second-stage (β_2) choice; β is when $\beta_1 = \beta_2$; perseveration for first-stage (κ_1) and second-stage (κ_2) choice; κ is when $\kappa_1 = \kappa_2$; eligibility trace (λ).

Table 3.4 **Model comparison results for the model-free *Q*-learning models**

Q-learning	Fixed-effects *BIC* sum		Mixed-effects BIC_{int}	
parameters	Subject C	Subject J	Subject C	Subject J
α, β	36445	35620	36415	35706
α, β, κ_1	35767	34668	35633	34574
α, β, κ_2	36508	35252	36354	35283
α, β, κ	35736	34162	35620	34101
α, β, λ	36357	35399	36245	35525
$\alpha_1, \alpha_2, \beta$	36484	35729	36428	35749
α, β_1, β_2	36501	35731	36408	35764
$\alpha_1, \alpha_2, \beta_1, \beta_2$	36609	35836	36431	35878
$\alpha_1, \alpha_2, \beta, \kappa_1$	35778	34788	35614	34607
$\alpha_1, \alpha_2, \beta, \kappa_2$	36566	35372	36373	35427
$\alpha_1, \alpha_2, \beta, \kappa$	35792	34297	35615	34088
$\alpha_1, \alpha_2, \beta, \lambda$	36420	35502	36265	35556
$\alpha, \beta_1, \beta_2, \kappa_1$	35761	34780	35483	34589
$\alpha, \beta_1, \beta_2, \kappa_2$	36587	35357	36370	35383
$\alpha, \beta_1, \beta_2, \kappa$	35878	34300	35609	34084
$\alpha, \beta_1, \beta_2, \lambda$	36458	35494	36367	35528
$\alpha, \beta, \kappa_1, \kappa_2$	35819	34281	35560	34127
$\alpha, \beta, \kappa_1, \lambda$	35721	34571	35486	34470
$\alpha, \beta, \kappa_2, \lambda$	36434	35068	36206	35160
$\alpha, \beta, \kappa, \lambda$	35709‡	34095‡	35502	33942
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_1$	35891	34909	35559	34545
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_2$	37797	35773	36742	35552
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa$	35925	34426	35829	34119
$\alpha_1, \alpha_2, \beta_1, \beta_2, \lambda$	36543	35590	36312	35612
$\alpha, \beta_1, \beta_2, \kappa_1, \kappa_2$	35849	34409	35451	34140
$\alpha, \beta_1, \beta_2, \kappa_1, \lambda$	35799	34713	35411	34352
$\alpha, \beta_1, \beta_2, \kappa_2, \lambda$	36540	35121	36301	35121
$\alpha, \beta_1, \beta_2, \kappa, \lambda$	35856	34220	35467	33935‡
$\alpha_1, \alpha_2, \beta, \kappa_1, \kappa_2$	35856	34413	35537	34169
$\alpha_1, \alpha_2, \beta, \kappa_1, \lambda$	35762	34704	35445	34423
$\alpha_1, \alpha_2, \beta, \kappa_2, \lambda$	36510	35143	36227	35169
$\alpha_1, \alpha_2, \beta, \kappa, \lambda$	35786	34224	35492	33969
$\alpha, \beta, \kappa_1, \kappa_2, \lambda$	35788	34217	35437	34006
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_1, \kappa_2$	35979	34546	36026	34133
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_1, \lambda$	35904	34844	35399	34432
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_2, \lambda$	36566	34994	36084	34888
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa, \lambda$	35907	34350	35524	33990
$\alpha, \beta_1, \beta_2, \kappa_1, \kappa_2, \lambda$	35884	34347	35429	34010
$\alpha_1, \alpha_2, \beta, \kappa_1, \kappa_2, \lambda$	35847	34346	35404	34038
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_1, \kappa_2, \lambda$	35992	34473	35376‡	34008

‡ Best fitting *Q*-learning model-free learning model for the respective analysis type. Abbreviations: Bayesian Information Criteria (*BIC*); integrated BIC (BIC_{int}); learning rate for first-stage (α_1) and second-stage (α_2) choice; α is when $\alpha_1 = \alpha_2$; inverse temperature for first-stage (β_1) and second-stage (β_2) choice; β is when $\beta_1 = \beta_2$; perseveration for first-stage (κ_1) and second-stage (κ_2) choice; κ is when $\kappa_1 = \kappa_2$; eligibility trace (λ).

Table 3.5 **Model comparison results for the three algorithms of model-based models**

Forward	Fixed-effects *BIC* sum		Mixed-effects BIC_{int}	
parameters	Subject C	Subject J	Subject C	Subject J
$Forward_1$				
α_2, β	35298	34824	35275	34868
$\alpha_2, \beta_1, \beta_2$	34630	33708	34548	33743
$\alpha_2, \beta, \kappa_1$	34737	34038	34610	33965
$\alpha_2, \beta, \kappa_2$	35425	34572	35271	34619
α_2, β, κ	34818	33634	34701	33616
$\alpha_2, \beta, \kappa_1, \kappa_2$	34856	33753	34595	33658
$\alpha_2, \beta_1, \beta_2, \kappa_1$	34418	33437	34176	33345
$\alpha_2, \beta_1, \beta_2, \kappa_2$	34715	33342	34499	33309
$\alpha_2, \beta_1, \beta_2, \kappa$	34360[§]	33182	34122[§]	33248
$\alpha_2, \beta_1, \beta_2, \kappa_1, \kappa_2$	34505	33071[§]	34143	32837[§]
$Forward_2$				
α_2, β	35304	34831	35279	34872
$\alpha_2, \beta_1, \beta_2$	34642	33732	34556	33758
$\alpha_2, \beta, \kappa_1$	34743	34046	34610	33959
$\alpha_2, \beta, \kappa_2$	35432	34579	35274	34612
α_2, β, κ	34824	33641	34702	33619
$\alpha_2, \beta, \kappa_1, \kappa_2$	34862	33761	34594	33661
$\alpha_2, \beta_1, \beta_2, \kappa_1$	34430	33462	34181	33359
$\alpha_2, \beta_1, \beta_2, \kappa_2$	34727	33366	34508	33324
$\alpha_2, \beta_1, \beta_2, \kappa$	34372	33205	34129	33256
$\alpha_2, \beta_1, \beta_2, \kappa_1, \kappa_2$	34517	33095	34152	32851
$Forward_3$				
α_2, β, ζ	35484	34993	35297	34883
$\alpha_2, \beta_1, \beta_2, \zeta$	34797	33841	34570	33748
$\alpha_2, \beta, \kappa_1, \zeta$	34923	34206	34630	33965
$\alpha_2, \beta, \kappa_2, \zeta$	35611	34742	35288	34633
$\alpha_2, \beta, \kappa, \zeta$	35004	33803	34717	33629
$\alpha_2, \beta, \kappa_1, \kappa_2, \zeta$	35042	33922	34624	33656
$\alpha_2, \beta_1, \beta_2, \kappa_1, \zeta$	34590	33579	34192	33346
$\alpha_2, \beta_1, \beta_2, \kappa_2, \zeta$	34882	33475	34516	33313
$\alpha_2, \beta_1, \beta_2, \kappa, \zeta$	34533	33335	34140	33255
$\alpha_2, \beta_1, \beta_2, \kappa_1, \kappa_2, \zeta$	35229	34091	34145	32841

[§] Best fitting model-based algorithm for the respective analysis type (see text for details about each algorithm). Abbreviations: Bayesian Information Criteria (*BIC*); integrated BIC (BIC_{int}); learning rate for second-stage (α_2) choice; inverse temperature for first-stage (β_1) and second-stage (β_2) choice; β is when $\beta_1 = \beta_2$; perseveration for first-stage (κ_1) and second-stage (κ_2) choice; κ is when $\kappa_1 = \kappa_2$; eligibility trace (λ); ζ is a weight given to state-transition model testing.

Table 3.6 **Model comparison results for the *Hybrid* models**

HYBRID	Fixed-effects *BIC* sum		Mixed-effects BIC_{int}	
parameters	Subject C	Subject J	Subject C	Subject J
$\alpha, \beta, \kappa_1, \omega$	34807	34144	34522	33930
$\alpha, \beta, \kappa_2, \omega$	35435	34642	35148	34592
$\alpha, \beta, \kappa, \omega$	34880	33742	34616	33600
$\alpha, \beta, \lambda, \omega$	35451	34900	35171	34775
$\alpha_1, \alpha_2, \beta, \omega$	35464	35028	35168	34814
$\alpha, \beta_1, \beta_2, \omega$	34515	33638	34246	33548
$\alpha_1, \alpha_2, \beta_1, \beta_2, \omega$	34577	33735	34313	33585
$\alpha_1, \alpha_2, \beta, \kappa_1, \omega$	34973	34292	34561	33923
$\alpha_1, \alpha_2, \beta, \kappa_2, \omega$	35593	34779	35195	34582
$\alpha_1, \alpha_2, \beta, \kappa, \omega$	35045	33891	34644	33617
$\alpha_1, \alpha_2, \beta, \lambda, \omega$	35602	35023	35191	34733
$\alpha, \beta_1, \beta_2, \kappa_1, \omega$	34380	33432	33948	33215
$\alpha, \beta_1, \beta_2, \kappa_2, \omega$	34602	33269	34194	33217
$\alpha, \beta_1, \beta_2, \kappa, \omega$	34326¶	33199	33898¶	33189
$\alpha, \beta_1, \beta_2, \lambda, \omega$	34652	33640	34244	33499
$\alpha, \beta, \kappa_1, \kappa_2, \omega$	34928	33861	34508	33673
$\alpha, \beta, \kappa_1, \lambda, \omega$	34966	34216	34542	33913
$\alpha, \beta, \kappa_2, \lambda, \omega$	35580	34650	35167	34542
$\alpha, \beta, \kappa, \lambda, \omega$	35036	33813	34640	33566
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_1, \omega$	34468	33553	33986	33252
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_2, \omega$	34663	33367	34265	33167
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa, \omega$	34422	33346	33948	33239
$\alpha_1, \alpha_2, \beta_1, \beta_2, \lambda, \omega$	34709	33748	34302	33541
$\alpha, \beta_1, \beta_2, \kappa_1, \kappa_2, \omega$	34468	33063¶	33904	32807¶
$\alpha, \beta_1, \beta_2, \kappa_1, \lambda, \omega$	34528	33473	33952	33182
$\alpha, \beta_1, \beta_2, \kappa_2, \lambda, \omega$	34739	33272	34202	33075
$\alpha, \beta_1, \beta_2, \kappa, \lambda, \omega$	34475	33249	33906	33197
$\alpha_1, \alpha_2, \beta, \kappa_1, \kappa_2, \omega$	35095	34010	34558	33642
$\alpha_1, \alpha_2, \beta, \kappa_1, \lambda, \omega$	35125	34355	34576	33889
$\alpha_1, \alpha_2, \beta, \kappa_2, \lambda, \omega$	35731	34769	35200	34552
$\alpha_1, \alpha_2, \beta, \kappa, \lambda, \omega$	35194	33951	34668	33566
$\alpha, \beta, \kappa_1, \kappa_2, \lambda, \omega$	35087	33935	34524	33612
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_1, \kappa_2, \omega$	34556	33186	33950	32843
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_1, \lambda, \omega$	34612	33599	33986	33234
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_2, \lambda, \omega$	34795	33380	34255	33180
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa, \lambda, \omega$	34568	33392	33940	33235
$\alpha, \beta_1, \beta_2, \kappa_1, \kappa_2, \lambda, \omega$	34616	33105	33903	32917
$\alpha_1, \alpha_2, \beta, \kappa_1, \kappa_2, \lambda, \omega$	35246	34071	34563	33606
$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_1, \kappa_2, \lambda, \omega$	34699	33231	33931	32821

All *Hybrid* models tested included the $SARSA$ algorithm as model-free strategy and the $Forward_1$ model-based algorithm (for more details see text). ¶ Best fitting *Hybrid* model for the respective analysis type. Abbreviations: Bayesian Information Criteria (*BIC*); integrated BIC (BIC_{int}); learning rate for first-stage (α_1) and second-stage (α_2) choice; α is when $\alpha_1 = \alpha_2$; inverse temperature for first-stage (β_1) and second-stage (β_2) choice; β is when $\beta_1 = \beta_2$; perseveration for first-stage (κ_1) and second-stage (κ_2) choice; κ is when $\kappa_1 = \kappa_2$; eligibility trace (λ); ω is the model-based weight.

Table 3.7 **Model comparison results for fixed-effects and mixed-effects best-fitting models from each reinforcement learning approach**

			Fixed-effects analysis			Mixed-effects analysis		
Model	Parameters	Subject	BIC sum	*% BIC Hybrid* better	*% LRT Hybrid* better	BICint	Exc Prob vs. *Hybrid*	Predictive choice probability
SARSA								
	$\alpha_1, \alpha_2, \beta_1, \beta_2, \kappa_1, \kappa_2, \lambda$	C	35697	100	100	35260	< 0.001	0.569
	$\alpha, \beta, \kappa, \lambda$	J	33931	93	100	33750	< 0.001	0.563
$Forward_1$								
	$\alpha_2, \beta_1, \beta_2, \kappa$	C	34360	43	70	34122	0.38	0.579
	$\alpha_2, \beta_1, \beta_2, \kappa$	J	33182	63	93	33248	< 0.001	0.568
Hybrid								
	$\alpha, \beta_1, \beta_2, \kappa, \omega$	C	34326	-	-	33898	-	0.581
	$\alpha, \beta_1, \beta_2, \kappa_1, \kappa_2, \omega$	J	33063	-	-	32807	-	0.572
Hybrid+								
	$\alpha, \beta_1, \beta_2, \kappa, L_1, L_2, L_3, \omega$	**C**	**33247**	**7**	**7**	**32441**	> 0.999	**0.595**
	$\alpha, \beta_1, \beta_2, \kappa_1, \kappa_2, L_1, L_2, L_3, \omega$	**J**	**28888**	**0**	**0**	**28659**	> 0.999	**0.614**

BIC, Bayesian Information Criterion (lower values correspond to better models) sum and % of sessions where the *Hybrid* model was better; % of sessions with *LRT*, likelihood-ratio test favouring the *Hybrid* model; BIC_{int}, is the integrated *BIC* (see text); Exc Prob, is the Bayesian exceedance probability (Stephan et al., 2009) measuring the likelihood that each model is the most common when tested against the *Hybrid* model. Both *Hybrid* and *Hybrid+* models (in bold as it was the overall best model) included the *SARSA* algorithm (as model-free) and the $Forward_1$ (as model-based). Abbreviations: learning rate for first-stage (α_1) and second-stage (α_2); α is when $\alpha_1 = \alpha_2$; inverse temperature for first-stage (β_1) and second-stage (β_2); β is when $\beta_1 = \beta_2$; perseveration for first-stage (κ_1) and second-stage (κ_2); κ is when $\kappa_1 = \kappa_2$; eligibility trace (λ); $L1_1$, $L2_2$ and L_3 are the reinforcement strength (or aversion) for high, medium and low outcome, respectively (see text for full details); ω is the model-based weight.

MF and MB control, the maximum a posteriori (MAP) ω hyperparameter was 86% for subject C and 88% for subject J (MAP estimates for the individual sessions different from 0 and 100% with $p < 0.001$ in all sign test for both subjects), very much in line with the dominance found in the logistic regression analysis. Both decision stages shared the same learning rate and the hyperparameter value was relatively high (C: $\alpha = 0.82$; J: $\alpha = 0.77$), probably due to the non-stationary and deterministic reward structure. On the other hand, differences in choice reliability were found for both decision stages with first-stage choices being more deterministic than second-stage ones (C: $\beta_1 = 6.39$ and $\beta_2 = 2.50$; J: $\beta_1 = 6.97$ and $\beta_2 = 1.68$). Finally, the perseverance parameter was small and positive (C: $\kappa = 0.05$; J: $\kappa_1 = 0.05$ and $\kappa_2 = 0.34$), reflecting the tendency to repeat recently chosen options that we mentioned above.

Model validation and simulation results

Logistic regression and computational results both matched the data relatively well, therefore more explicit analyses were performed to establish the relationship between the two. Such validation procedures also helped to confirm whether the models, together with their best-fitting parameters, could generate as well as fit observed choice behaviour. We therefore used the best RL models for each strategy to simulate choice data on the same two-step decision task and then analysed the simulated choices similarly (Figs. 3.2, 3.4 and 3.6).

Table 3.8 **Best fitting hyperparameter mixed-effects estimates from the best models of each reinforcement learning approach**

Model	Subject	α_1	α_2	β_1	β_2	κ_1	κ_2	λ	ω	L_1	L_2	L_3
SARSA												
	C	0.48 (0.32)	0.84 (0.46)	2.62 (0.04)	2.45 (0.10)	0.19 (0.05)	0.07 (0.02)	0.52 (0.32)	-	-	-	-
	J	0.62 (0.43)		1.93 (0.07)		0.28 (0.06)		0.58 (0.58)	-	-	-	-
$Forward_1$												
	C	-	0.80 (0.47)	6.06 (0.13)	2.52 (0.11)	0.06 (0.01)		-	-	-	-	-
	J	-	0.71 (0.69)	6.04 (0.18)	2.01 (0.12)	0.08 (0.02)		-	-	-	-	-
Hybrid												
	C	0.82 (0.40)		6.39 (0.12)	2.50 (0.11)	0.05 (0.01)		-	0.86 (0.23)	-	-	-
	J	0.77 (0.61)		6.97 (0.18)	1.68 (0.06)	0.05 (0.01)	0.34 (0.10)	-	0.88 (0.31)	-	-	-
Hybrid+												
	C	**0.78 (0.40)**		**4.57 (0.15)**	**2.54 (0.11)**	**0.06 (0.01)**		**-**	**0.86 (0.23)**	**0.25 (0.01)**	**-0.06 (0.02)**	**-0.08 (0.02)**
	J	**0.59 (0.51)**		**4.92 (0.20)**	**1.85 (0.08)**	**0.04 (0.01)**	**0.31 (0.10)**	**-**	**0.88 (0.27)**	**0.51 (0.03)**	**-0.10 (0.06)**	**-0.16 (0.05)**

Both *Hybrid* and *Hybrid+* (in bold as it was the best model) models included the *SARSA* algorithm as model-free strategy and the $Forward_1$ model-based algorithm (see full text for details). Values correspond to mean parameter estimates and in between brackets are the standard deviations of the parameters given on the transformed scale used for parameter fitting. Regarding the parameter nomenclature used (when placed in between parameters, the respective parameter estimate was shared between both first-stage and second-stage): learning rate for first-stage (α_1) and second-stage (α_2) choice; inverse temperature for first-stage (β_1) and second-stage (β_2) choice; perseveration for first-stage (κ_1) and second-stage (κ_2) choice; eligibility trace (λ); $L1_1$, $L2_2$ and L_3 are the reinforcement strength (or aversion) for high, medium and low outcome, respectively (see text for full details); ω is the model-based weight.

Not only did this generated data confirm the expected and previously described differences between MF and MB-RL, but the simulations also showed that when compared against pure MF and pure MB approaches, the *Hybrid* model matched better the repeat first-stage choice pattern (particularly the asymmetry between the high and the low outcome levels when compared to pure MB simulations) as well as the logistic regression results (which exhibited a more reasonable balance between the coefficients of the reward main effect and the reward × transition effect) of subjects' behaviour (Figs. 3.2C and 3.3; 3.4C and 3.5). Although these results supported the qualitative validity of the above computational modelling analyses, they also highlighted some quantitative limitations. One of the most striking differences between the *Hybrid* model simulations and the observed data was the weight given to the most recent trial and, consequently, the discrepancies in the exponential decays (decay constants for the reward main effect/reward × transition for C = -0.78/-0.94 versus simulations = -0.37/-0.22, and J = -1.62/-1.50 versus simulations -0.36/-0.17).

Motivated by these observations, a new model (*Hybrid+*) was conceived in which three additional parameters (one for each outcome level: L_1 for high, L_2 for medium and L_3 for low) were incorporated in the previously best *Hybrid* model. The goal was to improve the model's characterization of the influence of each reward level, particularly the one just received, on subsequent first-stage choices. This influence turned out to reflect something close to a one-step MB effect. That is, if the transition on the trial had been common, the Q_{Hybrid} value of the first-stage choice taken on that trial was boosted (or decreased) with a value dependent on the outcome level received. On the other hand, if the transition had been rare, then the increment (or reduction) was applied to the first-stage choice action not chosen on that trial. This way, a positive value ($L > 0$) denotes the strength of the reinforcement by reward, whereas a negative value ($L < 0$) quantifies the aversion for that particular outcome

level.

Comparisons between the models (Table 3.7) show that this new model fitted the data significantly better than chance (predictive probabilities significantly exceeded chance level in both subjects, $p < 0.05$) and it performed better than the previous ones (all exceedance probability values > 0.99). The extra parameters were justified according to the *BIC*, BIC_{int} and the exceedance probability. An important question is whether the original RL parameters remained stable after such adjustment. Indeed, as presented in Table 3.8, very few changes in the parameters were observed by comparing the *Hybrid* model against the *Hybrid+* model. Ultimately, this newly generated model could only be seen as a significant improvement if it also addressed the limitations found with the simulation results obtained with the previous best *Hybrid* model. Indeed the simulated choice data results generated by the *Hybrid+* model successfully captured not only the observed pattern of repeat probability at first-stage choice (compare Fig. 3.8A and 3.8D), but also the logistic regression profiles of both reward main effect (compare Fig. 3.8B and 3.8E) and reward $\times$ transition effect (compare Fig. 3.8C and 3.8F). Moreover, the best-fitted values of the additional parameters (hyperparameter mean values for C: L_1 = 0.25, L_2 = -0.06, L_3 = -0.08 and J: L_1 = 0.51, L_2 = -0.10, L_3 = -0.16) revealed that high outcome level had a high reinforcement strength but both medium and low outcome had an aversive impact, as previously noted. In conclusion, both model comparison and simulation results supported the validity of the newly generated *Hybrid+* model.

A final complementary approach to relating logistic and computational analyses is by explicitly comparing the coefficients obtained from the regression analysis with the RL parameters from the best fitted *Hybrid+* model. Such correlations can also be assessed on simulated data in order to provide a reference of what can be expected. When the analysis was applied to the choice behaviour across sessions of both experimental subjects, we found that the stronger the effect of the reward $\times$ transition interaction regression coefficients the greater the estimated MB ω parameter of the model (data: $r = -0.69$, $p < 0.001$; simulations: $r = -0.18$, $p < 0.001$; see Fig. 3.9A). We also found a significant negative correlation between the inverse temperature parameter at first-stage choice (with lower values revealing stochasticity in the choice) and the residuals from the regression model (data: $r = -0.44$, $p < 0.001$; simulations: $r = -0.41$, $p < 0.001$; see Fig. 3.9B), and a positive correlation between both logistic and computational first-stage choice perseverance measures (separately for each subject given the different number of κ parameters: data: r = 0.80/0.41, $p < 0.001/0.035$, simulations: r = 0.35/0.08, $p < 0.001/< 0.001$; see Fig. 3.9C).

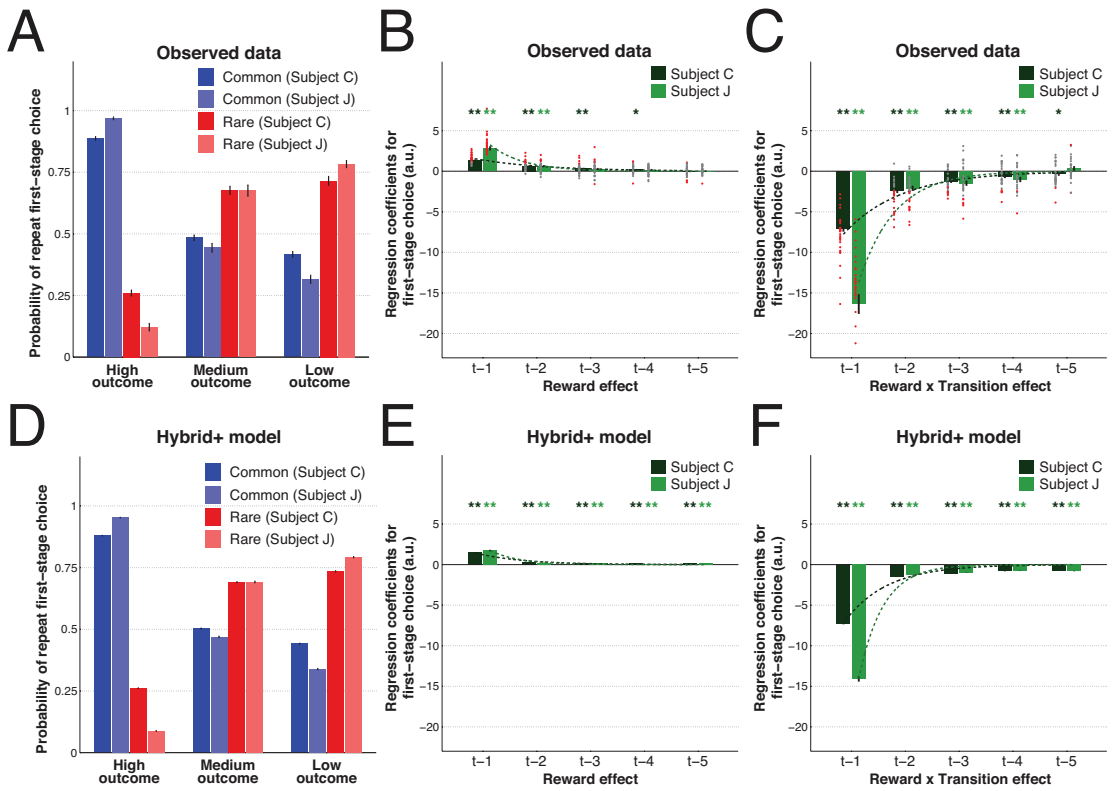


Fig. 3.8 **The impact of both reward and transition information on first-stage chosen picture behaviour.** (**A**) Likelihood of repeating the same first-stage picture choice, averaged across sessions, as a function of outcome and transition on the previous trial. Error bars depict SEM. (**B-C**) Logistic regression results on first-stage chosen picture with the contributions of the reward main effect (**B**) and reward $\times$ transition (**C**) from the five previous trials. (**D-F**) Similar results obtained from simulations (100 runs per session and respecting the exact reward structure subjects experienced) using the best fit *Hybrid+* model. Dots represent fixed-effects coefficients for each session (red when $p < 0.05$, grey otherwise). Bar and error bar values correspond, respectively, to mixed-effect coefficients and their SE. Dashed lines illustrate the exponential best fit on the mean fixed-effects coefficients of each trial into the past. ** $\alpha = 0.01$ and * $\alpha = 0.05$ in two-tailed one sample t-test with null-hypothesis mean equal to zero for the fixed-effects estimates.

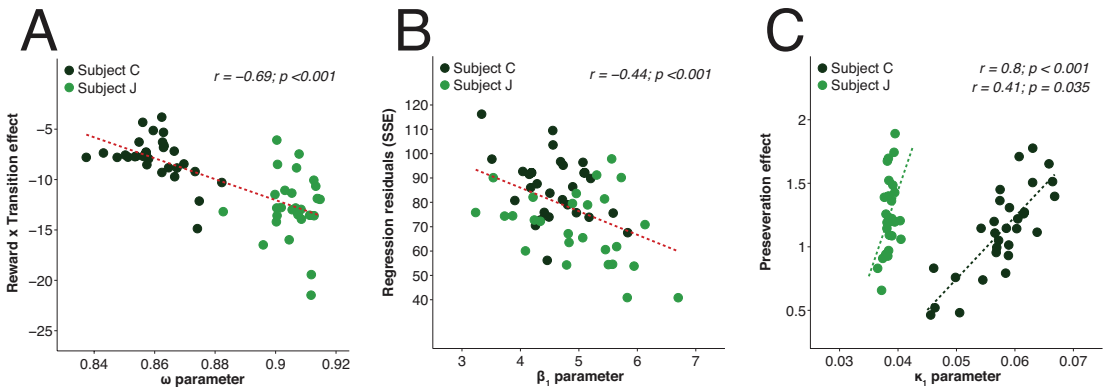


Fig. 3.9 **Correlation in both subjects between the logistic regression estimates and computational modelling parameters across sessions.** (**A**) Relationship between the model-based weight parameter ω obtained from the *Hybrid+* model fitting for each session and the corresponding regression coefficient for the reward $\times$ transition interaction term. The greater the MB weight, the stronger the reward $\times$ transition effect in the logistic regression. (**B**) Positive correlation between the computational preseveration κ_1 parameter and the regression coefficient for repeat first-stage choice independently of reward and transition. (**C**) Relationship between the inverse temperature parameter at first-stage choice β_1 obtained from the *Hybrid+* model fitting and the residual values from the regression model (the greater the β_1 parameter, the better was the logistic regression fit). Dashed lines represent the regression line of the fit for each individual subject or across subjects (in red). r is the Pearson's linear correlation coefficients and p is the p-values across subjects in (**A**) and (**B**), whereas in (**C**) top values are for subject C and bottom values are for subject J.

Lastly, in an attempt to test the possibility that habits were forming progressively (Dickinson, 1985), we assessed the correlations between relevant estimates and their respective session number. We did not find a significant linear increase with time for the main effect of reward (data: $r = -0.04/0.31$, $p = 0.844/0.115$; simulations: $r = -0.44/0.15$, $p = 0.016/0.457$; see Fig. 3.10A), or for a reduction in the ω parameter from the computational modelling (data: $r = 0.04/0.11$, $p = 0.836/0.568$; simulations: $r = 0.21/0.12$, $p = 0.259/0.539$; see Fig. 3.10B). Nonetheless, we did observe that the reward $\times$ transition interaction on first-stage choice tended to reduce with time significantly in both subjects (data: $r = 0.53/0.50$, $p = 0.003/0.008$; simulations: $r = -0.01/-0.24$, $p = 0.959/0.228$; see Fig. 3.10C).

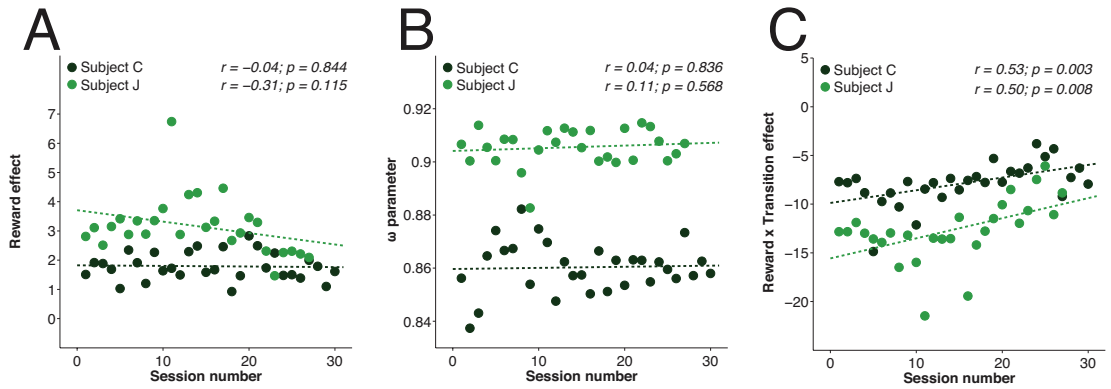


Fig. 3.10 **Evolution across sessions of logistic regression and computational modelling estimates.** Across time and for both subjects, no significant decrease in the regression coefficients for the reward effect (**A**) or model-based weight parameter ω (**B**) was found. However, a significant reduction was found for the effect of the regression coefficients for the reward $\times$ transition effect (**C**) with time (to note that the more positive the regression coefficient the weaker the effect). Dashed lines represent the regression line of the fit for each individual subject. r is the Pearson's linear correlation coefficients and p is the p-values; top values are for subject C and bottom values are for subject J.

These findings suggest other factors may also potentially influence choice across time without a corresponding interference with the MF/MB balance, in particular possible tiredness (cumulative fatigue across days given the concomitant electrophysiological experiments) or attention. Interestingly, first-stage choice behaviour in both subjects became slightly more stochastic with time, as revealed by a gradual decrement in the inverse temperature parameter (data: $r = -0.47/-0.68$, $p = 0.009/< 0.001$; simulations: $r = -0.02/-0.06$, $p = 0.898/0.768$; see Fig. 3.11A). Also relevant was the negative correlation found for first-stage reaction time mean and session number, indicating that subjects got gradually slower throughout the experiment (data: $r = 0.59/0.78$, $p < 0.001/< 0.001$; see Fig. 3.11B).

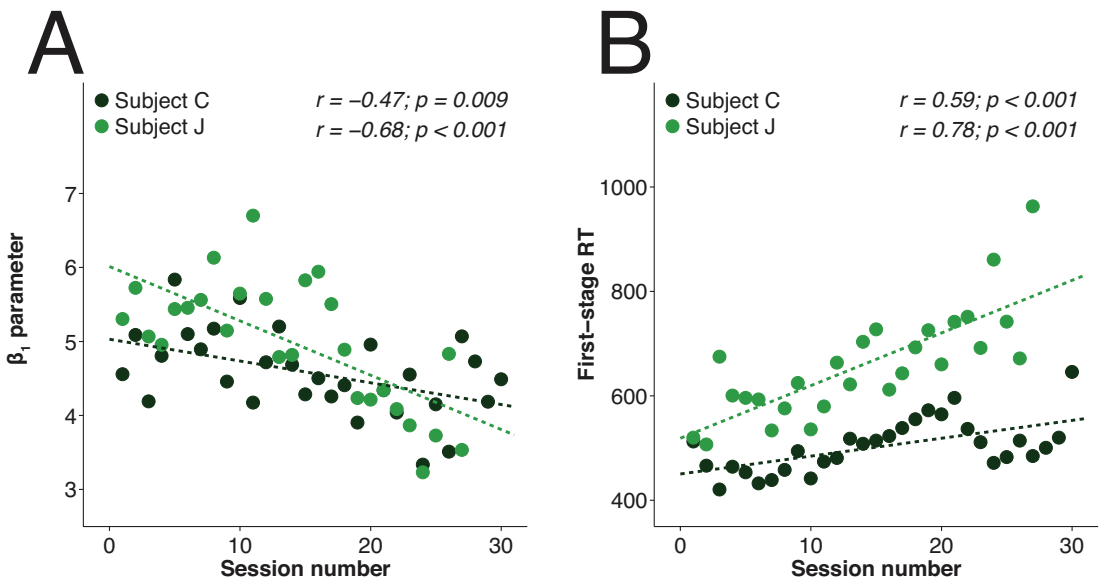


Fig. 3.11 **Evolution across sessions of the inverse temperature parameter as well as the reaction time for first-stage choice.** As the number of sessions performed increased, subjects got progressively more stochastic (smaller inverse temperature values) in their choice behaviour (**A**) and took longer to make a choice at first-stage (**B**). Dashed lines represent the regression line of the fit for each individual subject. r is the Pearson's linear correlation coefficients and p is the p-values; top values are for subject C and bottom values are for subject J.

Reaction time analysis

We have shown that choice behaviour takes into account both reward and transition information from more than just the last trial. From a motor planning perspective, adjusting one's behavioural strategy should presumably take longer than merely repeating a choice. Given the profile we observed of first-stage choice repeat probabilities as a function of reward and transition (Fig. 3.3), a similar dependency might therefore also be expected in first-stage reaction times. If present, such a response latency effect would further support the existence of MB-RL, and may even reflect prospective computation at choice time of the expected action values over the possible future states. However, it is important to note that the decision to switch or repeat first-stage choice can be made at the feedback epoch of the previous trial when the reward is revealed. For this reason, it would not be surprising if these effects are either small or absent. Based on the same line of thought, first-stage choices could also be in general faster than second-stage ones.

The first-stage choices were relatively fast, and were significantly shorter than second-stage ones (C: first-stage reaction time M = 499 ms SD = 201 versus reaction time second-stage M = 514 ms, SD = 210 with $t(31168) = -6.41$, $p < 0.001$, $d = 0.50$; J: first-stage reaction time M = 647 ms, SD = 191 versus second-stage reaction time M = 663 ms, SD = 194 with $t(29326) = -7.08$, $p < 0.001$, $g = 0.54$). Since the choice data exhibited action biases, we expected that they would also appear in reaction times (C: $F(2, 15582) = 530.96$, $p < 0.001$, $\eta^2 = 0.06$, with relative latencies of down < left < right surviving multiple comparison tests; J $F(2, 14661) = 66.67$, $p < 0.001$, $\eta^2 = 0.01$, with relative latencies of left < right < up, again surviving multiple comparison test). As a consequence, all analyses used z-scores of *log* transformed reaction time standardised according to each action for the respective session (high z-scores indicate slow responses and low z-scores denote fast reaction time).

The results in Figure 3.12A show clear differences between common and rare trials as a function of high or low outcomes received. The reaction time for first-stage choice following a high outcome obtained through a common transition was consistently faster than if it was through a rare one. On the other hand, after a low outcome the first-stage choice was faster if the previous trial had a rare transition. These findings are in agreement with the expectation that slower decisions occur in situations where the likelihood of choice switching is highest. Such situations are the ones where MB control is required, and the slower latency in response could be interpreted as the potential time costs associated with more demanding prospective computations. Of note, following a medium outcome, RTs were consistently slower with no significant differences between transition types. The differential

RT results following medium and low outcomes is somewhat surprising given the similar choice repetition profile across these outcomes (see Fig. 3.3). However, a medium outcome leads the agent to the challenging trade-off between exploration (aiming for the high reward) and exploitation (avoiding the low outcome). Therefore, this demanding computation could explain the relatively long RT.

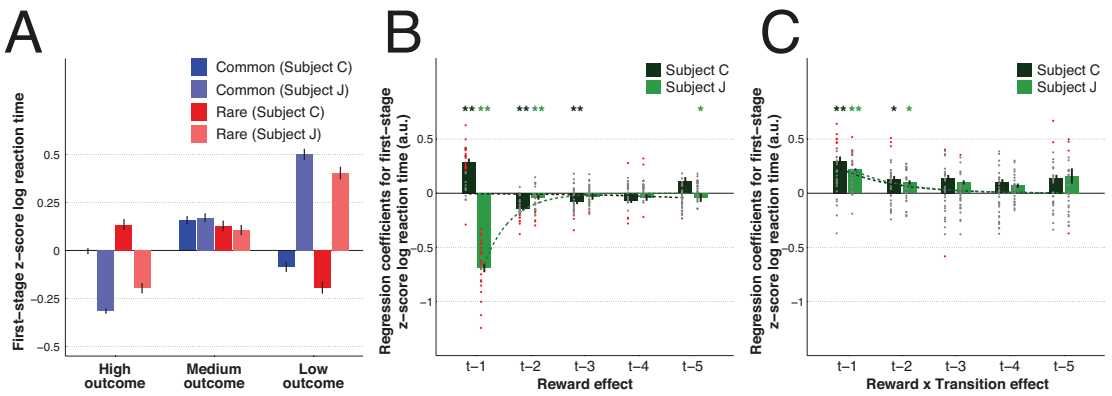


Fig. 3.12 **The impact of both reward and transition information on first-stage choice reaction time.** (**A**) The *log* transformed, z-scored, first-stage reaction time (high z-scores indicate slow responses), averaged across sessions, as a function of reward and transition type on the previous trial. Error bars depict SEM. (**B-C**) Multiple linear regression results on first-stage reaction time with the contributions of the reward (**B**) as well as the reward $\times$ transition (**C**) information from the five previous trials. Dots represent the fixed-effects coefficients for each session (coloured red when $p < 0.05$ and grey otherwise). Bar and error bar values correspond, respectively, to the mixed-effect coefficients and their SE. Dashed lines illustrate the exponential best fit on the mean fixed-effects coefficients of each trial into the past. ** $\alpha = 0.01$ and * $\alpha = 0.05$ in two-tailed one sample t-test with null-hypothesis mean equal to zero for the fixed-effects coefficients.

To complement this analysis, we performed linear regressions on reaction time for the first-stage choice using similar principles as the ones used for choice behaviour (Table 3.9 and 3.10; Fig. 3.12 B and C). The data are consistent with a significant reaction time modulation, from more than one trial into the past, by both the reward main effect (i.e., the effect of previous reward on first-stage RT; Fig. 3.12B) as well as the reward $\times$ transition effect (i.e., the effect of previous reward as well as previous transition on first-stage RT; Fig. 3.12C). Although the interaction effect was similar between the subjects, the main effect of reward on the last trial differed (positive coefficients for subject C but negative ones for subject J): a high reward (independently from the transition) made the reaction time faster on the following trial in subject J, but increased the response latency in subject C (also observed in Fig. 3.12A). To note that the main effect of the type of transition (i.e., the effect

of previous transition on first-stage RT) on the latency of future responses was small or non-existent (see effects of T predictors in Tables 3.9 and 3.10). One might have expected that encountering a rare trial, on its own, could slow down the following decisions given its surprise effect; but this was not apparent.

Table 3.9 **Fixed and mixed-effects linear regression results for the previous trial predictors of first-stage reaction time**

	Fixed-effects analysis		Random-effects analysis	
	Subject C	Subject J	Subject C	Subject J
Const	0.01 (<0.01)**	0.02 (<0.01)**	0.01 (0.01)**	0.02** (0.01)**
F_t	0.15 (0.04)**	0.38 (0.06)**	0.23 (0.03)**	0.50 (0.05)**
R_{t-1}	*0.16 (0.04)***	*-0.68 (0.05)***	*0.24 (0.03)***	*-0.71 (0.11)***
T_{t-1}	0.06 (0.02)**	0.03 (0.02)	0.06 (0.03)**	0.06 (0.02)**
$R_{t-1} \times T_{t-1}$	*0.28 (0.04)***	*0.23 (0.03)***	*0.31 (0.05)***	*0.24 (0.04)***

The dependent variable was z-scores of *log* transformed reaction time at trial t standardised according to each side of response for the respective session (high z-scores indicate slow responses and low z-scores denote fast reaction time). The variables used as predictors were: F is a linear function corresponding to the trial number; R is previous outcome level (assumed as continuous and with low=1, medium=2, high=3); and T is previous transition (rare=1, common=0). Const is the constant term. Predictors were mean centred and continuous variables were also scaled by dividing them by two standard deviations (adjustments made before the computation of the interaction terms). In bold are the predictors of interest, such as the reward main effect (reward_{t-1}) and the reward $\times$ transition effect ($\text{reward}_{t-1} \times \text{transition}_{t-1}$). Values of fixed-effects results are the mean and in between brackets the standard error of the mean of the regression coefficients across sessions; mixed-effects results are the estimated regression coefficients and in between brackets their standard error. ** for $\alpha = 0.01$ and * for $\alpha = 0.05$ in either two-tailed one sample t-test with null-hypothesis mean equal to zero for the fixed-effects results or confidence interval estimation for the mixed-effects results.

To explore motivational levels and response vigour in another way, we examined the time to the first attempt to eye fixation at first-stage. Interestingly, we found a pattern in subject C very similar to that observed in subject J's reaction time, where the reward from the previous trial made the latency to fixation shorter (Fig. 3.13A). However, there was no effect of the reward $\times$ transition interaction term, as this action does not require prospective evaluation of choices (Fig. 3.13B). Therefore, we believe that the previous trial reaction time differences could be secondary to a particular learning feature during the training protocol (despite similar protocols, subject C took significantly longer than J in the last steps of training).

Table 3.10 **Fixed and mixed-effects linear regression results for predictors of first-stage reaction time up to five trials back**

	Fixed-effects analysis		Mixed-effects analysis	
	Subject C	Subject J	Subject C	Subject J
Const	0.02 (< 0.01)**	0.02 (0.01)**	0.02 (0.01)**	0.03 (0.01)**
F_t	0.14 (0.04)**	0.41 (0.07)**	0.23 (0.03)**	0.51 (0.05)**
R_{t-1}	*0.24 (0.04)***	*-0.66 (0.05)***	*0.29 (0.03)***	*-0.69 (0.07)***
T_{t-1}	0.08 (0.02)**	0.03 (0.02)	0.09 (0.02)**	0.06 (0.02)**
$R_{t-1} \times T_{t-1}$	*0.23 (0.05)***	*0.20 (0.03)***	*0.30 (0.04)***	*0.22 (0.04)***
R_{t-2}	*-0.14 (0.02)***	*-0.09 (0.02)***	*-0.14 (0.02)***	*-0.04 (0.02)***
T_{t-2}	0.03 (0.02)	-0.03 (0.02)	0.06 (0.02)**	0.08 (0.01)**
$R_{t-2} \times T_{t-2}$	*0.09 (0.04)**	*0.07 (0.03)*	*0.13 (0.03)***	*0.10 (0.03)***
R_{t-3}	*-0.07 (0.02)***	*-0.02 (0.02)*	*-0.08 (0.02)***	*-0.03 (0.02)***
T_{t-3}	0.01 (0.02)	-0.01 (0.02)	0.05 (0.04)**	0.09 (0.02)**
$R_{t-3} \times T_{t-3}$	*0.02 (0.04)*	*-0.01 (0.03)*	*0.14 (0.03)***	*0.10 (0.03)***
R_{t-4}	*-0.01 (0.02)*	*0.04 (0.02)*	*-0.07 (0.02)***	*-0.04 (0.02)***
T_{t-4}	0.01 (0.02)	0.01 (0.02)	0.06 (0.02)**	0.03 (0.01)**
$R_{t-4} \times T_{t-4}$	*-0.02 (0.04)*	*0.04 (0.02)*	*0.10 (0.03)***	*0.07 (0.02)***
R_{t-5}	*-0.03 (0.02)*	*0.04 (0.02)**	*0.11 (0.04)***	*-0.04 (0.01)***
T_{t-5}	< 0.01 (0.02)	0.02 (0.02)	0.06 (0.03)**	0.02 (0.02)**
$R_{t-5} \times T_{t-5}$	*0.06 (0.04)*	*0.04 (0.05)*	*0.14 (0.03)***	*0.16 (0.04)***

The dependent variable was z-scores of *log* transformed reaction time at trial t standardised according to each side of response for the respective session (high z-scores indicate slow responses and low z-scores denote fast reaction time). The variables used as predictors were: F is a linear function corresponding to the trial number; R is outcome level (assumed as continuous and with low=1, medium=2, high=3); and T is transition (rare=1, common=0). Const is the constant term. Predictors were mean centred and continuous variables were also scaled by dividing them by two standard deviations (adjustments made before the computation of the interaction terms). In bold are the predictors of interest from i trials back, such as the reward main effect (reward_{t-i}) and the reward $\times$ transition effect ($\text{reward}_{t-i} \times \text{transition}_{t-i}$). Values of fixed-effects results are the mean and in between brackets the standard error of the mean of the regression coefficients across sessions; mixed-effects results are the estimated regression coefficients and in between brackets their standard error. ** for $\alpha = 0.01$ and * for $\alpha = 0.05$ in either two-tailed one sample t-test with null-hypothesis mean equal to zero for the fixed-effects results or confidence interval estimation for the mixed-effects results.

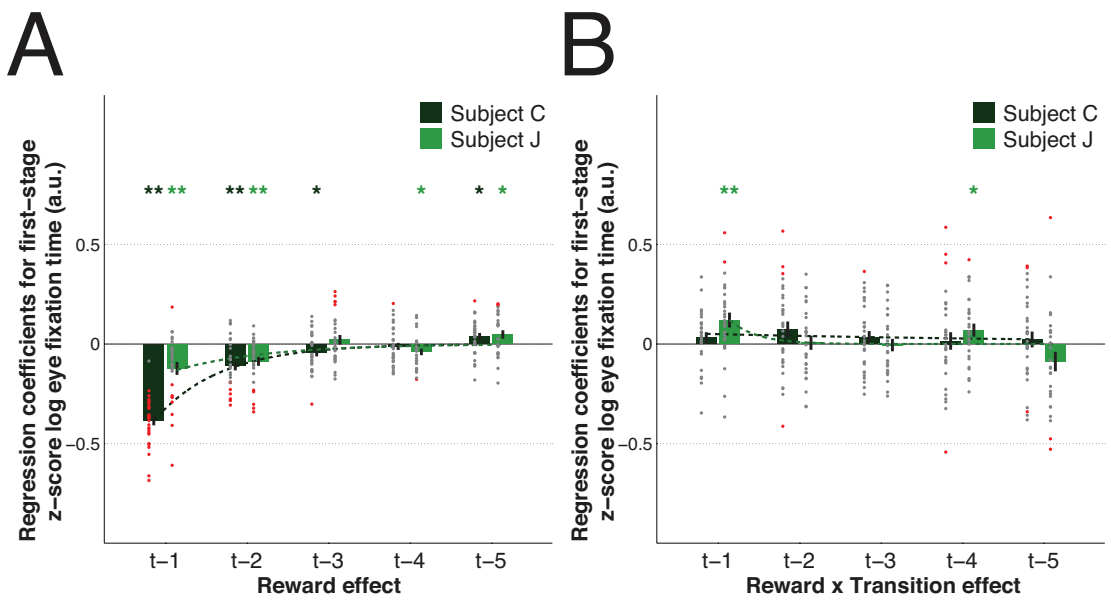


Fig. 3.13 **The impact of both reward and transition information on the first attempt to eye fixation at first-stage.** Multiple linear regression results on z-scores of *log* transformed first-stage eye fixation time (high z-scores indicate slow first eye fixation attempt) with the contributions of the reward (**A**) as well as reward $\times$ transition interaction (**B**) information from the five previous trials. Dots represent the fixed-effects coefficients for each session (coloured red when $p < 0.05$ and grey otherwise). Bar and error bar values correspond, respectively, to the mean value of the fixed-effect coefficients and its SEM. Dashed lines illustrate the exponential best fit on the mean fixed-effects coefficients of each trial into the past. ** for $\alpha = 0.01$ and * for $\alpha = 0.05$ in two-tailed one sample t-test with null-hypothesis mean equal to zero for the fixed-effects coefficients.

3.5 Discussion

Despite the extensive work in both MF-RL (or habitual) and MB-RL (or goal-directed) behaviour (Daw and Dayan, 2014; Dolan and Dayan, 2013; Doya et al., 2002), very few studies attempted to elicit simultaneous signatures of both learning strategies (Daw et al., 2011; Gershman et al., 2012; Otto et al., 2013; Wunderlich et al., 2012a). The evidence and consistency of both behavioural and computational modelling results support the idea that, like human subjects performing an equivalent decision-making task (Daw et al., 2011), non-human primates employ both MF and MB-RL strategies. In both subjects, reward history (relevant for both learning strategies) and state-transition knowledge (used in MB computations) had a significant impact on behaviour and, as also expected by the theory, such influence decayed exponentially as a function of trials into the past. This was evident in the logistic analysis and reassured the appropriateness in the use of the RL modelling.

One of the most important conclusions from the trial-by-trial computational analysis was the overall better fit of the *Hybrid* model when tested against pure MF or MB methods. This is consistent with human reports in a similar two-step task (Daw et al., 2011; Deserno et al., 2015; Wunderlich et al., 2012b). In addition to previous reports of MF-RL existence in these animals (Bayer and Glimcher, 2005; Samejima et al., 2005; Schultz et al., 1997), this study reports choice and reaction time signatures of MB-RL as formally defined by the theory. Moreover, the rigour of our model comparison procedure allowed us to not only test variants of individual classes of RL models (e.g. several variants of each learning approach), but also compare how well different classes of RL models predicted choice.

The MF $SARSA$ algorithm fit better than Q-learning. This bolsters the previous finding, in a completely different task, of neural substrates consistent with the implementation of $SARSA$ in non-human primates (Morris et al., 2006). It does, however, conflict with evidence from rodents favouring Q-learning (Roesch et al., 2007). This observation is relevant to future studies and theories, and merits further investigation.

The best MB-RL approach involved the explicit state-transition probability distributions being known from start. A manifest cause for this is the extensive exposure of the subjects to the task. Another possible consequence of the extensive task experience is the observed dominance of MB over MF behaviour as given by the high ω parameter. In fact, this much stronger MB control remarkably differed from the human studies (here ω was close to 90% versus 40-60% found in humans by Daw et al., 2011; Deserno et al., 2015; Wunderlich et al., 2012b). Even if partially instructed at start, human transition learning in an equivalent task seemed to only settle after few trials (Daw et al., 2011; Deserno et al., 2015; Wunderlich

et al., 2012b); although in other studies human subjects implemented incremental learning of transition matrices (Gläscher et al., 2010). Any working memory load potentially required for state-transition learning, as it seemed to have happened in humans, could increase reliance on a MF-RL strategy (Otto et al., 2013). By knowing the task structure well, our subjects could reduce such finite executive burden and increase MB control (Economides et al., 2015).

However, methodological and computational reasons might also explain MB dominance. The reward in the task were three different levels and our random walk design evolved such that there were sequences of a few trials with similar reward levels which alternated with discrete changes (in similar human studies reward probabilities that diffused as a random walk were used). We used the discrete reward jumps to help the subjects learn what is already a difficult task for a non-human primate. However, previous computational work has suggested that low uncertainty in estimates is a determinant factor for behavioural control, while arbitrating between the two systems (Daw et al., 2005). Therefore, the more non-stationary properties of the environment could have increased MF uncertainty, at the same time that having sequences of known outcomes may have reduced MB uncertainty. Finally, this MB-RL supremacy remained constant across session with no evidence for a significant emergence of habitual behaviour (Dickinson, 1985; Gläscher et al., 2010).

Our validation procedures assessed parallelisms between regression and computational modelling results. Hence, the procedures used quantified how well the models predicted subject's choices as well as their ability to generate the observed behavioural patterns. Although commonly used in RL studies, few studies have directly investigated the relationship between both types of analysis (Katahira, 2015). As in other behavioural studies (Corrado et al., 2005; Lau and Glimcher, 2005), this approach was fruitful in identifying discrepancies between adequate model predictive performance and generative limitations, evident, for instance, in the failure of the original *Hybrid* model (Daw et al., 2011) to replicate the stronger previous trial's influence on first-stage choice, prompting modelling refinement (*Hybrid+* model). In general, despite some disparities, both descriptive and computational approaches were mutually consistent, for instance with the expected (also by the simulated data) correlation between the reward $\times$ transition interaction coefficient and the MB weight parameter (ω).

The novel *Hybrid+* model, which fits best, closely reproduced the observed choice behaviour. Its additional parameters modelled the reinforcement (or aversion) strength of each reward level on either chosen or unchosen first-stage option, given the agent's transition knowledge. Similarly, other authors have also used and found success by adding, to MF-

RL methods, forgetting to actions not chosen and allowing aversion from a lack of reward (Ito and Doya, 2009). In fact, these latter adjustments seem to be essential for equivalence between regression and computational results (Katahira, 2015). Our new adjustment can also be regarded as some regularization procedure given potential knowledge of changes in the reward settings. This way, the added parameters could be related to other processes going beyond MF-RL but also not quite fulfilling the theoretical definition of MB-RL, as it happens in serial reversal tasks (Doll et al., 2012). To solve such tasks, some form of heuristic directly or indirectly relating the choice with the counterfactual option could be used to detect changes in the reward structure and promote adaptive behaviour. Even so, modern theoretical accounts of RL seem to incorporate these issues. Some authors have considered either multiple MB-RL systems or prior knowledge about high-level structures of behaviour as a way of augmenting more classical RL methods (Botvinick and Weinstein, 2014; Doya et al., 2002). Nevertheless, we view these extra parameters as a MF implementation of a MB effect (Akam et al., 2015) – MF, since it depends on an effect of the past trial rather than an assessment of a future one; MB, since it includes a one-step working-memory-for-state version of MB reasoning.

Importantly, though, is to acknowledge an alternative to our *Hybrid+* model as the possibility of habituation to sophisticated forms of conditional evaluation based on relationships between current state, outcome and the next action. In line with this, a less classical but alternative MF learner with time could just use second-stage state associated with reward information to choose the next first-stage choice. In fact, in our formal logistic definition of reward $\times$ transition interaction effect ($R_{t-1} \times T_{t-1} \times C_{t-1}$) as a predictor of next first-stage choice, it is embedded the second-stage state definition ($T_{t-1} \times C_{t-1}$). It is challenging to disentangle both possibilities in choice behaviour. However, this hypothesis does not explain well some choice and RT effects observed in our data. We not only observed a long temporal tail of MB effects in the choice regression analyses, but we also found that the effect of full MB learner remained after refitting all parameters with the Hybrid+ model. The RT differences between common and rare trials as a function of outcome are also not expected by such alternative strategy.

Theoretical accounts have suggested a speed accuracy trade-off between MF and MB computations, with the former being fast and the latter relatively slow (Keramati et al., 2011). It was then expected that decisions taking into account both reward and transition structure would take longer. Indeed, first-stage reaction time analysis confirmed this hypothesis, regardless of the possibility that a decision can be made on the previous trial when the outcome is known. Furthermore, this latter effect followed a similar exponential decay with

trials into the past as in choice data. Such finding does not fit well with alternative accounts for reaction time in sequential decision making that consider sequential action chunking as an explanation for faster responses (Dezfouli and Balleine, 2013). It also seems to go against other MB accounts emphasising pre-computations at the time of outcome where the re-evaluation of the utility of the states given the received reward helps future choice (Daw and Dayan, 2014; Gershman et al., 2014; Moore and Atkeson, 1993). Overall, the reaction time evidence is then supportive of a forward looking MB valuation process happening at the time of choice (Doll et al., 2015; Johnson and Redish, 2007).

A final reaction time result deserves attention, particularly given its disparity with choice data. We found a stronger influence of the main effect of reward on reaction time (subject J) and eye fixation time (both subjects) in comparison to the reward $\times$ transition effect. In other words, the influence of MF control on response vigour was stronger than MB control, and these effects were seen with a behaviour dominated by MB-RL. Others have also reported that even in over-trained scenarios response latency can still be influenced by learned action values (Wunderlich et al., 2012a). It is then appealing to relate this finding with theoretical proposals suggesting that MF control, by being the main reporter of the average rate of reinforcement, is the main mediator of the vigour of actions (Niv et al., 2007).

In conclusion, we have implemented a decision learning task in non-human primates with formal definitions of reinforcement learning MF and MB approaches and reported a detailed quantitative and computational analysis. We were able to show, to our knowledge for the first time, clear evidence of combined MF and MB-RL behaviour in those animals. Future studies focusing on the neural signals may uncover the biological substrates of these computational mechanisms.

Chapter 4

Value-based pupil responses in non-human primates performing a reinforcement learning task

4.1 Abstract

While interacting with the environment, animals learn to estimate the future values available for different possible actions, and then choose accordingly. There is evidence that multiple, partially separate, systems are involved, including model-based (MB) methods, for which estimates are prospective, based on a learned characterization of the environment and its affordances; and model-free (MF) methods, for which estimates are calculated from direct, retrospective, experience of rewards, using a temporally sophisticated form of prediction error. In a sequential decision task where both RL strategies were employed, we found that pupil dilation at particular time points during performance of the task reflected key factors associated with values and learning. At both pre- and post-choice epochs, pupil diameter reflected the expected value of the action that was to be chosen. The pupil response was best correlated with value estimates derived from a MB system; pupil diameter also reflected further task-specific features of MB calculations. Finally, when feedback was provided, pupil diameter reflected a reward prediction error signal. Overall, our data suggest that pupillary responses encode several key elements of value-based reinforcement learning processes.

4.2 Introduction

In decision-making tasks in which subjects must make a sequence of choices in distinct states of the world before gaining affectively important outcomes, prediction of future rewards plays a critical role in optimal choice. Reinforcement learning (RL) is concerned with learning to make such predictions and thus appropriate choices (see chapter 1); it provides a foundational framework for understanding psychological and neural underpinnings of decision-making (chapter 2). As already introduced in the previous chapters, there are at least two different RL methods for making predictions. Model-based (MB) valuation exploits a model of the structure of the environment, including how actions influence the (typically stochastic) transitions among different states. By contrast, model-free (MF) methods are blind to the underlying structure of the environment and instead learn by creating a temporally-sophisticated form of prediction error from sampled experience. It has been shown that the choices made by humans (Daw et al., 2011) and other animals (see chapter 3) typically exhibit characteristics of both MF and MB-RL strategies; the suggestion has been made that their estimates are integrated according to a weighted combination. However, much remains to be discovered about the systems and their combination.

Here, we consider whether pupil dilation might contribute to this evolving understanding. Changes in pupil diameter have long been used as a marker of arousal, attention, memory load, fear or novelty (Beatty, 1982; Bradshaw, 1967; Hess, 1972; Hess and Polt, 1964; Kahneman and Beatty, 1966). The pupil has been shown to dilate and/or constrict in ways that reflect these cognitive operations. Dilation is under the control of several pathways, including noradrenergic and cholinergic neuromodulation, and indeed has been used as a peripheral expression of the central activation of these substances (Aston-Jones and Cohen, 2005; Samuels and Szabadi, 2008; Sara, 2009). More recently, an effort has been made to link these pupil changes to modern theories of learning and decision-making (J.Yu and Dayan, 2005; Schultz et al., 1997). Pupil metrics have been associated with computational factors that drive learning in unpredictable or volatile environments, like uncertainty and learning rate adjustments (Nassar et al., 2012; O'Reilly et al., 2013; Preuschoff et al., 2011; Silvetti et al., 2013a). Other motivational learning factors, such as the expected reward, also seem to modulate pupil size (Gilzenrat et al., 2010; Kennerley and Wallis, 2009b; O'Doherty et al., 2003; Satterthwaite et al., 2007; Varazzani et al., 2015); but this relationship is controversial (Lavin et al., 2014; Preuschoff et al., 2011).

We therefore investigated pupil dynamics in two adult rhesus monkeys performing a value-based decision task designed to detect simultaneous signals of both MF and MB-

RL approaches (Fig. 4.1). Several aspects of the task made it particularly suitable for relating pupil diameter to some key components of reinforcement learning and decision making. First, we could assess whether pupil diameter encoded the expected value even before choices were made, and/or after those choices. Secondly, taking advantage of the two different RL computations elicited by the task, we could investigate whether dilation was more closely linked with MF or MB estimates. Finally, we could examine whether pupil diameter was modulated by previous reward history, as in the form of a prediction error. This would be expected if the pupil diameter reflected learning adjustments.

4.3 Methods

The data presented in this chapter comes from a further analysis of the data set shown in chapter 3. Therefore, all information pertaining to *Subjects and experimental apparatus*, *Task: design and timeline*, *Choice behaviour and reaction time analysis*, *Computational modelling*, *Model fitting procedures*, and *Model comparison and validation procedures* are discussed in detail in the Methods section of chapter 3. Information not particularly detailed there and relevant for the present chapter, as well as specific methodological features related to the present pupil data analysis is described.

Stimuli

All visual stimuli used were the same across sessions for both subjects, and were presented at pre-determined degrees of visual angle on a 19-inch computer screen (800×600 screen resolution and 60Hz video refresh rate) positioned 62 cm in front of the subjects eyes. A red square (0.4° in width) in the centre position was used as fixation cue. Six decision option pictures (5° in size) were chosen from the internet, reduced in size and modified through a custom-made image processing algorithm to make the average luminance equivalent for all. Similarly, the background colours used (grey, violet and brown) were measured with a SpectraScan PR650 luminance meter (Micron Techniques Ltd.) and adjusted to obtain equivalent values (one-way ANOVA: $F(2,9) = 0.66$, $p < 0.5417$; overall mean of 54.2 cd/m^2). Finally, three stimuli used as secondary reinforcers (5° in size) were generated as different spatial combinations of the same number of dark pixels in a white background, also to assure luminance equality across the three (although not equivalent to the remaining stimuli used).

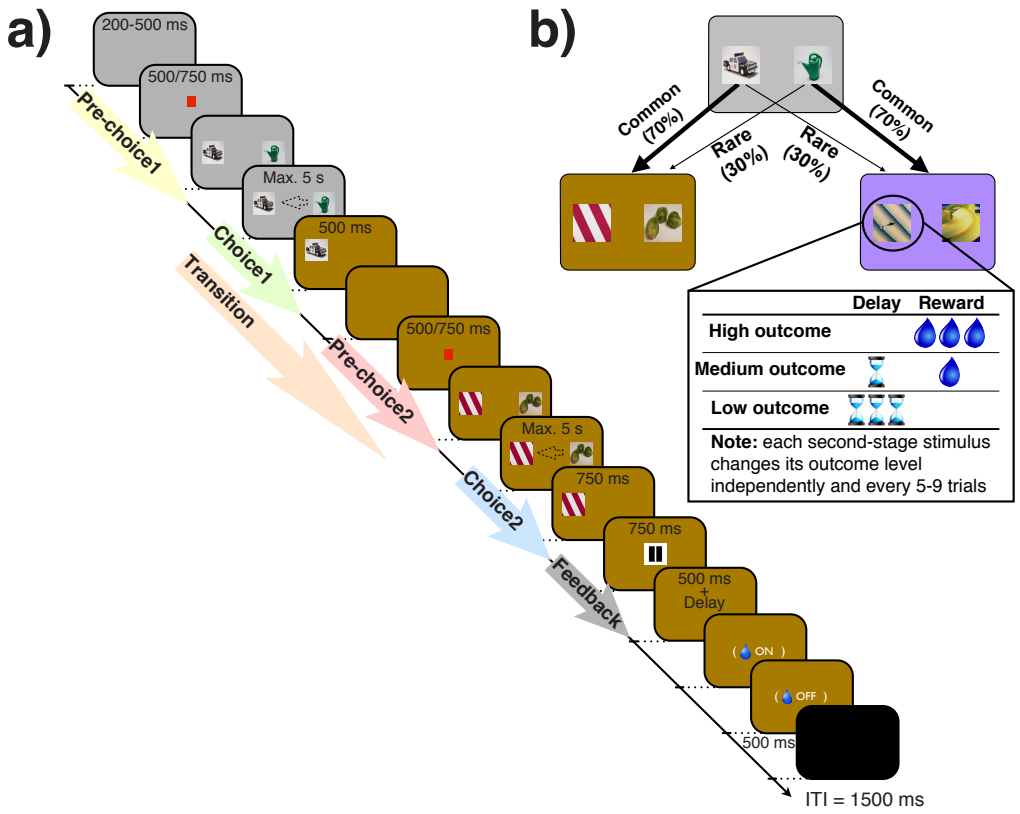


Fig. 4.1 **Two-stage decision task. (a)** Timeline of events and epochs for analysis. Eye fixation was required while a red fixation cue was shown (500/750 ms for C/J, respectively), otherwise subjects could saccade freely and indicate their decision (dotted arrow as an example) with a manual joystick movement. Once the second-stage choice had been made, the nature of the outcome was revealed by a secondary reinforcer cue (here, the pause symbol represents high outcome). Once the latter cue was removed and depending on the outcome level, there could be an additional delay followed by reward delivery. The inter-trial interval (ITI) was 1.5 s. Epochs for pupil analysis are marked along the timeline and each one is represented by colour. **(b)** The state-transition structure (kept fixed throughout the experiment). Each second-stage stimuli had an independent reward structure: the outcome level (defined by the magnitude of the reward and the delay to its delivery) remained the same for a minimum number of trials (a uniformly distributed pseudorandom integer between 5 and 9) and then, either stayed in the same level (with one-third probability) or changed randomly to one of the other two possible outcome levels.

Eye data acquisition and pupil data preprocessing.

Subjects were seated in a primate chair inside a silent and dark room with their heads fixed, and facing the computer screen. Each subject's eye position and pupil dilation was monitored with an infrared eye tracking system having a sampling rate of 240 Hz (ISCAN ETL-200). We used Monkeylogic software (http://www.monkeylogic.net/): to control the presentation of stimuli and task contingencies; to generate timestamps of behaviourally-relevant events; and to acquire as well as calibrate joystick and eye data (1000 Hz of analog data acquisition).

Blinks were identified on the basis of pupil diameter as well as eye position and removed through linear interpolation of values before and after each identified blink (interpolation time window: from 30 ms before until 30 ms after each blink) using a custom-built code. Blink-filtered pupil data were then low-pass filtered by applying a second-order Butterworth filter with a normalized cut-off frequency of 3.75 Hz. To eliminate small day-by-day fluctuations in tonic pupil diameter or position of eye recording equipment, pupil data were z-scored in each session using the across-session overall mean and standard deviation.

Our pupil analyses focused on six different task epochs (each one coloured in Fig. 4.1a):

1. **PreChoice1 epoch:** from 900 ms before until 100 ms after first-stage pictures presentation;

2. **Choice1 epoch:** from 100 ms before until 700 ms after first-stage choice;

3. **Transition epoch:** from 100 ms before until 1500 ms after background colour changed from first-stage (grey) to the respective state in second-stage (brown or violet);

4. **PreChoice2 epoch:** from 900 ms before until 100 ms after second-stage pictures presentation; 5) Choice2: from 100 ms before until 700 ms after second-stage choice;

5. **Choice2 epoch**: from 100 ms before until 700 ms after second-stage choice;

6. **Feedback epoch:** from 100 ms before until 700 ms after the secondary reinforcer picture presentation (i.e. when the upcoming outcome level is revealed).

It is important to note that PreChoice1 is not adjoining Choice1 and PreChoice2 is not adjoining Choice2, with some time missing between each.The reason for this was because during the actual decision period subjects saccade to see the stimuli location and this has an impact on pupil metrics (due to changes in light between background and stimuli).

During the task, in the absence of a fixation cue, the animal was free to look around. If gaze strayed from the screen (defined as x or y positions $> 15°$; nearest integer values of the mean of both width and height screen limits), light levels would decrease, leading to a marked increase in pupil size (Fig. 4.2). We therefore excluded trials in which subjects looked off the screen for more than 50 ms during the analysed epochs (we initially approximated this value by inspecting the raw data, but then formally tested it; see Fig. 4.3). Figure 4.4 shows the proportions and timings of the trials excluded. Table 4.1 displays the trial conditions of the excluded trials for the transition (proportion of trials excluded of 0.09 ± 0.03 for subject C and 0.39 ± 0.10 for subject J) and feedback (proportion of trials excluded of 0.73 ± 0.08 for C and 0.14 ± 0.06 for J) epochs, the ones with more off-screen gazes.

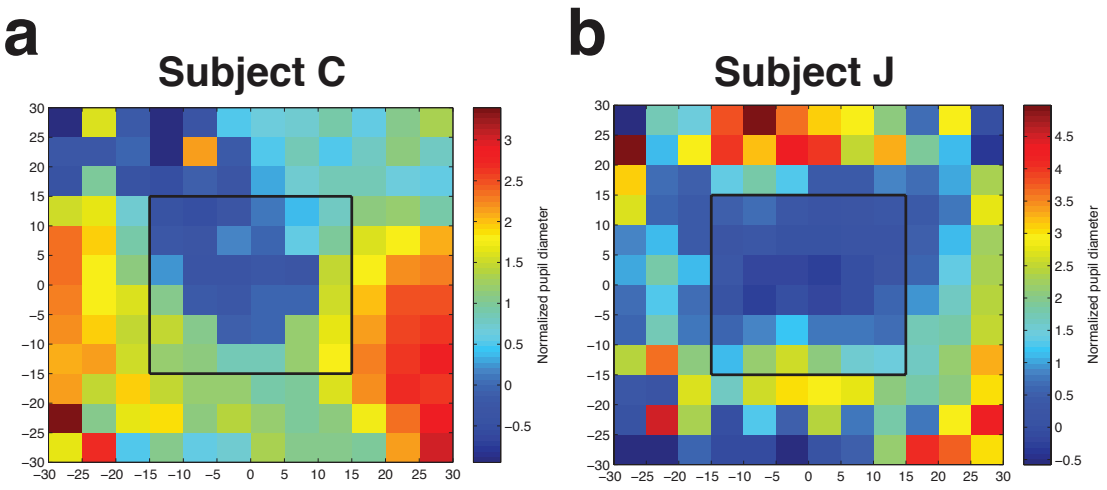


Fig. 4.2 **Relationship between eye position (expressed in visual degrees away from zero) and subsequent pupil size (z-scored based on within-session mean and standard deviation) during all trials (from beginning to end of each trial) from all sessions.** Off-screen positions led to higher subsequent pupil size. For every recorded eye position, the average pupil size associated with the subsequent 250 milliseconds was stored as a value in the square belonging to that particular eye position. Each square spanned 5 visual degrees along both X and Y axes and the analysis covered eye positions ranging from -30 to +30 visual degrees along both X and Y axes such that 144 squares (12×12) were obtained. The black lines mark the off-screen limits we defined. The displayed color of each square is based on the mean z-scored pupil diameter value obtained for that particular square. a) Subject C; b) Subject J.

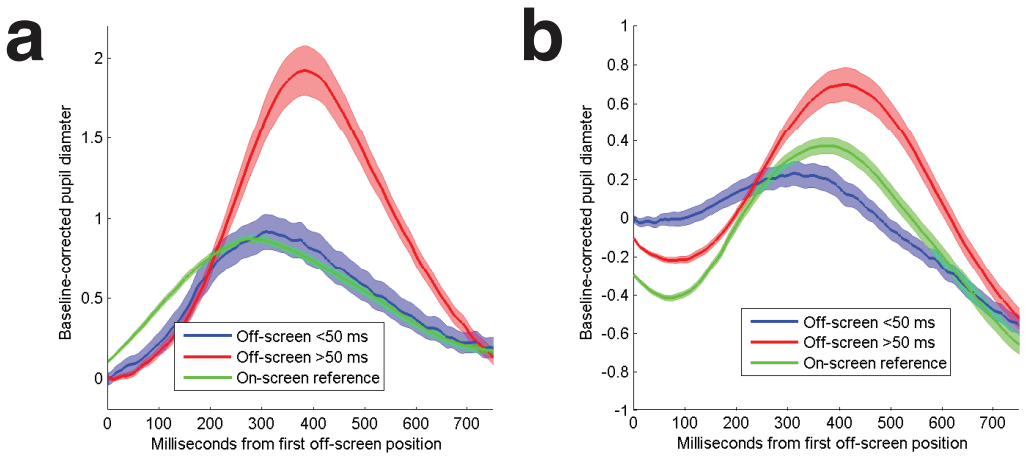


Fig. 4.3 **Pupil dilation when gaze goes off-screen.** The pupil dilated significantly more for off-screen gazes lasting longer than 50 ms compared with an on-screen reference; whereas no significant difference can be observed from the reference when the gaze was off-screen for less than 50 ms. This threshold time was based on a careful initial exploratory inspection of the data. The procedure to create this figure was as follows: for each off-screen trial, the adjusted pupil dilation for 750 ms after the first moment when the gaze went off the screen was recorded, with the adjustment being the subtraction of the mean of the dilation for 150 ms before this moment. A similar procedure was applied for on-screen trials, but the starting point of the analysis was the average of the first off-screen time across sessions. Pupil responses for the respective conditions were first averaged within session and then averaged across sessions. Shaded areas display s.e.m.. a) Subject C; b) Subject J.

Table 4.1 **Relevant trial conditions for the two epochs where off-screen gazes were most frequently observed**

		Proportion of excluded trials	
Epoch	Trial condition	Subject C	Subject J
Transition			
	High outcome, common transition	0.11 (0.05)	0.55 (0.13)
	High outcome, rare transition	0.03 (0.04)	0.31 (0.11)
	Medium outcome, common transition	0.09 (0.05)	0.36 (0.12)
	Medium outcome, rare transition	0.03 (0.03)	0.33 (0.12)
	Low outcome, common transition	0.11 (0.05)	0.33 (0.10)
	Low outcome, rare transition	0.05 (0.04)	0.27 (0.11)
Feedback			
	High outcome	0.81 (0.10)	0.12 (0.07)
	Medium outcome	0.71 (0.12)	0.20 (0.10)
	Low outcome	0.29 (0.11)	0.15 (0.08)

Results are mean (s.e.m) values across sessions.

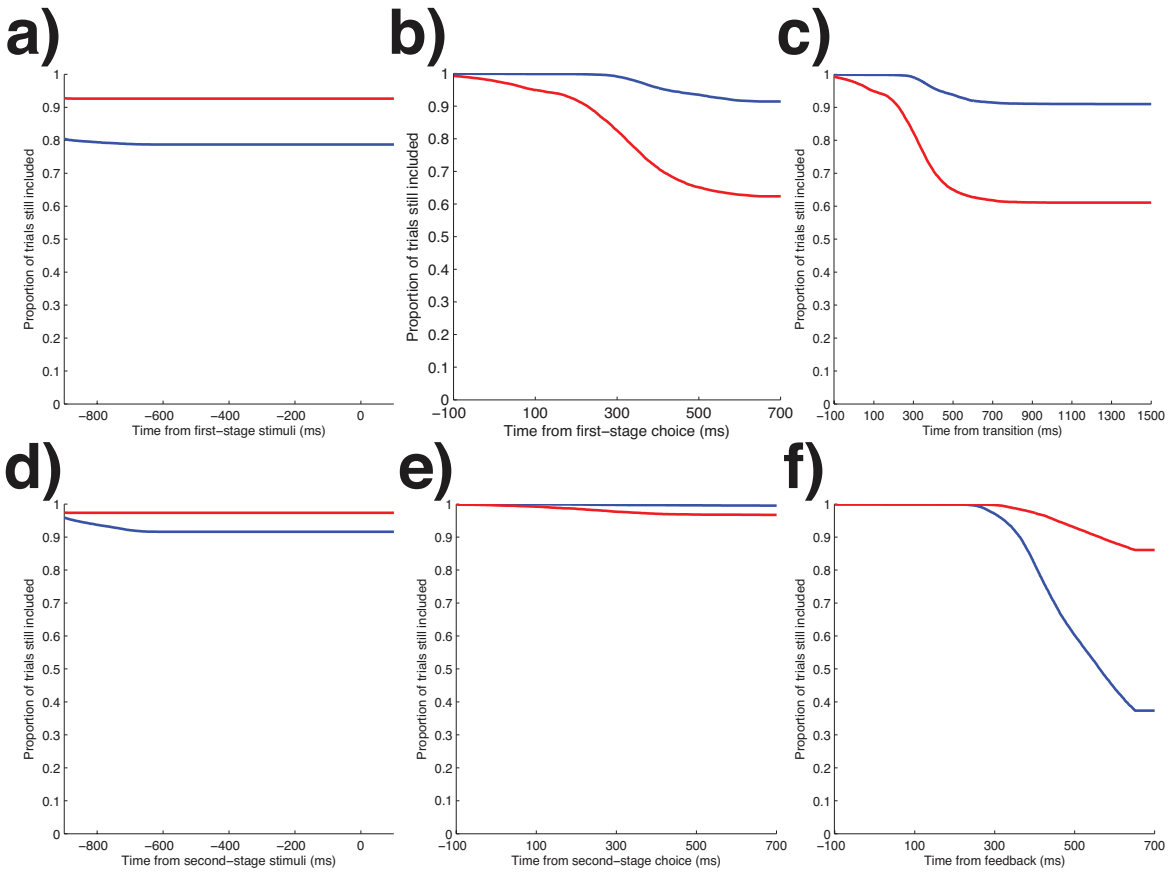


Fig. 4.4 **Timings of the first off-screen gazes for each analysed epoch.** Proportion of trials still included in the analysis for: **(a)** pre-choice1 epoch; **(b)** choice1 epoch; **(c)** transition epoch; **(d)** pre-choice2 epoch; **(e)** choice2 epoch; and **(f)** feedback epoch. Subject C in blue and subject J in red.

Pupil diameter analyses

Several multiple linear regression models were performed separately for each subject and session to quantify predictors of pupil change at different epochs. For graphical purposes they were fit as sliding regression analysis to the mean of z-scored pupil diameter during the preceding 100 ms time-window, and then shifted in 1 ms steps until we had analysed the entire epoch. We then displayed the across sessions mean (and s.e.m.) of the resultant time-series of regression coefficients (β) for the regressors of interest. When assessing statistical significance of the predictors from each regression, 200 ms non-overlapping time-windows were used instead and the same model applied. Then, one-sample t-tests with Bonferroni multiple comparison correction were performed for the null hypothesis that the mean of the resultant regression coefficients differed significantly from zero.

Two types of predictors were used. One set comprised observable behavioural variables: ternary outcome level R or $RewPic2_t$ (low=1, medium=2, high=3); binary T transition (rare=1, common=0); $C1$ and $C2$ are first- and second-stage choices, respectively. The other set of variables was derived from the best-fitting computational modelling (see chapter 3), and comprised: $Chosen1Qhyb$ and $Unchosen1Qhyb$, which are the $Hybrid$+ Q-values of the chosen and unchosen first-stage options, respectively; $Chosen1Qmf$ and $Unchosen1Qmf$, which are the MF-RL ($SARSA$ algorithm) Q-values of the chosen and unchosen first-stage options, respectively; $Chosen1Qmb$ and $Unchosen1Qmb$, which are the MB-RL ($Forward_1$ algorithm) Q-values of the chosen and unchosen first-stage options, respectively; $Chosen2Q$ and $Unchosen2Q$, which are the (MB and MF) Q-values of the chosen and unchosen second-stage options, respectively; $RPE2$ is the reward prediction error at second-stage. Due to the high linear correlation between the $Chosen1Qhyb$ and $Unchosen1Qhyb$ (mean r across sessions = 0.55/0.43 for C/J, all sessions with $p < 0.05$), we orthogonalised $Unchosen1Qhyb$ with respect to $Chosen1Qhyb$ for regression analyses in which both variables were predictors. For similar reasons, we orthogonalised $Unchosen1Qmb$ with respect to $Chosen1Qmb$ (mean r across sessions = 0.76/0.75 for C/J, all sessions with $p < 0.05$). All these variables, when used as predictors in regression analysis, were mean centered and continuous (and ternary) variables were also scaled by dividing them by two standard deviations so that the magnitudes of regression coefficients could be directly compared (Gelman, 2008). These adjustments were performed before the computation of the interaction terms. The error terms of the regression models are represented by ε.

We built ten regression models. The first set tested the possibility that z-scored pupil diameter ($Pupil$) at trial t encoded upcoming expected value at pre-choice1 (Eq. 4.1) and

at pre-choice2 (Eq. 4.2). Forced first- and second-stage choice trials were excluded from these two previous regression analyses, because animals did not know before options were presented whether the choice would be a forced one. We examined pupil modulation by expected value at the time of choice with two similar regressions applied at choice1 epoch (Eq. 4.3) and choice2 epoch (Eq. 4.4).

$$\begin{aligned} Pupil_t^{\text{Prechoice1}} =& \beta_R R_{t-1} + \beta_T T_{t-1} + \beta_{RT} R_{t-1} \times T_{t-1} + \\ & \beta_{Chosen1Qhyb} Chosen1Qhyb_t + \beta_{Unchosen1Qhyb} Unchosen1Qhyb_t + \varepsilon_t \end{aligned} \tag{4.1}$$

$$\begin{aligned} Pupil_t^{\text{Prechoice2}} =& \beta_R R_{t-1} + \beta_T T_t + \beta_{RT} R_{t-1} \times T_t + \\ & \beta_{Chosen2Q} Chosen2Q_t + \beta_{Unchosen2Q} Unchosen2Q_t + \varepsilon_t \end{aligned} \tag{4.2}$$

$$\begin{aligned} Pupil_t^{\text{Choice1}} =& \beta_R R_{t-1} + \beta_T T_t + \beta_{RT} R_{t-1} \times T_t + \\ & \beta_{Chosen1Qhyb} Chosen1Qhyb_t + \beta_{Unchosen1Qhyb} Unchosen1Qhyb_t + \varepsilon_t \end{aligned} \tag{4.3}$$

$$\begin{aligned} Pupil_t^{\text{Choice2}} =& \beta_R R_{t-1} + \beta_T T_t + \beta_{RT} R_{t-1} \times T_t + \\ & \beta_{Chosen2Q} Chosen2Q_t + \beta_{Unchosen2Q} Unchosen2Q_t + \varepsilon_t \end{aligned} \tag{4.4}$$

Next, two complementary multiple regressions were conducted to investigate whether pupil changes reflected value estimates from MF-RL, MB-RL, or a combination of both. In the first model, we aimed to directly compare MF and MB-RL expected values at prechoice1 epoch (Eq. 4.5). As we will see, this showed a preponderant influence of MB values in predicting the pupil response. Therefore, in a second analysis, we tested the null hypothesis that expected value coding in pupil was purely explained by this learning strategy. For this, we created a regressor, $DiffChosenQmbQmf$, defined at the same time points as the $ChosenQmb$, using the partial derivative with respect to the MB/MF weighting parameter ω that captures how $ChosenQmb$ would change if it had been computed according to a different value of ω (i.e. $\omega = 0$ for MF-RL). In this case, it is just the difference between $ChosenQmb$ and $ChosenQmf$, and it characterises how pupil response would differ if it were correlated with $ChosenQmf$ or any weighted mixture of both learning strategies. This approach was adopted to reduce the correlation between the regressors of interest ($ChosenQmb$ and $ChosenQmf$). We then run the regression model (Eq. 4.6) and, if the null hypothesis was true, then the effect of $DiffChosenQmbQmf$ on pupil dilation would not differ significantly from zero (one sample two tailed t-test with Bonferroni correction for multiple comparisons). For the same reasons as previously described, forced first-stage choice trials were again not included in these two regressions.

$$\begin{aligned} Pupil_t^{\text{Prechoice1}} =& \beta_R R_{t-1} + \beta_T T_{t-1} + \beta_{RT} R_{t-1} \times T_{t-1} + \\ & \beta_{Chosen1Qmf} Chosen1Qmf_t + \beta_{Unchosen1Qmf} Unchosen1Qmf_t + \\ & \beta_{Chosen1Qmb} Chosen1Qmb_t + \beta_{Unchosen1Qmb} Unchosen1Qmb_t + \varepsilon_t \end{aligned} \tag{4.5}$$

$$\begin{aligned} Pupil_t^{\text{Prechoice1}} =& \beta_R R_{t-1} + \beta_T T_{t-1} + \beta_{RT} R_{t-1} \times T_{t-1} + \\ & + \beta_{Chosen1Qmb} Chosen1Qmb_t + \beta_{DiffChosenQmbQmf} DiffChosenQmbQmf_t \end{aligned} \tag{4.6}$$

The next step was to confirm whether pupil changes reflected knowledge about the state-transition structure, as well as the interaction between this and first-stage chosen and unchosen values at transition epoch (Eq. 4.7).

$$\begin{aligned} Pupil_t^{\text{Transition}} =& \beta_T T_t + \\ & \beta_{Chosen1Qmb} Chosen1Qmb_t + \beta_{Unchosen1Qmb} Unchosen1Qmb_t + \\ & \beta_{Chosen1Qmb \times T} Chosen1Qmb_t \times T_t + \\ & \beta_{Unchosen1Qmb \times T} Unchosen1Qmb_t \times T_t + \varepsilon_t \end{aligned} \tag{4.7}$$

As a non-parametric test of the effect of learning about rewards, we regressed pupil diameter at feedback against the upcoming reward, and also the previous reward history of the current second-stage choice. At trial t, a given second-stage choice $C2_t$ will have its $RewPic2_t$ or upcoming outcome level (same as R variable but renamed here for consistency with the other predictors), its most-recent, $Lag1RewPic2_t$, second-most recent, $Lag2RewPic2_t$, and third-most recent, $Lag3RewPic2_t$, rewards obtained with that choice (zero was assumed for lag variables that were associated with no observation). This regression can be described as (Eq. 4.8).

$$\begin{aligned} Pupil_t^{\text{Feedback}} =& \beta_{RewPic2} RewPic2_t + \beta_{Lag1RewPic2} Lag1RewPic2_t + \beta_{Lag2RewPic2} Lag2RewPic2_t \\ & \beta_{Lag3RewPic2} Lag3RewPic2_t + \beta_T T_t + \varepsilon_t \end{aligned} \tag{4.8}$$

The prediction error at this point in a trial would be the difference between the current reward and a weighted sum of previous rewards. As a non-parametric test of this, we examined the values of the $\beta_{RewPic2}$, $\beta_{Lag1RewPic2}$, $\beta_{Lag2RewPic2}$ and $\beta_{Lag3RewPic2}$ obtained from this

regression model (Eq. 4.8). Some of these predictors were different from zero before the moment the upcoming reward was known to the subject, and since we were interested in the change consequent on the new information from the feedback signal, we baseline corrected each predictor β time-series (by subtracting the mean β values during the first 200 ms of the epoch). The last step of this analysis, was to calculate the mean values (and s.e.m) across sessions of these baseline corrected β coefficients from the moment the upcoming reward coding in the pupil started (approximated, by inspection of the results, to 400ms after the secondary reinforcer onset in both subjects) until the end of the feedback epoch.

A more parametric examination of the prediction error comes from looking at the relationship between pupil diameter and the two components R and $Chosen2Q$ which are subtracted to form $RPE2$ (Eq. 4.9). A pattern expected for a reward prediction error signal at feedback epoch would be: a positive correlation between pupil diameter and R; and negative correlation between pupil diameter and $Chosen2Q$. A contrast of parameter estimates was also used to examine the influence on pupil size of the difference between the expected value of the second-stage choice and the value of the outcome that was actually received (that is, [1 -1] contrast on R and $Chosen2Q$ values, i.e., testing the difference between both variables regression coefficients).

$$Pupil_t^{\text{Feedback}} = \beta_R R_t + \beta_{Chosen2Q} Chosen2Q_t + \beta_T T_t + \varepsilon_t \tag{4.9}$$

Finally, we also correlated pupil dynamics at feedback epoch directly with the $RPE2$ derived by the computational model (Eq. 4.10).

$$Pupil_t = \beta_{RPE2} RPE2_t + \beta_T T_t + \varepsilon_t \tag{4.10}$$

Importantly, for each of the regressions, we considered an additional model which also included two further control predictors: $xpos_t$ and $ypos_t$, corresponding to the average absolute x and y eye position at trial t for the analysed time-window, respectively. This acted as a control for the potential confounding effect of eye position, in addition to excluding trials with off-scree gaze. The results of the original model were reported, unless significant differences were noted with this controlling measure.

4.4 Results

Choice behaviour and computational modelling

Subjects solved a two-stage decision task (Fig. 4.1). The task and the analysis of the behaviour are described in detail in chapter 3; we repeat just the details necessary for characterizing the pupil response. In the task, two sequential choices between isoluminant stimulus pairs had to be made before subjects received a reward. A grey background represented the first-stage state and each choice was between two options presented as pictures. Each of these first-stage choices could lead to either a common (70% transition probability) or rare (30% transition probability) second-stage state, represented by different isoluminant background colours (brown and violet). In the second-stage, another two-option choice between pictures was required, and it was reinforced according to one of three different levels of reward (high, medium or low outcomes). Subjects then returned to the first-stage choice following an inter-trial interval. Importantly, to encourage learning, the outcome level of each second-stage option remained constant for five to nine trials and subsequently switched with a probability of two-thirds to one of the other two possible outcomes.

As analysed in detail in chapter 3, subjects exhibited a hybrid mixture of MB and MF control. MB control was relatively dominant, but not to the complete exclusion of MF influence, with the mean value of the parameter MF= $0 \leq \omega \leq 1$ =MB governing the relative weighting being near 0.9. Here, we used the model to quantify the various components of choice, such as expected values associated with the MB and MF components of the best-fitting hybrid accounts and outcome prediction errors, and regressed these against changes in pupil diameter.

Expected value coding in the pupil

A key characteristic of the task is that subjects can start planning aspects of their first- and second-stage decisions before the options were actually presented (albeit not the actual motor actions required). For the first-stage choice, the decision to switch or repeat the previous selection could be made from the moment the feedback at the previous trial was provided. For second-stage choice, subjects could start planning their decision as soon as the second-stage state was known. Thus, given evidence that pupil diameter encodes features before (de Gee et al., 2014; Fiedler and Glöckner, 2012) as well as after (Einhauser et al., 2010; Einhäuser et al., 2008) choices are made, we used multiple linear regression to assess the extent to which trial-wise estimates of first-stage (Fig. 4.5a-b; see Eq. 4.1) as well as

second-stage (Fig. 4.5c-d; see Eq. 4.2) upcoming chosen and unchosen action-values (*Q*-values), derived from the best fitting *Hybrid+* model, predicted pupil response at epochs pre-choice1 and pre-choice2, respectively (Fig. 4.1 to see epochs along the timeline). In this regression, we controlled for the most recent reward, the most recent transition and the interaction between the two. We found that the value of the choice about to be made, at both decision stages, significantly increased pupil diameter before the options were presented. The value of the option not chosen was also associated with significant pupil dilation, but to a much lesser degree (Fig. 4.5).

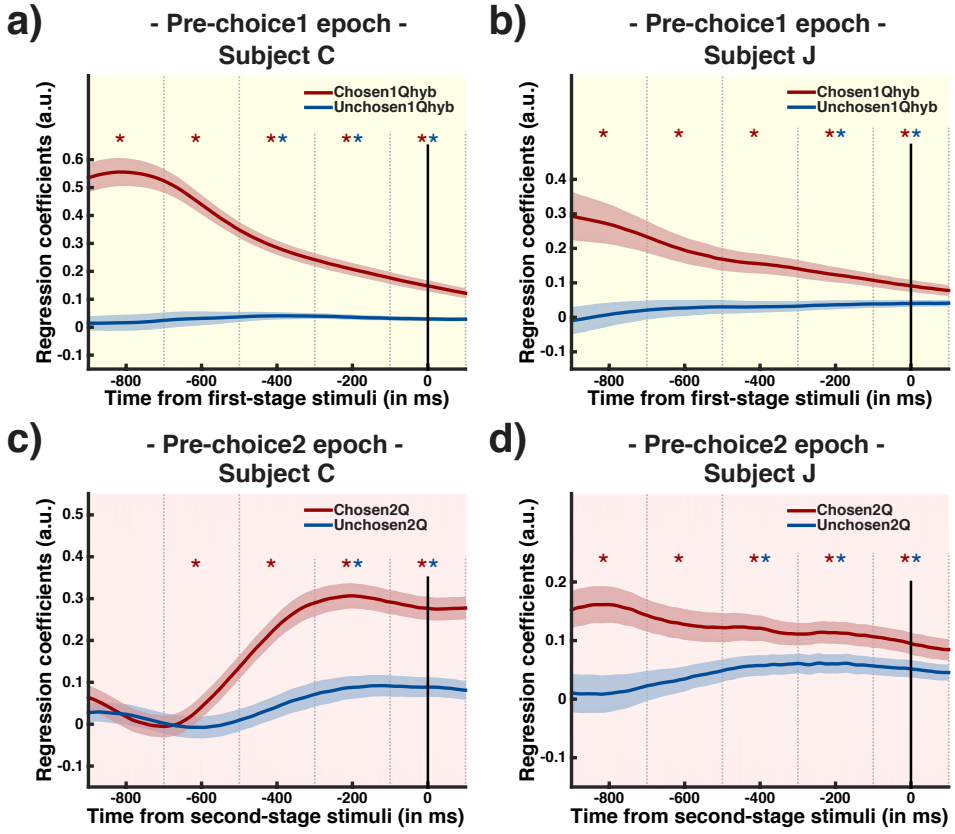


Fig. 4.5 **Pupil diameter encodes expected value of upcoming choices.** Time courses of the mean across sessions (shaded regions represent s.e.m.) for the effects of first-stage **(a-b)** and second-stage **(c-d)** chosen as well as unchosen *Q*-values derived by the *Hybrid+* model regressed on z-scored pupil size at pre-choice1 and pre-choice2 epochs, respectively (controlling for recent reward, recent transition and the interaction between the two; see Eq. 4.1 and Eq. 4.2 in Methods for details). Asterisks indicate 200 ms time bins in which the coefficients differed significantly from zero (two tailed t-test with Bonferroni correction for multiple comparisons).

We then used a similar strategy to examine the relationship between pupil diameter and expected values at the time of first-stage (Fig. 4.6a-b; see Eq. 4.3) and second-stage (Fig.

4.6c-d; see Eq. 4.4) choices. It is important to note that pre-choice (Fig. 4.5) and choice (Fig. 4.6) epochs are non-contiguous, as they are separated by times at which the subjects made saccades to look at the pictures associated with the choices. We observed that the pupil response started to encode the expected chosen value for first- and second-stage positively shortly after the choice was made, and this remained relatively constant throughout the remaining part of the epoch. In regard to the value of the unchosen option, there was a tendency for an initial weak positive effect to be lost after the decision was made. Overall, these findings suggest that pupil size reflected the expected value of the choice, both before and after it is made.

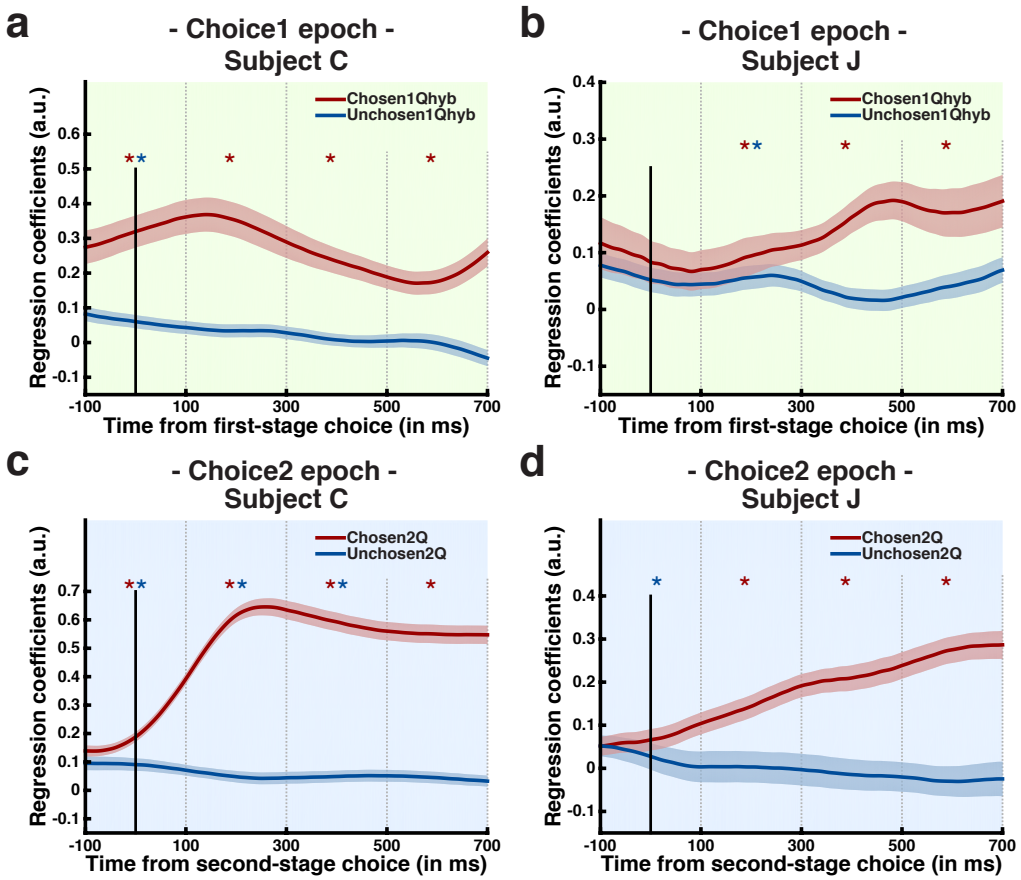


Fig. 4.6 **Pupil diameter encodes expected chosen value at choice time.** Time courses of the mean across sessions (shaded regions represent s.e.m.) for the effects of first-stage **(a-b)** and second-stage **(c-d)** chosen as well as unchosen Q-values derived by the *Hybrid+* model regressed on z-scored pupil size at choice1 and choice2 epochs, respectively (controlling for most recent reward, most recent transition and the interaction of both; see Eq. 4.3 and Eq. 4.4 in Methods for details). Asterisks indicate 200 ms time bins in which the coefficients differed significantly from zero (two tailed t-test with Bonferroni correction for multiple comparisons).

Pupil and model-based reinforcement learning

We next investigated whether the expected value coding in pupil dilation was related to both MF and MB values. These only differ for the first stage choice, so we confined our analysis to the pre-choice1 epoch. We eschewed the choice1 epoch because after the first-stage choice, the signal could be contaminated by the upcoming second-stage chosen value. We constructed two regression models. The first included as predictors of interest chosen MF and MB Q-values, as well as unchosen MF and MB Q-values (controlling for the previous reward, the previous transition and the interaction between the two; see Eq. 4.5). The results of this analysis showed a strong significant positive effect of chosen MB Q-values on pupil size, but no correlation with unchosen or MF Q-values (Fig. 4.7). This finding suggests that the previously observed expected value coding in the pupil (Fig. 4.5 and Fig. 4.6), predominantly reflected MB rather than MF value computations.

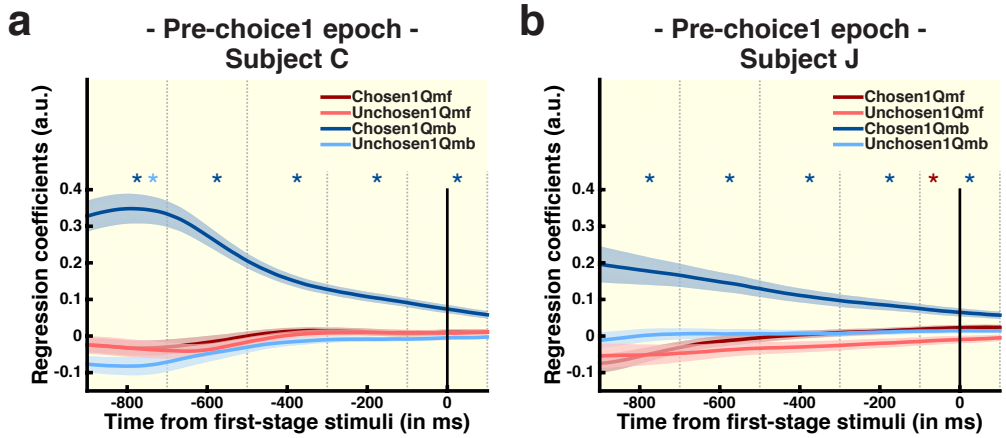


Fig. 4.7 **The relationship between pupil response and model-free and model-based value estimates.** Time courses of the mean across sessions (shaded regions represent s.e.m.) for the effects of trial-wise estimates of first-stage chosen/unchosen model-free (Chosen1Qmf/Unchosen1Qmf) as well as chosen/unchosen model-based (Chosen1Qmb/Unchosen1Qmb) Q-values derived by the *Hybrid+* model regressed on z-scored pupil size at pre-choice1 epoch (controlling for previous reward, previous transition and the interaction between the two; see Eq. 4.5 in Methods for details). Asterisks indicate 200 ms time bins in which the coefficients differed significantly from zero (two tailed t-test with Bonferroni correction for multiple comparisons).

To examine this observation further, we built a second regression model comprised by two regressors of interest: the chosen MB Q-values and the difference between the chosen MF and MB Q-values (controlling for also previous reward, current transition and the interaction of these two variables; see Eq. 4.6). We used this second approach in addition to the

previous model (where valuations of both learning strategies were included) to make sure potential multicolinearity issues due to correlations between MF and MB value estimates are not interfering with our regression estimates. Furthermore, it is also a way of testing the null hypothesis that the previously seen chosen value coding in the pupil was purely MB. If the null hypothesis was true, then the pupil signal would be accounted for entirely by the trial-by-trial chosen MB Q-values, and the difference regressor should not be significantly positive or negative. By contrast, if the null hypothesis is falsified, it means that the pupil response was better characterized, on average, by also having some MF valuation (i.e., a mixed MF and MB pupil signal). We found that the modulation in pupil diameter of both subjects was, overall, not better characterised when MF valuation was also taken into account (non-significant effect of the difference regressor coefficient on Bonferroni-corrected one-sample t-tests for the entire epoch in subject C and from 500 ms before to 100ms before first-stage stimuli presentation in subject J; see Fig. 4.8).

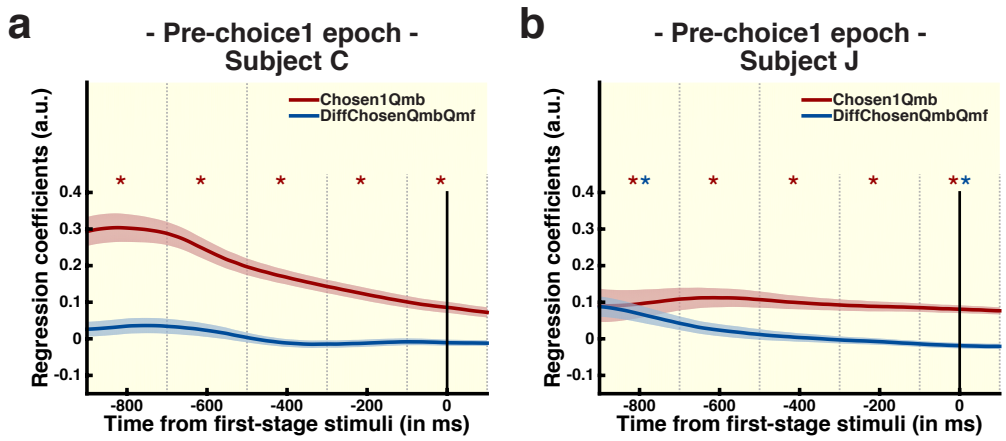


Fig. 4.8 **The expected value coding in pupil size is explained by pure model-based predictions.** Time courses of the mean across sessions (shaded regions represent s.e.m.) for the effect, on pupil size at pre-choice1 epoch, of trial-wise estimates of first-stage chosen model-based Q-values (Chosen1Qmb) as well as a difference regressor (DiffChosenQmbQmf) derived by subtracting to Chosen1Qmb the first-stage chosen model-free values (see Eq. 4.6 in Methods for details). Asterisks indicate 200 ms time bins in which the coefficients differed significantly from zero (two tailed t-test with Bonferroni correction for multiple comparisons).

These regressions suggest that pupil diameter reflected valuations computed by pure MB-RL methods. To investigate this hypothesis further, we related pupil metrics with other features linked to MB computations. A distinctive signature of MB-RL is the combined effect of knowledge about state transitions and observations of changes to rewards. We therefore compared pupil dynamics during the transition epoch as a function of the tran-

sition type on the current trial (Fig. 4.9a-b). In both subjects, pupil size increased to a significantly greater degree following a common than a rare transition. Given that the pupil encoded the MB value of the chosen option, one could interpret this pupil dilation following a common transition as a high-arousal state, because it is an opportunity to pursue the expected chosen outcome (given the first-stage choice pattern shown by both subjects, the expected outcome corresponds to the highest reward possible). Similarly, when the vicissitude of a rare transition interfered with the reward expectation (i.e. the high reward) one could expect some disappointment or task disengagement, and hence a smaller increase in pupil size (Chiew and Braver, 2013; Hess and Polt, 1964).

To assess the possibility that the modulation of the pupil diameter according to transition type also took into account the expected values, we built a regression model in which pupil dilation after the transition could be accounted for by the following explanatory variables: transition (coded rare=1 and common=0), chosen and unchosen first-stage MB Q-values, and the interaction terms, namely first-stage chosen MB Q-values $\times$ transition and first-stage unchosen MB Q-values $\times$ transition (see Eq. 4.7). The last two are the regressors of interest: the higher the value of the first-stage choice the more disappointment a rare transition would cause, and the less pupil dilation would be expected (i.e., a negative chosen MB Q-value $\times$ transition effect is observed); on the other hand, a high-arousal state and substantial pupil dilation would be expected if the value of the unchosen first-stage option is high (such as when errors in first-stage choices are made) when a rare transition occurs (i.e., a positive unchosen MB Q-value $\times$ transition effect occurs), as the subject still has an opportunity to obtain a valuable outcome. Consistent with our hypothesis, we found in both subjects a significant negative chosen MB Q-value $\times$ transition effect, as well as a significant positive unchosen MB Q-value $\times$ transition effect once the transition was revealed (Fig. 4.9c-d). Furthermore, it is important to note that the main effect of transition remained significant even after adding the MB-RL Q-values and corresponding interaction terms.

As a whole, the above evidence is consistent with the pupil size reflecting reward $\times$ transition knowledge, or MB-RL correlates, as well as information about the state-transition consequences in the expected value.

Pupil encoded a reward prediction error at feedback

The final epoch of critical interest is after the subject knows what level of reward will be provided – indicated by the secondary reinforcer picture. The pupil started to increase in diameter 300-400ms after this picture was presented (Fig. 4.10). When we compared pupil

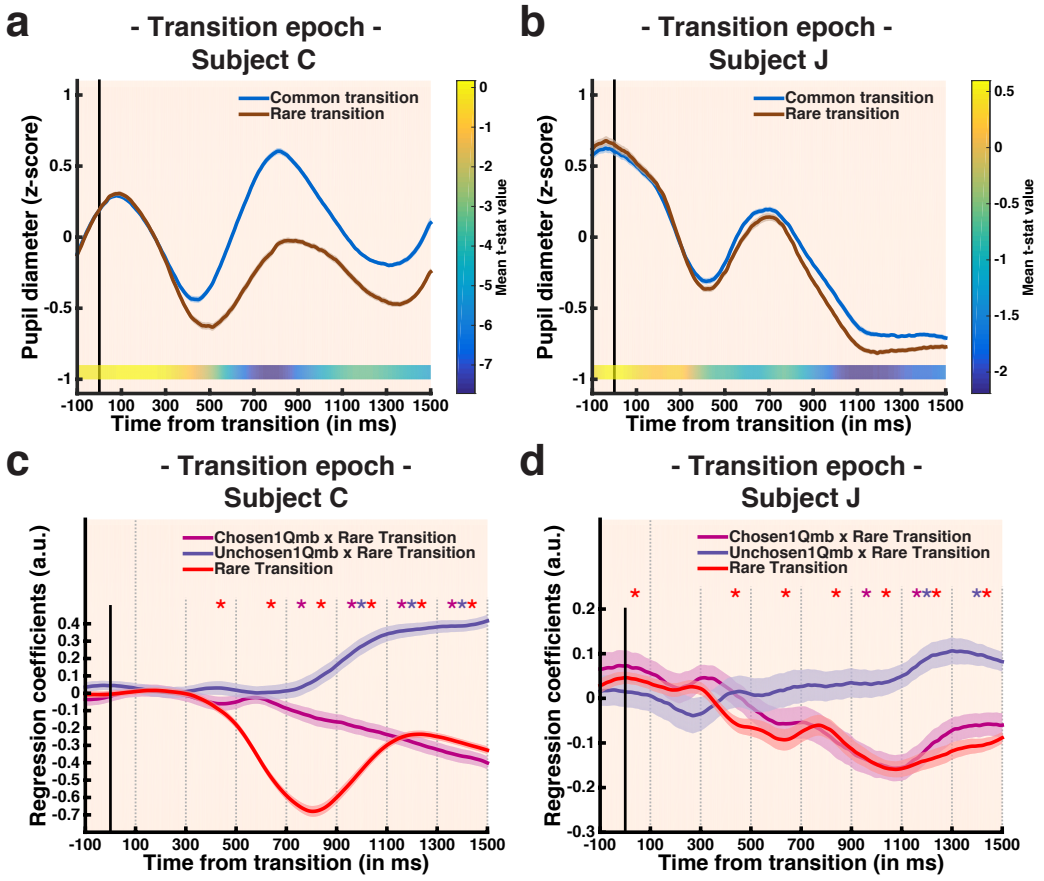


Fig. 4.9 **Pupil size at transition epoch reflected knowledge about the state-transition structure as well as its impact on expected value. (a-b)** Time course of z-scored pupil diameter mean across sessions (shaded regions represent s.e.m.), aligned to transition (i.e. change in the background colour) and sorted by common and rare trials. The horizontal colour bar represents the mean t-statistic for a test that the transition coefficient (coded rare=1 and common=0) is zero on a multiple regression that also included as predictors the model-based first-stage chosen (Chosen1Qmb) and unchosen (Unchosen1Qmb) action-values as well as their interaction with transition (see Eq. 4.7 in Methods for details). **(c-d)** Time courses of the mean across sessions (shaded regions represent s.e.m.) for the effects of the regression interaction terms Chosen1Qmb x Rare transition and Unchosen1Qmb x Rare transition are shown in. Asterisks indicate 200ms time bins in which coefficients differed significantly from zero (two tailed t-test with Bonferroni correction for multiple comparisons).

response of both subjects across the three different outcome levels, we found that expectation of higher reward elicited a significantly stronger pupil dilation (Bonferroni-corrected one-sample t-tests).

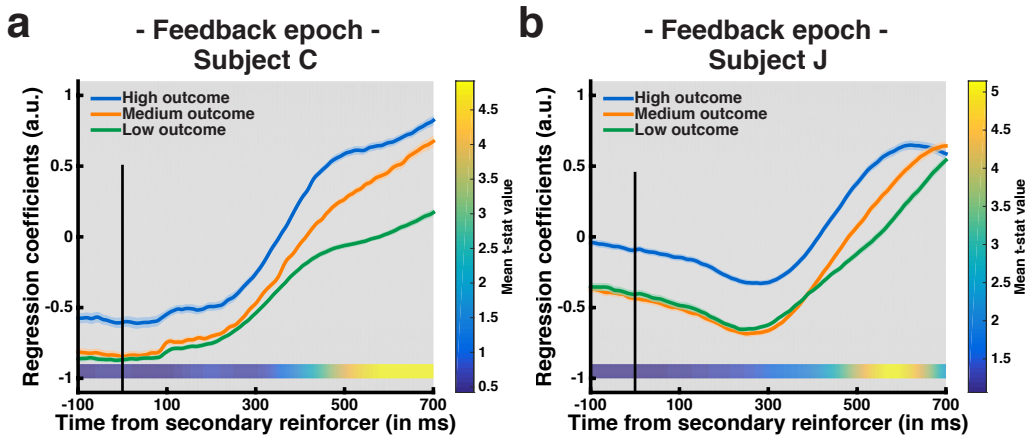


Fig. 4.10 **Pupil dilation as a function of the upcoming reward.** Time course of z-scored pupil diameter mean across sessions (shaded regions represent s.e.m.), aligned to the presentation of the secondary reinforcer cue and sorted by outcome level. The horizontal colour bar represents the mean t-statistic for a test that the outcome level coefficient is zero in a multiple regression on pupil at feedback epoch (also controlling for transition, the outcome level $\times$ transition interaction term, as well as chosen and unchosen second-stage Q-values).

However, also noticeable in Fig. 4.10, is the much higher pupil diameter for the big reward condition even before the secondary reinforcer is presented, consistent with the expected value coding previously found. The reward structure of our task included sequences of a few trials with similar reward levels, making it possible for subjects to anticipate the upcoming reward at feedback as a function of past reward experience. This normally happened when the subject discovered a high reward second-stage option and, taking advantage of its knowledge about the task's structure, subsequently exploited the appropriate first-stage choice. However, if a rare transition had occurred on that trial, this expectation should no longer be supported. Therefore, we performed the same analyses but separating common from rare trial conditions (Fig. 4.11). Consistent with our hypothesis, the anticipatory signal prior to secondary reinforcer onset was only present if a high outcome was received following a common transition.

For the pupil diameter to reflect expectation of reward, it must show evidence of learning from experience. To test this, we not only considered the effect of the current outcome level on pupil at feedback epoch, as before, but also extended as possible predictors the three most recent rewards obtained with the current second-stage choice. In contrast to the prominent

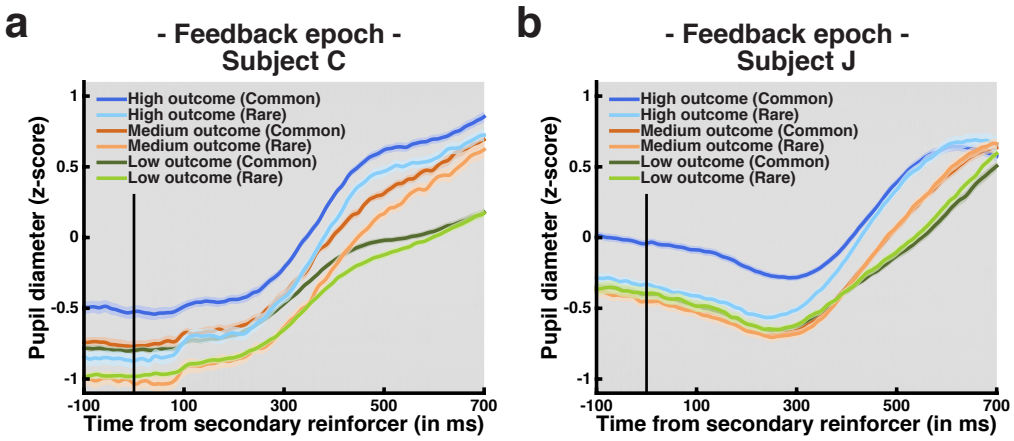


Fig. 4.11 **Pupil dilation as a function of the upcoming reward and transition.** Time course of z-scored pupil diameter mean across sessions (shaded regions represent s.e.m.), aligned to the presentation of the secondary reinforcer cue and sorted by outcome level and transition type on the trial.

positive effect of upcoming reward, the pupil coding for the three preceding outcomes with the same second-stage choice became significantly less positive once the reward level was made known to the subjects (Fig. 4.12a-b; see also Eq. 4.8). In fact, the observed pattern of negative reward weights decaying gradually with trials into the past (Fig. 4.12c), resembled the quantitative reward prediction error encoded by midbrain dopamine neurons (Bayer and Glimcher, 2005).

In order to test this more quantitatively, we regressed the pupil response at feedback against the inputs for the second-stage reward prediction error computation: the expected second-stage chosen Q-value and the outcome level actually received (controlling for transition type). The feedback pupil signal revealed hallmarks of a reward prediction error (Fig. 4.13; see Eq. 4.9) -– an initial positive coding of the second-stage choice value expectation, and a subsequent encoding of the reward outcome. Moreover, we also observed a significant difference between both regression coefficients for the expected value of the second-stage choice and the value of the outcome that was actually received (contrast Outcome level - Chosen2Q in Fig. 4.13).

Finally, we directly regressed pupil size at feedback against the computationally derived reward prediction error at the second-stage choice (controlling for transition type as well as for outcome level and expected value of second-stage choice; the two latter regressors were orthogonalised in respect to the second-stage reward prediction error) and also found the expected positive encoding (Fig. 4.14; see Eq. 4.10).

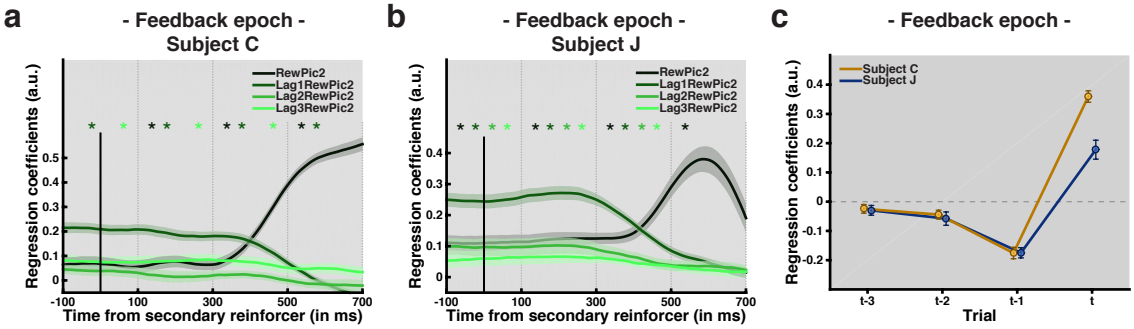


Fig. 4.12 **Pupil changes at feedback epoch encoded the difference between the current reward and a weighted average of previous rewards. (a-b)** Time courses of the mean across sessions (shaded regions represent s.e.m.) for the effects of the upcoming (RewPic2), the most-recent (Lag1RewPic2), the second-most-recent (Lag2RewPic2) and the third-most-recent (Lag3RewPic2) reward experienced with the current second-stage choice regressed on z-scored pupil size at feedback epoch (controlling for transition type; see Eq. 4.8 in Methods for details). Asterisks indicate 200 ms time bins in which the coefficients differed significantly from zero (two tailed t-test with Bonferroni correction for multiple comparisons). The results in **(c)** correspond to the mean (error bars depict s.e.m) value of the RewPic2(t), Lag1RewPic2(t-1), Lag2RewPic2(t-2) and Lag3RewPic2(t-3) baseline adjusted regression coefficients (the first 200 ms mean served as baseline) for the period of 400 to 700 ms of the feedback epoch.

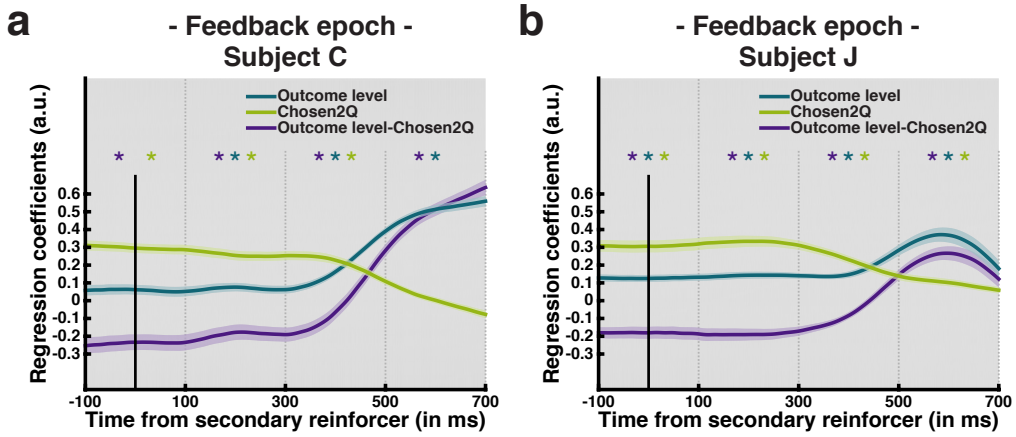


Fig. 4.13 **Pupil size correlated with the inputs for the second-stage reward prediction error computation.** Time courses of the mean across sessions (shaded regions represent s.e.m.) for the effects of the outcome level (Outcome level), second-stage chosen Q-value (Chosen2Q) and the contrast (Outcome level - Chosen2Q) regressed on z-scored pupil size at feedback epoch (controlling for transition type; see Eq. 4.9 in Methods for details). Asterisks indicate 200ms time bins in which the coefficients differed significantly from zero (two tailed t-test with Bonferroni correction for multiple comparisons).

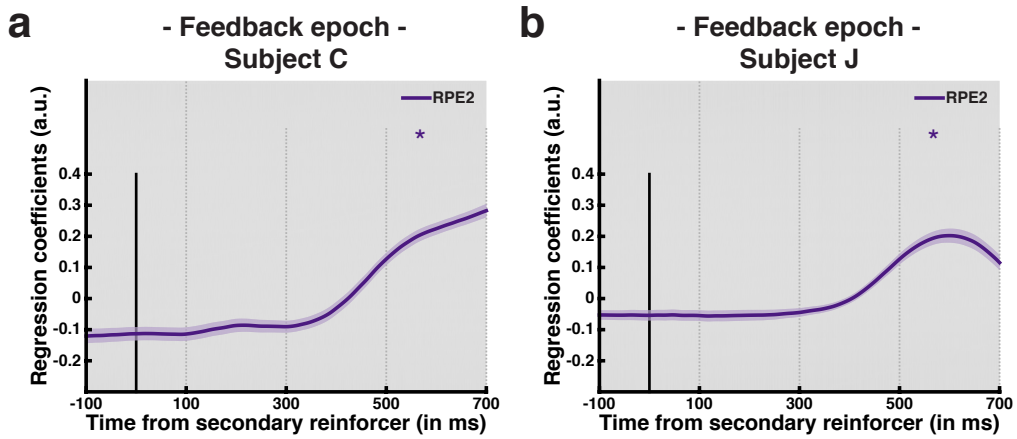


Fig. 4.14 **Pupil size encoded second-stage reward prediction error.** Time courses of the mean across sessions (shaded regions represent s.e.m.) for the effect of the computationally derived second-stage reward prediction error (RPE2) regressed on z-scored pupil size at feedback epoch (controlling for transition type; see Eq. 4.10 in Methods for details). Asterisks indicate 200 ms time bins in which the coefficients differed significantly from zero (one tailed t-test with Bonferroni correction for multiple comparisons).

In conclusion, outcome pupil data reflected the difference between the current reward and a weighted average of previous rewards, consistent with a reward prediction error signal defined by RL theory.

4.5 Discussion

We investigated pupil responses while two non-human primates performed a sequential decision task believed to induce trial-by-trial adjustments in choice that combine both MF and MB-RL control. Our results establish an association between pupil signals and key components of RL at different moments of the task. First, pupil diameter tracked the expected value of choices at both pre- and post-decision moments. Second, these value-related pupil modulations reflected MB-RL computations, as they revealed knowledge of the state-transition structure as well as the impact of this information on upcoming reward expectations. Finally, pupil changes at feedback were consistent with the second-stage reward prediction error signal crucial for value updating.

Effective decision-making requires that the values of possible choices are assessed and upated as they change. It has previously been reported that pupil dilation tracks expected reward; however, those studies did not formally address instrumental learning (Kennerley and Wallis, 2009b; O'Doherty et al., 2003; Varazzani et al., 2015), used indirect measures

of expected value (Gilzenrat et al., 2010) or reported interactions with other variables such as uncertainty (Nassar et al., 2012; Satterthwaite et al., 2007). In fact, this latter association has lead some authors to reject what they call the conjecture of expected reward coding in pupil diameter altogether (Lavin et al., 2014; Preuschoff et al., 2011). Even though our task included some non-stationary elements, we observed consistent choice behaviour that allowed us to establish quantitative links between pupil diameter and trial-by-trial value estimates produced by an RL model, with uncertainty being only a minor factor.

Our data duly suggested that pupil diameter encoded expected chosen values, at different stages and moments of the decision process. There was also a weak, yet significant, effect of the unchosen value as the subject got closer in time to the choice epoch, but that was lost once the decision was made. Some subtleties of our pupil dynamics deserve discussion. The chosen value effect on pupil size remained prominent throughout the pre-choice epochs, even during eye fixation (given the expected pupil constriction with eye accommodation while attempting fixation, any significant effect here is noteworthy). On a previous study, pupil metrics were split into relative size changes and overall average diameter (Nassar et al., 2012), as measures of the different phasic and tonic pupil-linked neuromodulatory activity (Usher et al., 1999), respectively. On that account, one could interpret the sustained correlation between expected value and pupil observed during both pre-choice epochs as a motivational arousal effect on performance, due to a tonic locus coeruleus (LC) activity and norepinephrine (NE) release (Aston-Jones and Cohen, 2005; Bouret and Sara, 2005; Usher et al., 1999). A candidate area for the modulation of such effect is the dorsal anterior cingulate cortex, a brain region not only known to be involved in expected value coding (Shenhav et al., 2013) and pupil dilation (Critchley et al., 2005; Ebitz and Platt, 2015), but also hypothesized to modulate the firing pattern of the LC (Aston-Jones and Cohen, 2005).

The task was designed to elicit both MF and MB RL, allowing us to examine whether the pupil signal was preferentially modulated by one of the two. Consistent with the dominant influence of MB-RL in choice, we found that the pupil response before first-stage choice was better correlated with the computationally-derived MB value estimates. Despite the novelty of the finding, we were more surprised to see that the same pupil changes were not better explained when MF-RL valuation was also taken into account.

Motivated by this, we tested and, indeed, found that pupil response varied with the type of state-transition, once again, in favour of the pupil signal being a MB correlate. This particular difference is strong evidence for the effect of value rather than unexpectedness or uncertainty in our task. This is because a rare trial, as a low probability event given the structure of the environment, would have been a surprise, and so, under those notions,

would have been expected to have elicited more pupil dilation (Nassar et al., 2012; O'Reilly et al., 2013; Preuschoff et al., 2011). Instead, we observed less pupil dilation in rare trials. Perhaps consistent with this, a recent study into the effects of surprise and model change also found a similar late (latencies of around 700-1200 ms in both) decrease in pupil size that covaried with the need to adjust internal representations (O'Reilly et al., 2013).

Similarly, in our task, observing the less expected background colour (requiring sensory processing, and hence time), led to an expected change in the inference about the upcoming reward (Dayan, 2012b; J.Yu and Dayan, 2005). Further evidence for this came in the interaction between transition and expected value: in rare trials, the higher the value of the chosen first-stage option, the more the pupil constricted. This high value would have been supported by the now unavailable, high probability, second-stage state. Conversely, the higher the value of the unchosen first-stage option, the more the pupil dilated. This high value would have been supported by the now available, low probability, second-stage state.

The independent effects on pupil size observed with transition as well as expected value interaction with transition, also allude to links between sensory attention and reward (Gottlieb and Balan, 2010). One possibility for such pupil modulation is the contribution of the cholinergic system (J.Yu and Dayan, 2005), given its role in fast enhancement of visual perception and cortical modulation (Pinto et al., 2013), its phasic response after the detection of cues which change the circumstances of the upcoming reward (Parikh et al., 2007) and its effect in constricting the pupil diameter (Little et al., 1998).

Finally, we considered how current and previous rewards obtained with the same second-stage choice influenced the pupil. While we observed a positive effect of the upcoming reward, the coding of outcomes obtained on previous trials was negative and decayed gradually with increasing trial lags. Such coding of current and previous rewards was reminiscent of the dopamine reward prediction error (Bayer and Glimcher, 2005; Schultz et al., 1997), suggesting possible interneuromodulatory interactions between the dopaminergic circuit and the pupil control systems (Dayan, 2012b; Weinshenker and Schroeder, 2007). While an earlier study pointed to a relationship between pupil size and the learning effects of uncertainty (Nassar et al., 2012), our data thus pointed to the coding of another key computational element that drives learning – the reward prediction error.

Although we observed that the correlates of pupil dilation around the first-stage choice were MB rather than MF, it is not possible to draw any conclusion about the second stage. MB and MF values and prediction errors are essentially the same at the second stage – both follow the same Rescorla-Wagner rule. Thus, although the prediction error is often considered to be MF, we cannot determine whether the pupil had access to both learning

systems just at different times points in the task, or whether it reflected the second-stage MB error signal, and so was always reporting MB-RL.

In conclusion, we found that key elements of RL computations in an adaptive decision-making task were encoded in pupil responses at different time points in the task. The study reinforced the utility of pupillometry as a non-invasive method to track task-induced variables dynamically. However, that the pupil diameter reflected multitudinous signals in the task, which were likely to be calculated and reported by different brain areas, implies one should be cautious in the conclusions drawn from dilation alone.

Chapter 5

Model-free and model-based reinforcement learning in prefrontal cortex and striatal neurons

5.1 Abstract

Growing evidence supports the idea that animals use model-free (MF) and model-based (MB) reinforcement learning (RL) valuations to make choices. Despite the known involvement of multiple and partially separate cortico-subcortical circuits, the underlying mechanisms of the interaction between the strategies remain unknown. Therefore, we recorded single-neuron activity from three prefrontal areas (frontal pole, anterior cingulate cortex and dorsolateral prefrontal cortex) and two dorsal striatal regions (caudate and putamen) while two subjects performed a sequential decision task in which both RL strategies were employed. Here we show neural representations at different time points of key components of MF and MB valuation, such as reward history, state-transition knowledge and choice information. Taking advantage of the two different RL values elicited in the task, we also investigated whether single-neuron activity at the time of choice correlated with MF, MB or both action-value signals. Finally, we related the firing rate of neurons in each recorded population to prediction error signals at two different moments of the task. Prediction errors are essential in updating value expectations or for global reinforcement. Taken together, our finding extends the computational and algorithmic work from previous chapters to provide insight as to how the brain implements MF and MB RL strategies to optimise behaviour.

5.2 Introduction

Humans and other animals seem to use two major competing and cooperating reinforcement learning (RL) systems for behavioural control in sequential decision-making: a goal-directed or model-based (MB) and an habitual or model-free (MF). As explained in detail in chapter 1, both methods rely on previous experience and converge to the same behaviour given enough experience, but they differ as to how this information is used to infer the values of choices. MB-RL computes prospective estimates by integrating reward information with knowledge about the state-transition function, which specifies how the state of the world evolves probabilistically given particular actions. As a less flexible but simpler approach, MF-RL learns without any model of the environment just by bootstrapping sampled experience and taking changes in expectations as indications of errors in its cached value predictions. Either due to computational constraints or a speed-accuracy trade-off, it is advantageous for learning agents to have both strategies and take advantage of each according to the required task.

The neural substrates for these different reward-learning systems involve complex connections between a midbrain-striatal-prefrontal network, which we discussed extensively in chapter 2. The phasic activity of midbrain dopaminergic neurons is known to report a reward prediction error that could drive the updating of MF-RL predictions (Schultz et al., 1997). Evidence from either lesion-based studies in animals or functional neuroimaging techniques in humans, suggests an involvement of the primate putamen or more posterolateral striatum (or dorsolateral striatum in rodents) also in MF learning, whereas more anterior regions of the caudate (or dorsomedial striatum in rodents) and the prefrontal cortex has been implicated in MB-RL (Balleine and O'Doherty, 2009; Daw and Dayan, 2014; Daw et al., 2005; Hampton et al., 2006). Furthermore, neural representations of action-values, which map state-action pairs into expected returns, are essential for the implementation of action selection in RL theory. Neurons throughout dorsal regions of the primate striatum were found to encode cached action-values as computed by MR-RL methods and useful in guiding the choice process (Lau and Glimcher, 2008; Samejima et al., 2005). However, it is important to note that although these neural representations of action-values were considered to be learned using MF methods, the tasks did not elicit differences between MF and MB value computations. Thus, it is not possible to draw any conclusion whether these neurons are actually just MF action-value neurons. In any case, there is no similar reports of neuronal representations for MB action-values – with the closest perhaps being the hippocampal firing patterns consistent with forward search of paths ahead of the animal's current location

(Johnson and Redish, 2007). Finally, very few studies, and none not involving human subjects, have focused on detecting simultaneous neuronal signals of both learning strategies (Daw et al., 2011; Gershman et al., 2014).

Despite the evidence of these two methods for learning and calculating values, several important questions remain unanswered: 1) is there single-neuron evidence of MB-RL as formally proposed in RL theory; 2) how do the two forms of instrumental control interact and determine action choice; 3) how do these signals evolve through the decision-making process? Teasing these various questions apart, and understanding the neural properties and substrates shared or unique to each learning strategy is critical to move modern learning theory forward. Moreover, this knowledge could benefit the application of these RL models to psychiatric and neurological disorders (Maia and Frank, 2011).

To answer these open questions, two rhesus monkeys were trained to perform a two-stage decision task (Fig. 5.1), while single-neuron data was recorded simultaneously from three prefrontal regions (frontal pole, FP; dorsal anterior cingulate cortex, ACC; dorsolateral prefrontal cortex, DLPFC) and from two dorsal striatal areas (caudate and putamen). The simultaneous recordings from different brain regions allowed us to examine whether MF and MB-RL computations are specific or not to different areas, whereas the behavioural paradigm used provided a novel insight into the RL problem as it induced trial-by-trial adjustments in choice that combined both learning strategies.

Although already introduced in previous chapters (chapter 3 and chapter 4), a brief description of the task is provided for general guidance. Two sequential binary choices had to be made before subjects received a reward. A grey background represented the first-stage state and each choice was between two options presented as pictures. Each of these first-stage choices could lead to either a common (70% transition probability) or rare (30% transition probability) second-stage state, represented by different background colours (brown and violet). In the second-stage, another two-option choice between pictures was required and it was reinforced according to three different levels of reward (high, medium or low outcomes). Subjects then returned to the first-stage choice following an inter-trial interval. Importantly, to encourage learning, the outcome level of each second-stage option remained constant for five to nine trials and subsequently switched with probability 2/3 to one of the other two possible outcomes.

In chapter 3 we reported in detail the behavioural evidence that, just like humans (Daw et al., 2011), our subjects solved the two-stage decision task using a hybrid mixture of MB and MF control. In both subjects, reward history (relevant for both learning strategies) and state-transition knowledge (used in MB computations) had a significant impact on be-

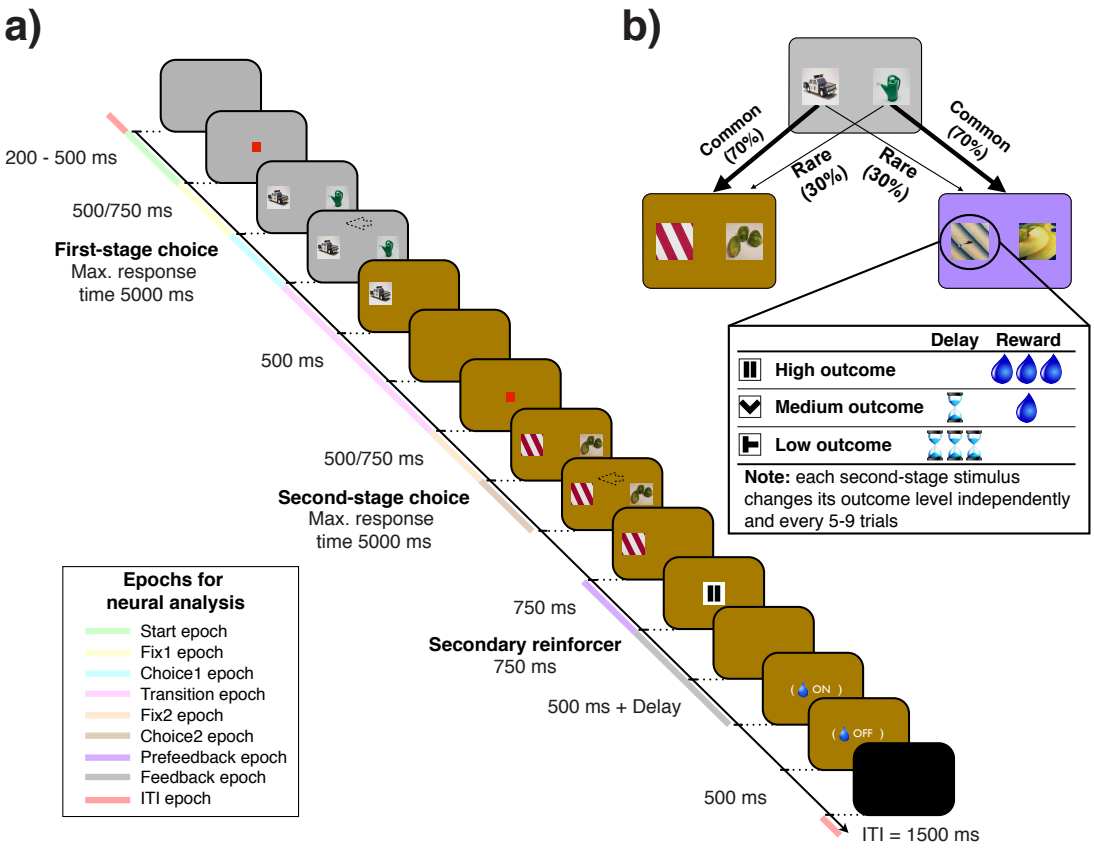


Fig. 5.1 **Two-stage decision task. a**) Timeline of events. Eye fixation was required while a red fixation cue was shown, otherwise subjects could saccade freely and indicate their decision (arrow as an example) with a manual joystick movement. Once the second-stage choice had been made, the nature of the outcome was revealed by a secondary reinforcer cue (here, the pause symbol represents high outcome). Once the latter cue was off the screen, there was a fixed 500 ms delay and the possibility of a further delay (for both medium and low outcomes) before juice was provided (for both high and medium outcomes). The inter-trial interval (ITI) was 1.5s. Epochs for neural analysis are marked along the timeline and each one is represented by colour. **b**) The state-transition matrix (kept fixed throughout the experiment). Each second-stage stimuli had an independent reward structure (with outcomes being defined by the magnitude of the reward and the delay to its delivery) according to a form of random walk sampled afresh on each session. Task design influenced by Daw et al. (2011).

haviour, with their influence decaying exponentially as a function of trials into the past. Here we show the neuronal correlates in prefrontal cortex and striatum of variables such as reward, state-transition and first-stage choice crucial RL computations in this behavioural paradigm. These factors were found either independently or in a combined way throughout all the regions from which we recorded, but some areas exhibited specific functional segregation. We then report single-neuron representations within each brain area of MF and MB action-values, which control action selection. Finally, we focus on outcome updating and prediction error signals known to guide reward learning. This way, we aimed to dissect how the several ingredients of MF and MB-RL are combined and work together in the brain to generate the choice behaviour we observed. Overall, our neurophysiological evidence unifies findings from several other studies suggesting the idea of parallel MF and MB-RL systems. However, the thoroughly intertwined neuronal architecture we found also presents new challenges.

5.3 Materials and Methods

The data presented in this chapter comes from a further analysis of the data set shown in chapter 3. Therefore, all information pertaining to *Subjects and experimental apparatus*, *Task: design and timeline*, *Choice behaviour and reaction time analysis*, *Computational modelling*, *Model fitting procedures*, and *Model comparison and validation procedures* are discussed in detail in the Methods section of chapter 3. Specific methodological features related to the present neuronal data analysis is described.

Neurophysiological Procedures

Surgical procedures were performed using aseptic techniques and under general anaesthesia. Each monkey was implanted with a custom-designed titanium head holder. Following the behavioural training, two recording chambers (C: in PEEK; J: in titanium) were stereotactically (Kopf Instruments, Tujunga, USA) implanted on each animal: one over the left hemisphere at AP = 38(C)/37(J) mm, ML = 20.2(C)/18.1(J) mm (AP, anterior-posterior; ML, medio-lateral; inter-aural line used as reference) and tilted laterally by 21°(C)/26°(J) from vertical; and one over the right hemisphere at AP = 27(C)/27.5(J) mm, ML = 19.7(C)/17.9(J) mm and tilted laterally by 22.5°(C)/28°(J) from vertical. We used a 1.5-T magnetic resonance imaging (MRI) scanner for pre-op planning of the position of these chambers.

In subject C, we recorded simultaneously from the lateral FP, the dorsal bank of ACC

and the DLPFC (dorsal bank of the principal sulcus) in the left hemisphere; and from the the dorsal bank of ACC, the DLPFC (dorsal bank of the principal sulcus), the dorsal caudate and the dorsal putamen from the right hemisphere. In Subject J, we recorded from the lateral FP, the dorsal bank of ACC and the DLPFC (dorsal bank of the principal sulcus) in the left hemisphere; and from the dorsal caudate and the dorsal putamen from the right hemisphere.

For single-neuron recording we used epoxy-coated (FHC Instruments, Bowdoin, USA) or glass-coated (AlphaOmega Engineering, Nazareth, Israel) tungsten microelectrodes inserted through a stainless-steel guide tube mounted on a custom-designed plastic grid with 1 mm spacing between adjacent locations inside the recording chamber. Electrodes were acutely and slowly advanced through the intact dura at the beginning of every recording session using custom-built micro-drive assemblies manually controlled that lowered electrodes in pairs or triplets from a single screw; or motorised microdrives (Flex MT™ and EPS™ by Alpha Omega Engineering, Nazareth, Israel) with individual digital control of electrodes. The approximate distance to lower the electrodes was determined from the MRI images. Once into the desired location, time was given for the brain to settle and thus ensure stability during the recording session. Neurons were randomly sampled; no attempt was made to select neurons on the basis of responsiveness or specific cortical layer. This procedure ensured an unbiased estimate of neuronal activity, thereby allowing a fair comparison of neuronal properties between the different brain regions. After each recording session, the microelectrodes were retracted and the microdrive assemblies were removed from the recording chambers.

Neuronal signals were acquired, amplified, filtered and digitised (OmniPlex™ D Neural Data Acquisition System by Plexon Instruments, Dallas, USA). Spike waveform sorting was performed off-line using principal component analysis-based method (Offline Sorter™ by Plexon Instruments, Dallas, USA). Channels were discarded if either neuronal waveforms could not be clearly separated or because the waveforms did not remain stable throughout the session.

Recording locations

We recorded single-unit spiking activity from a total of 941 neurons (C: 695 and J: 246) in 57 recording sessions (C: 30 and J: 27) across all five investigated regions: FP: 278 neurons, ACC: 240 neurons, DLPFC: 187 neurons, Caudate: 116 neurons, Putamen: 120 neurons (c.f., Appendix Appendix A). We recorded from up to 28 electrodes simultaneously per session (mean/SD: 10/6 electrodes). Across all sessions, all pairs of regions were recorded

simultaneously at least once.

To reconstruct and confirm the correspondence between the MRI sections and our recording chambers and electrode locations we combined information from both pre- and post-op MRI scans, the stereotactic measurements determined at surgery and the neurophysiological mapping of sulci and gray and white matter boundaries during the acute recordings. In the post-op MRI scanning we mounted on the chamber a custom-built grid with several linear tracts that were filled with the magnetic resonance contrast-agent (1:1200 in 0.9% saline) Magnevist® (Bayer HealthCare Pharmaceuticals, Leverkusen, Germany) to visualize the chamber and the angle of the grid tracts. Anatomical MRI data was preprocessed (registration and realignment) using SPM8 software (http://www.fil.ion.ucl.ac.uk/spm/) in order to make sure the interaural axis was in the precise stereotactic location. The chamber positioning, together with the grid tracts, were then plotted onto the post-processing MRI sections using commercial graphics software (Adobe Illustrator, San Jose, CA). A custom-built MATLAB® version R2014b (The MathWorks, Massachusetts, USA) algorithm was used to project each recording location (depth from dura penetration and respective AP and LM grid position) on pre-op MRI scan images (as they had better quality and less artefacts). Figures 5.2 and 5.3 display the full reconstruction with the location of the recorded neurons for subject C and J, respectively.

Task epochs for neuronal analysis

For our neuronal analyses nine different task epochs (each one coloured in Fig. 5.1a) were defined: 1) Start epoch: from 300ms before until 500ms after the presentation of the grey background (first-stage state), which represented the start of the trial; 2) Fix1 epoch: from 500ms before until first-stage pictures presentation; 3) Choice1 epoch: from 300ms before until 750ms after first-stage stimuli presentation; 4) Transition epoch: from 300ms before until 1000ms after background colour changed from first-stage (grey) to the respective state in second-stage (brown or violet); 5) Fix2 epoch: from 500ms before until second-stage pictures presentation; 6) Choice2 epoch: from 300ms before until 750ms after second-stage stimuli presentation; 7) Prefeedback epoch: from 750ms before until the presentation of the secondary reinforcer picture; 8) Feedback epoch: from 300ms before until 1250ms after the presentation of the secondary reinforcer stimulus; 9) ITI epoch: from 1500ms before until the presentation of the grey background (first-stage state), which represented the start of the trial.

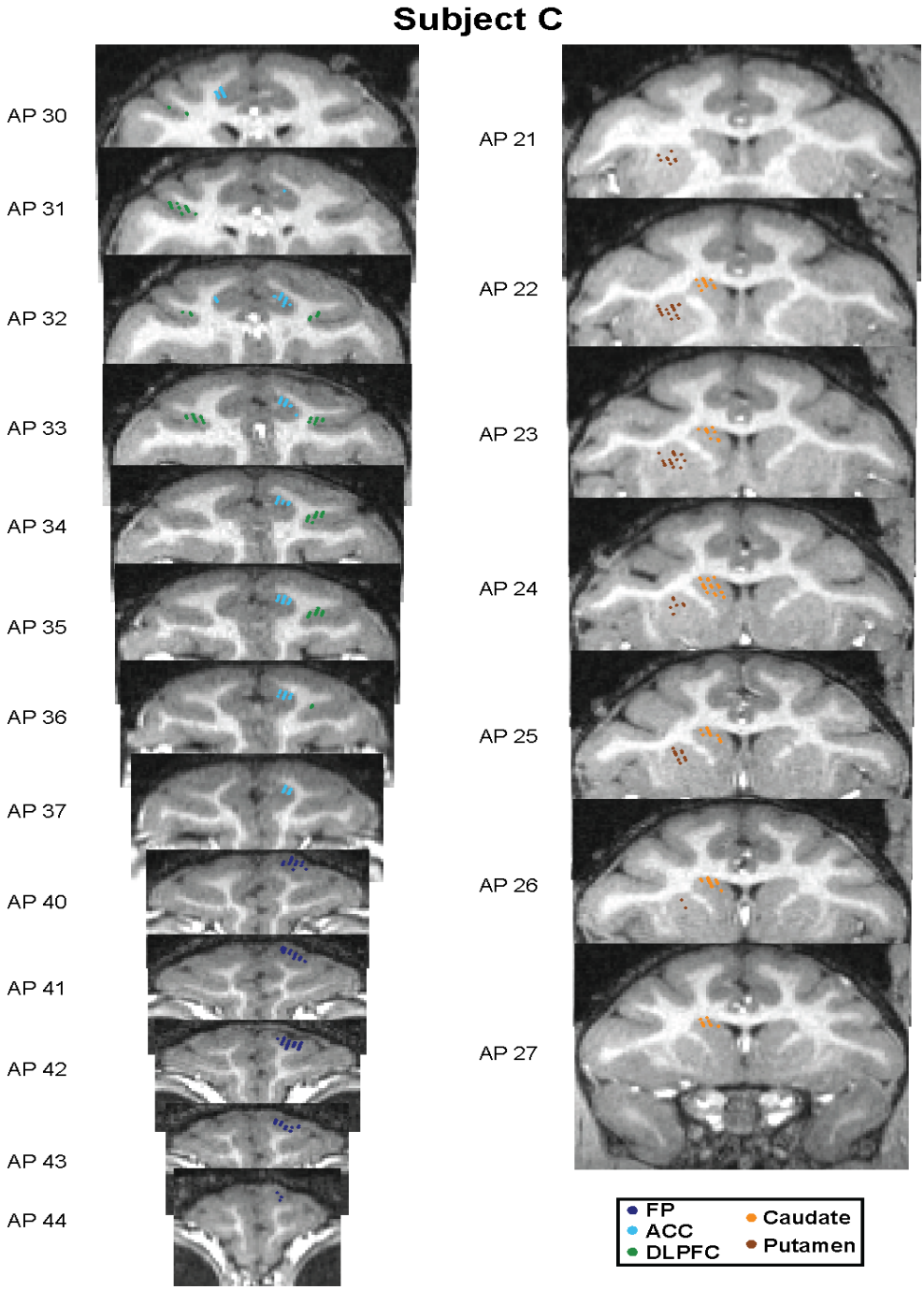


Fig. 5.2 **Locations of neurons recorded from subject C.** Each dot represents the location of one neuron, but the spatial resolution was insufficient to differentiate two very close neurons. Each location was estimated based on depth of penetration, electrophysiological observations during recordings and registration of the recording grid to pre and post-operative MRI scans. Numbers correspond to the distance in millimetres anterior of the inter-aural line (AP).

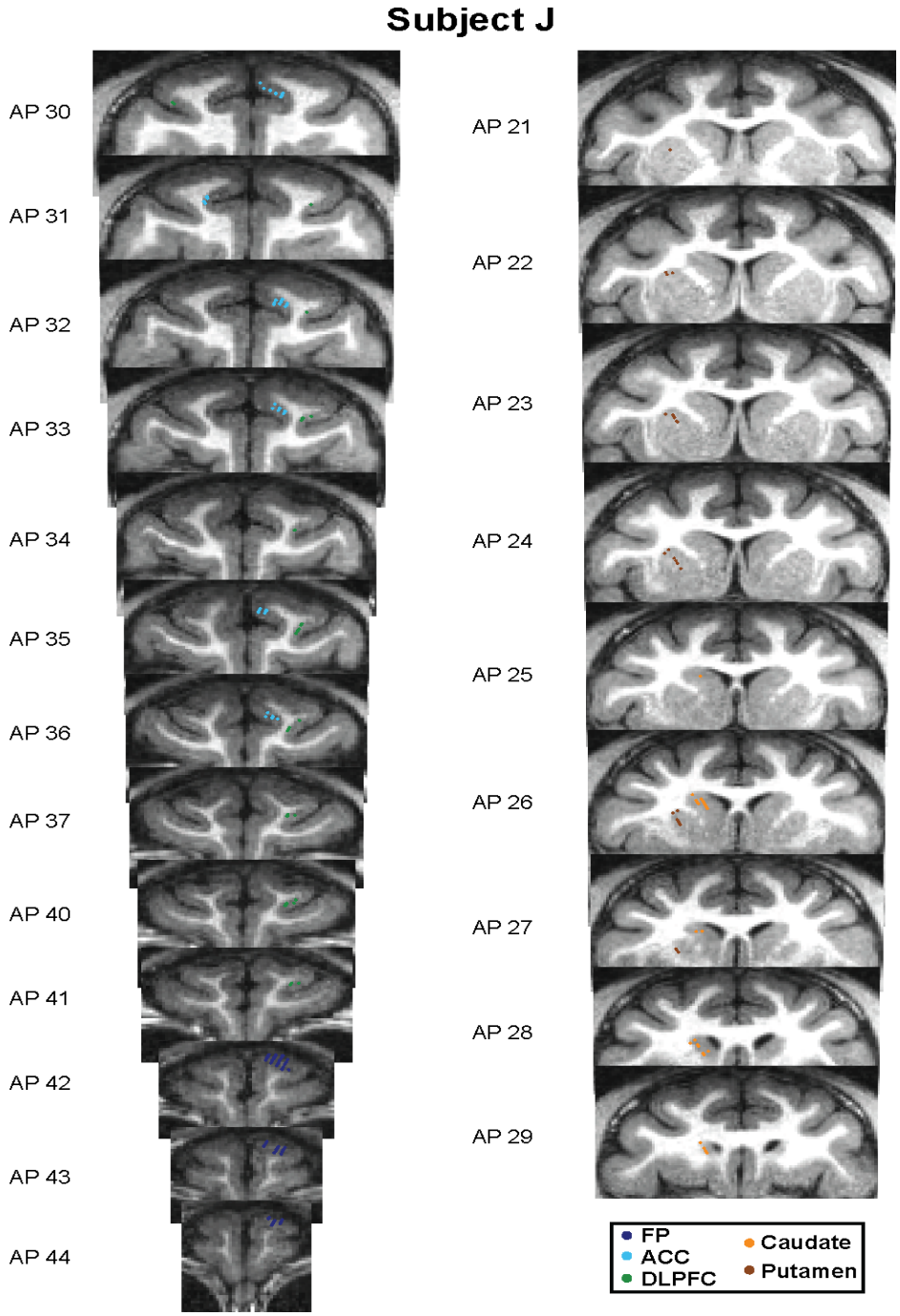


Fig. 5.3 **Locations of neurons recorded from subject J.** Each dot represents the location of one neuron, but the spatial resolution was insufficient to differentiate two very close neurons. Each location was estimated based on depth of penetration, electrophysiological observations during recordings and registration of the recording grid to pre and post-operative MRI scans. Numbers correspond to the distance in millimetres anterior of the inter-aural line (AP).

Neuronal analysis

All statistical and data analyses were conducted using MATLAB® version R2014b (The MathWorks, Massachusetts, USA). To construct spike density histograms, each cell's neuronal activity was standardised by subtracting the mean over all time points and dividing by the standard deviation, averaged across appropriate conditions and smoothed with a sliding Gaussian kernel with σ = 100 ms. Averages of standardised firing rate across all neurons were used for population analysis.

To test whether neurons encoded relevant behavioural or computationally-derived variables general linear models (GLMs) were used, more specifically multiple linear regression (Equation 5.1), that expressed the dependent or observed response variable (neuronal firing rate) y in terms of a linear combination of a constant term (β_0), each regressor (variables of interest or confounds) X_n weighted by coefficients β_n and an error term ε.

$$y = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n + \varepsilon \tag{5.1}$$

For each epoch, the raw neuronal firing rates were averaged during 200-ms time windows, starting in the first 200 ms of the respective epoch and then shifted in 10-ms steps until the end of the entire epoch. Each time window was then standardised by subtracting the mean and dividing by the standard deviation of all time windows across all trials. Regression was then performed at each time bin. All regressors were mean-centred; continuous ones were also scaled by dividing them by two standard deviations so that the magnitudes of coefficients for binary and continuous regressors could be directly compared (Gelman, 2008). Such adjustments in the variables were performed before the computation of the interaction terms.

In the task, the subjects were free to look around when there was no fixation cue. It was then important to control for eye position as this could cause neural activity changes and confound our analysis. To rule out activity due to eye movements, the respective epoch of analysis was divided into non-overlapping 200 ms bins and the mean values of x- and y-positions of the eyes were computed and used as nuisance predictors in all regression models (they are not mentioned as predictors later just for convenience). The only exceptions were the ITI and the start epochs due to technical problems in acquiring calibrated eye data at those times.

To evaluate statistical significance and correct for multiple comparisons across time in our GLMs, we performed a permutation test for every regressor and neuron. To do this we permuted the trial order of the averaged firing rate matrix and then performed the same GLM

analysis on this permuted data. This was repeated for 1000 permutations. For each permuted regression and each regressor, the maximum and minimum t-statistic observed over all time bins was calculated to create an empirical distribution of maximum and minimum t-statistics observed from random data. The 97.5th percentile of the maximum t-statistic distribution and the 2.5th percentile of the minimum t-statistic distribution were taken as the upper and lower significance thresholds respectively. This protocol meant that all regressors and neurons had individual significance threshold t-statistics. Using this permutation method, the typical computed t-statistic threshold (e.g. at feedback epoch in ACC for outcome coding the upper threshold ranged between 2.4 and 2.9), which was considerably stricter than the typical upper threshold computed from a standard t-distribution. The point of maximum selectivity for each neuron was defined as the time bin in each sliding GLM that had the largest absolute t-statistic value for each regressor. The sign of the regression coefficient associated with this bin was then used.

To determine the contribution of each regressor (or group of regressors) in explaining the variance in a neuron's firing rate, we calculated the coefficient of partial determination (CPD) and then averaged this value across all neurons in each brain area to obtain an estimate of the amount of variance explained for each regressor (or group of regressors) at the population level. The CPD for regressor(s) X_i is defined by:

$$CPD(X_i) = \frac{SSE(X_{-i}) - SSE(X_{-i}, X_i)}{SSE(X_{-i})} \tag{5.2}$$

where $SSE(X)$ refers to the sum of squared errors in a regression model that includes a set of regressors X, and X_{-i} is a set of all the regressors included in the full model except X_i (Cai et al., 2011). To note that a neuron could encode significantly a variable but have a relatively low CPD value. This could be explained by the fact that the t-statistics, which determine selectivity, indicate the role of a predictor given that other predictors are in the model. If there is some correlation between predictors in the model, the marginal contribution of a predictor when all other predictors are already included in the model (i.e., the CPD value) could be small. As we report CPD as a percentage, we multiply the CPD value by one hundred. To avoid a bias in the population mean CPD for each regressor being driven by either very selective or non-selective neurons, we excluded the top and bottom 5% of neuronal CPDs and respective SEM in each brain area; we thus refer to population CPD values as the "5% trimmed absolute mean CPD". Further, we sought to understand whether firing rates in a region were positively or negatively related to a regressor on average. To look at this, we first sorted neurons by the sign of their maximum absolute t-statistic (with

negative coders having a negative CPD by convention) and then by averaging we calculated the net CPD for each population.

The latency of coding for a particular regressor of interest was the time at which the population CPD reached half its maximum (determined by continuously falling down from the maximum value). This measure was used as it is robust to differences in the strength of coding or the amount of data and less sensitive to noise compared to measures based on the time at which statistical significance is reached. Outcome coding was already significant before the presentation of the secondary reinforcer, particularly in ACC and putamen. Therefore, to rule out that differences in such pre-stimulus coding between regions lead to spurious latency differences, we subtracted the baseline outcome coding before estimating half maximum latencies. For statistical comparisons of latencies between regions we estimated the SE of latencies by bootstrap across units (100 resamples), and then assessed the significance of latency differences using one-way ANOVA testing.

GLMs for reward, transition and first-stage choice coding

In GLM-behav1, we focused on how each neuron's firing rate was influenced by the most recently experienced reward, transition and first-stage choice information. The regressors used were: a constant (to model the mean firing rate across trials); three main effects regressors defined the most recent information about reward (time the juice pump was on), transition (rare=1; common=0) and first-stage choice (1=car picture or PicA, 0=watering can picture or PicB); and four regressors defined the interactions between the different types of most recent information: reward $\times$ transition (to model multiplicative effects of both reward and transition), reward $\times$ first-stage choice (to model multiplicative effects of both reward and first-stage choice), transition $\times$ first-stage choice (to model the second-stage state), and reward $\times$ transition $\times$ first-stage choice (to model multiplicative effects of reward, transition and first-stage choice). The model was applied to all epochs between feedback and choice1 epochs. Note that the most recent information about reward at feedback epoch corresponds to the reward about to be delivered, transition information at transition epoch is the transition revealed on that epoch and first-stage choice at choice1 epoch is the first-stage choice made on the previous trial. It is important to emphasise that a decision regarding the next first-stage choice can be made from the moment the reward is revealed at the feedback epoch. Hence, we included for the feedback, ITI, start and choice1 epochs two additional regressors: the first-stage choice about to be made (1=car picture or PicA, 0=watering can picture or PicB) and a regressor for whether the subject will repeat or switch its previous first-stage choice (i.e. the interaction term between the most recent and the next first-stage

choices). Because the animal could not predict forced trials, only choice trials on the next first-stage decision were included on these latter four epochs (for the remaining epochs both forced and choice types of trials were included). Finally, at choice1 and choice2 epochs the animal has to perform an action by moving a joystick (towards one of three possible locations) and, therefore, we also included the respective response side of the action taken (coded as a dummy variable and using the left side as the reference group) as well as the reaction time (in ms).

To better investigate how the most recent information about transition on a given trial may influence the way neurons encode the reward information, we applied to the feedback epoch a slightly modified version of the previous GLM-behav1. Therefore, our GLM-behav2 included as regressors: the constant term; a second regressor differentiating the most recent transition information (rare=1; common=0); a third regressor modelled a linear relationship between firing rate and the reward on common trials only; a fourth modelled a linear relationship between firing rate and reward on rare trials only; the remaining predictors were the same and followed the same principles as in GLM-behav1 (such as first-stage recent and next choice, switch or repeat as well as response side and reaction time). The third and fourth regressors were our predictors of interest and each one of them was orthogonalized with respect to the first and second, to ensure that parameter estimates for the second regressor would reflect differences in mean firing rate between common and rare trials.

In order to better understand how single-neuron computations across regions integrate two different types of information obtained at distinct epochs in the trial, we tested whether the the overall strength (or direction) of the encoding of the transition regressor at transition epoch was related to the overall strength (or direction) of coding of reward at feedback. To achieve this goal, we used the mean of either the absolute (for testing overall strength) or signed (to take direction of coding into consideration) beta coefficient values for the transition regressor at transition epoch as a ”population transition code”. We then regressed (we used linear robust regression to avoid the higher sensitivity of least square methods to very selective neurons) this fixed population code against the sliding absolute (for testing overall strength) or signed (to take direction of coding into consideration) beta coefficient values for the reward regressor at feedback epoch.

GLMs for action-value coding

In GLM-comp1, we focused on how each neuron’s firing rate at choice1 epoch was influenced by the trial-by-trial computational estimates of MF and MB action values (or *Q*-

values) for first-stage choice stimuli (car picture or PicA, watering can picture or PicB). The regressors used were: a constant (to model the mean firing rate across trials); the four regressors of interest composed by PicA and PicB Q-values for each MF and MB approaches; the chosen first-stage choice (1=PicA, 0=PicB); the first-stage response side of the action taken (coded as a dummy variable and using the left side as the reference group); and the reaction time (in ms).

We defined the following types of neurons: Q_{MF}-value only as any neuron with a significant regressor at any time of the epoch for either PicA or PicB MF Q-values but not for MB Q-values; Q_{MB}-value only as any neuron with a significant regressor at any time of the epoch for either PicA or PicB MB Q-values but not for MF Q-values; Q_{HYBRID}-value as any neuron with a significant regressor at any time of the epoch for either PicA or PicB MF Q-values or either PicA or PicB MB Q-values. We also looked at the specific selectivity for a particular first-stage stimuli within each of these neuron types, by identifying the neurons where the significant regressors were only significant for one of the stimuli but never for the other one. Finally, we assessed whether the three main types of neurons additionally had selectivity for chosen first-stage choice, chosen first-stage response side, reaction time and eye position.

GLMs for value-based neural signals at transition epoch

The GLM-behav3 assessed the impact of current common versus rare transition on the neural coding at transition epoch for the previous reward, taking into account the current common vs rare transition, and whether the first-stage choice aimed to repeat or switch the second-stage state where that reward was received (i.e., whether first-stage choice was the one more likely associated, given the 70%/30% state-transition matrix used, with the previous trial's second-stage state and hence reward). This model used as regressors: the constant term (to model the mean firing rate across trials); a second regressor differentiating the current transition type (rare=1; common=0); a third regressor differentiating whether the first-stage choice aimed to repeat or switch the previous second-stage state (repeat=1; switch=0); the next for regressors were our regressors of interest and modelled a linear relationship between firing rate and the reward on common and repeat trials only, rare and repeat trials only, common and switch trials only and, rare and switch trials only, respectively; the remaining predictors were the previous and current trial's chosen first-stage stimulus (1=car picture or PicA, 0=watering can picture or PicB) and the transition $\times$ current current trial's chosen first-stage stimulus (to model the second-stage state). The fourth, fifth, sixth and seventh regressors were our predictors of interest and each one of them was orthogonalized

with respect to the first three regressors, to ensure that parameter estimates for the second regressor would reflect differences in mean firing rate between common and rare trials and for the third regressor would reflect differences in mean firing rate between repeat and switch trials.

In line with the previous regression model and to assess directly the effect of transition on expected chosen value neural coding at transition epoch, we used both the first-stage chosen Q-value estimates as well as the second-stage chosen Q-value estimates derived by the best-fit computational model (*HYBRID+* model). In GLM-comp2, the regressors used were then: the constant term (to model the mean firing rate across trials); a second regressor differentiating the current transition type (rare=1; common=0); a third regressor modelled a linear relationship between firing rate and the chosen first-stage Q-value on common trials only; a fourth regressor modelled a linear relationship between firing rate and the chosen first-stage Q-value on rare trials only; a fifth regressor modelled a linear relationship between firing rate and the second-stage chosen Q-value on common trials only; a sixth regressor modelled a linear relationship between firing rate and the second-stage chosen Q-value on rare trials only; the remaining predictors were the current trial's chosen first-stage stimulus (1=car picture or PicA, 0=watering can picture or PicB) and the transition $\times$ current current trial's chosen first-stage stimulus (to model the second-stage state). The third, fourth, fifth and sixth regressors were our predictors of interest and each one of them was orthogonalized with respect to the first two regressors, to ensure that parameter estimates for the second regressor would reflect differences in mean firing rate between common and rare trials.

GLMs for neural representations of second-stage prediction error signals

In regard to the second-stage reward prediction error coding across regions, we aimed to investigate how the firing rate at feedback was modulated not only by the upcoming reward, but also by the previous reward history of the current second-stage choice. At trial t, a given second-stage choice will have its upcoming reward (Rew), its most-recent (Lag1Rew), second-most recent (Lag2Rew) and third-most recent (Lag3Rew) rewards obtained with that choice (zero was assumed if no experienced reward was observed for each lag variables). Therefore, the GLM-behav4 model was conducted using the following regressors: the constant term (to model the mean firing rate across trials); the Rew variable; the Lag1Rew variable; the Lag2Rew variable; and the Lag3Rew variable; the remaining predictors were transition (rare=1; common=0), reward $\times$ transition, current first-stage choice (1=car picture or PicA, 0=watering can picture or PicB), reward $\times$ current first-stage choice, transition

× current first-stage choice, reward × transition × current first-stage choice, next trial first-stage choice and repeat or switch the current first-stage choice on the next trial.

Given the observed results with this last regression, we aimed to test if the neural response at feedback epoch reflected a difference between the current or upcoming reward and the weighted average of previous rewards, consistent with a theoretical reward prediction error signal. For this, we calculated and plotted the mean value (and respective SEM) of the beta coefficients across all neurons of each region for Rew, Lag1Rew, Lag2Rew and Lag3Rew during the period of time of 200 to 600ms post-secondary reinforcer (this period of time was based on the striatal phasic response findings of the GLM-behav4 but it is also consistent with the approximate latency and duration of dopamine increase in striatum after stimulation of dopaminergic neurons; see Schultz 2007).

Similarly to what we did before, we also aimed to confirm the second-stage reward prediction error signal by using the trial-by-trial computational estimates inferred from behaviour, and used two approaches for this. First, we assessed the relationship between firing rate at feedback epoch with the computational inputs of the prediction error signal: the current or upcoming reward, Rew, and the expected chosen Q-value for second-stage, Chosen2Q. Hence, the GLM-comp3 used as regressors: the constant term (to model the mean firing rate across trials); the Rew variable; the Chosen2Q variable; and the remaining predictors were transition (rare=1; common=0), current first-stage choice (1=car picture or PicA, 0=watering can picture or PicB), transition × current first-stage choice, next trial first-stage choice and repeat or switch the current first-stage choice on the next trial. Finally, we regresses neural activity at feedback epoch directly with the second-stage reward prediction error signal derived by the best-fit computational model.

5.4 Results

We recorded simultaneous single-cell activity from 278, 240, 187, 116 and 120 neurons located in FP, ACC, LPFC, Caudate and Putamen, respectively (Fig. 5.2 and 5.3 for recording locations). Our analysis methods focused on determining the percentages of single neurons in each area with significant effects for our variables of interest, as well as on quantifying, at the population level, the averaged amount of variance in neuronal firing rates explained by such factors. We applied these procedures across different task epochs to reveal temporal dynamics of relevant information.

Neuronal encoding of reward, transition and first-stage choice

First, we focused on the relationship between neuronal activity and key elements of choice behaviour, such as reward (from the current trial and into the next trial), transition, the combination of these two variables and first-stage choice.

Reward coding

Because the task design meant that current choice was influenced by both the reward and transition of the previous trial, most of our neuronal analyses explored how task relevant variables were encoded across the epochs linking trials, particularly from feedback of the current trial through to the choice of the next trial. Following the presentation of the secondary reinforcer and during the feedback epoch of the current trial, the proportion of neurons encoding the magnitude of the upcoming reward (Fig. 5.4a) was significantly above chance level (all binomial tests, $p < 0.05$) and relatively high (not less than 75%) in all recorded regions (Fig. 5.4b). The areas with higher percentage of those cells were ACC (94%) and putamen (93%), then caudate (84%), DLPFC (78%) and FP (76%). At the population level, reward size explained more of the variance in the firing rate of ACC neurons than any other region, followed by both striatal regions and finally FP and DLPFC (Fig. 5.4e). Regional differences in the latencies of the reward signals were also observed ($F(4, 495) = 183.7$, $p < 0.05$; see Fig. 5.4c), but here putamen (latency to reach half maximum: 251$\pm$31 ms SE) was the fastest, followed by caudate (280$\pm$13), then ACC (302$\pm$14), FP (327$\pm$10) and finally DLPFC (305$\pm$45). Although none of these areas showed a significant bias in the number of selective neurons encoding reward either positively or negatively (all binomial test, $p > 0.05$), it is possible to find differences in relative strength between the two subpopulations of selective cells with opposite coding schemes. To clarify this, we assessed the net population reward signal for each brain area by averaging together the neuronal variance of significant neurons that encoded reward with either positive (CPD defined as positive value) or negative (CPD defined as negative value) regression coefficients (Fig. 5.4f). Both caudate and putamen had greater tendency to increase their average firing rate for higher outcomes, whereas the relationship in ACC and FP was negative, slightly more prominent and transient in the latter.

We then examined the temporal dynamics of the reward selectivity from feedback until the next first-stage choice (Fig. 5.4d, 5.4e and5.4f). After its initial peak, the encoding of reward in ACC neurons remained substantial during the bulk of the feedback epoch. Despite a subsequent gradual reduction, it remained the region with the highest CPD strongest until

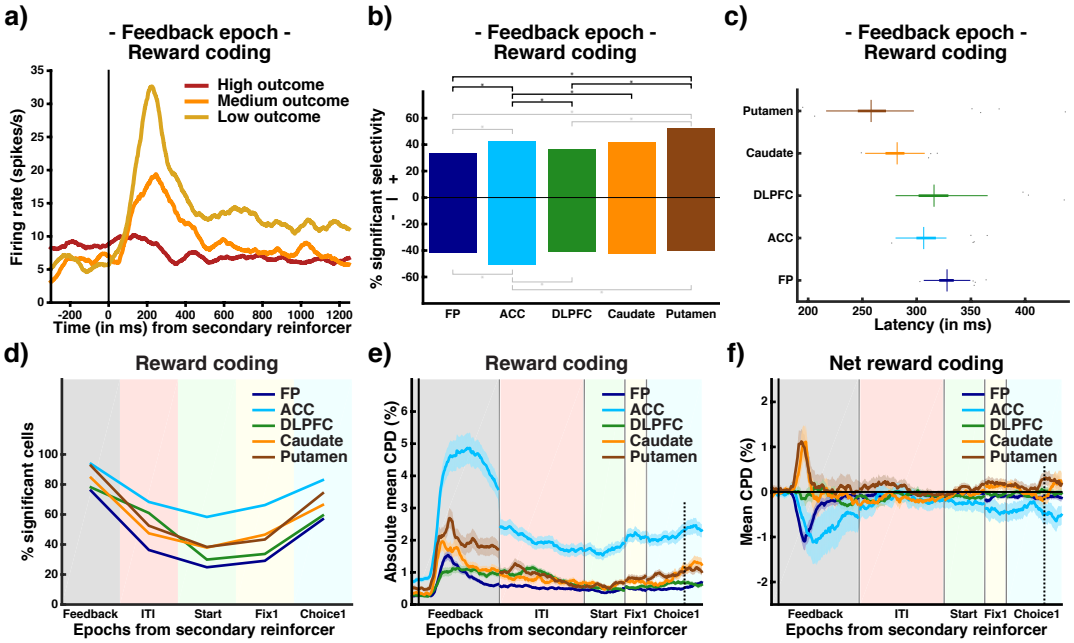


Fig. 5.4 **Population encoding of reward. a)** A spike density histogram of an ACC single-neuron example encoding (negatively) the outcome level at feedback epoch. **b)** Bar plot with the prevalence of neurons significantly encoding reward, based on the sign of the regression coefficient (+ or -). Single black or grey lines and asterisks, $p < 0.05$ (chi-squared tests), for differences between areas in the number of selective cells overall, and each type of signed coding, respectively. No significant differences from the chance 50%-50% split were found for the proportion of neurons with positive or negative regression coefficients (all regions binomial test with $p > 0.05$). **c)** Comparison of reward coding latencies (time to reach half maximum coding) between regions. Grey dots are outliers; vertical thin lines are median values. **d)** Prevalence of neurons significantly encoding reward across epochs from secondary reinforcer presentation to choice1 epoch. **e)** Time course of the reward coding at the population level, as determined by the absolute coefficient of partial determination (CPD) value, from feedback to choice1 epochs. **f)** Time course of the population net reward CPD value, averaged across neurons that significantly encoded reward with either positive or negative regression coefficients, from feedback to choice1 epochs. For **e-f)** The 5% trimmed absolute mean (solid coloured line) and respective SEM (shading) across recorded neurons was used; solid vertical line corresponds to secondary reinforcer presentation; dotted vertical line represents the mean first-stage reaction time across subjects and sessions.

the next first-stage choice. Neuronal activity during feedback in DLPFC showed a sustained post-maximum reward selectivity after reaching its maximum, whereas the coding in both caudate and putamen reduced shortly after their respective peaks. The reward coding in FP was particularly transient and confined solely to the feedback epoch. Finally, it is important to note that in all regions, the neural encoding of the most recent reward increased with the arrival of the next first-stage choice. The time course of this later rise was fastest in ACC (at the start epoch), then followed by caudate, putamen, DLPFC and finally FP.

Transition coding

The information about reward is important for both RL approaches, but the knowledge of the state-transition structure differentiates MB from MF-RL. Therefore, we investigated whether there were neurons that significantly discriminated a common from a rare transition during the transition epoch (Fig. 5.5a). A greater proportion of such neurons was found in ACC (69% of cells) when compared to all other regions ($p < 0.05$ in all pairwise χ^2 tests), with no significant difference found between the number of selective cells increasing their firing rate for a rare (positive coding) or a common (negative coding) transition (χ^2 test with $p > 0.05$; Fig. 5.5b). Putamen and caudate both had slightly fewer selective cells, but also did not show a bias for positive or negative coders (both with χ^2 test with $p > 0.05$; Fig. 5.5b). However, FP and DLPFC were found to have significantly more neurons increasing their firing rate when rare transitions occurred (both χ^2 tests with $p < 0.05$; Fig. 5.5b). At the population-level and when compared to the remaining areas, the neuronal activity in ACC was not only the first (303 ± 13 ms) to show such transition coding ($F(4, 495) = 57.27$, $p < 0.05$; multiple comparison tests all with $p < 0.05$; Fig. 5.5c) but it was also the better explained by the transition regressor (see Transition epoch in Fig. 5.5e). With the exception of putamen, which showed a slight preference at the population level to increase its spiking for common transitions, all other regions increased their firing rate for rare transitions (i.e., positive coding) but the signal was much more pronounced in FP (see Transition epoch in Fig. 5.5f).

For MB-RL value updating, the state-transition information requires integration with the received reward and, therefore, it is important to hold this information until the outcome is revealed. When we analysed how this transition coding evolved in time, we found that the strong coding signal present in ACC at the time of transition persisted until about the moment the outcome was made known to the subject (Fig. 5.5d-e). In fact, when the analysis of transition selective cells was restricted to the short period in the feedback epoch just before the secondary reinforcer presentation (note that Fig. 5.5d feedback results correspond to the

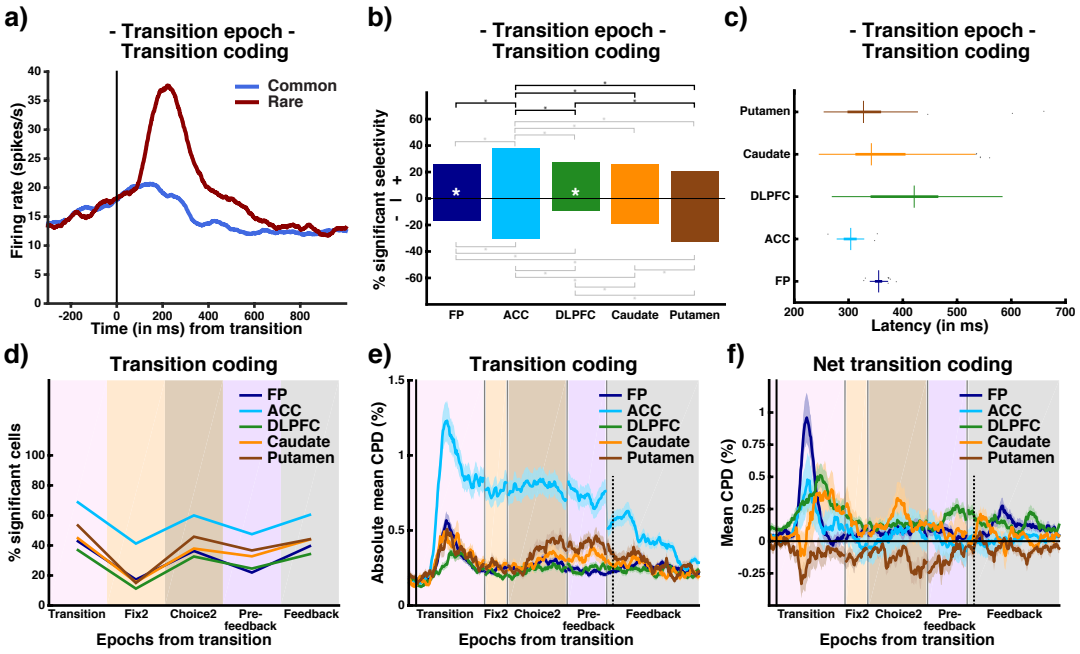


Fig. 5.5 **Population encoding of state-transition information.** **a)** A spike density histogram of an ACC single-neuron example encoding (negatively by convention) the transition type (common versus rare) at transition epoch. **b)** Bar plot with the prevalence of neurons significantly encoding transition type, based on the sign of the regression coefficient (+/− for increased firing rate if rare/common transition, respectively). Single black or grey lines and asterisks, $p < 0.05$ (chi-squared tests), for differences between areas in the number of selective cells overall, and each type of signed coding, respectively; double white asterisks, $p < 0.05$ for the proportion of neurons with positive or negative regression coefficients different from the chance 50%-50% split (binomial test); position of white asterisks indicates the larger population. **c)** Comparison of transition coding latencies (time to reach half maximum coding) between regions. Grey dots are outliers; vertical thin lines are median values. **d)** Prevalence of neurons significantly encoding transition across epochs from transition to feedback epoch. **e)** Time course of the transition coding at the population level, as determined by the absolute coefficient of partial determination (CPD) value, from transition to feedback epochs. **f)** Time course of the population net transition CPD value, averaged across neurons that significantly encoded transition with either positive or negative regression coefficients, from transition to feedback epochs. For **e-f)** The 5% trimmed absolute mean (solid coloured line) and respective SEM (shading) across recorded neurons was used; solid vertical line corresponds to the moment the background colour changes indicating the second-stage state, and hence the transition type; dotted vertical line represents the secondary reinforcer presentation.

entire feedback period), a significantly greater percentage of ACC neurons (22%) compared to all other regions was found (pairwise χ^2 test, all p<0.05), and caudate (11%) and putamen (12%) were the only other areas with a number above chance level of such selective cells (binomial test, all $p < 0.05$).

Combined reward and transition coding

Having found neurons, predominantly in ACC, with modulations in their firing rate caused by either the reward or the transition information, we next focused on the combined effects of these two factors.

First, we focused on potential additive effects. We started by looking at how many cells that significantly differentiated a common from a rare trial at the transition epoch, also encoded the upcoming reward later at feedback. We found that this was the case in most recorded regions (103/119 in FP, 157/165 in ACC, 59/69 in DLPFC, 49/52 in caudate and 58/64 in putamen). We then looked at the same relationship but taking into account positive and negative coding neurons. We did not find a significant difference in the proportion of neurons significantly encoding reward at feedback as a function of whether the transition coding at transition was either positive or negative (all binomial tests with $p > 0.05$). However, positive coding neurons at transition epoch (i.e. neurons increasing their firing rate for rare transitions) were more likely (64% in FP, 71% in ACC, 52% in DLPFC, 57% in caudate and 60% in putamen) to encode negatively reward at feedback (i.e. decrease their firing rate for higher rewards). On the other hand, cells that significantly increased their firing rate for a common transition at transition epoch (negative coders for transition) showed higher tendency (62% in FP, 61% in ACC, 59% in caudate and 59% in putamen) to spike more for high reward (positive reward coders), with the exception of cells from DLPFC (35% in DLPFC).

The next step was to examine similar effects but at the population level. We first asked how much the transition coding at the time of transition could predict the reward coding at feedback across the population of neurons of each recorded region. For this, we regressed the mean regression coefficients for the transition variable across the transition epoch of all neurons of the population against the sliding regression coefficients for reward at feedback epoch (Fig. 5.6a). Because this analysis takes into account the direction of how each factor is encoded (positively or negatively for either transition or reward), we also run other similar regressions but looking at the predictive effect of: 1) the overall coding of transition (i.e. the mean of the absolute regression coefficient values for transition at transition epoch; Fig. 5.6b) on the signed reward coding; 2) the signed transition coding on the overall coding of

reward (i.e. the sliding absolute regression coefficients values for reward at the feedback epoch; Fig. 5.6c);and 3) the overall coding of transition on the overall coding of reward (Fig. 5.6d).

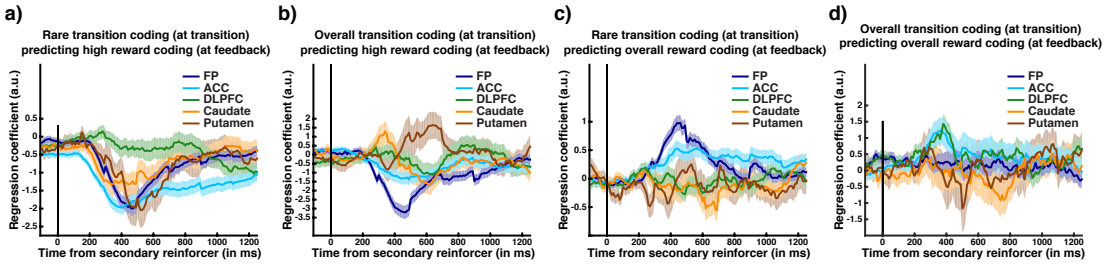


Fig. 5.6 **Relationship between transition coding at transition epoch and reward coding at feedback epoch.** Sliding robust linear regression results of: **a)** the mean value of the transition regression coefficients at transition epoch against reward coding regression coefficients at feedback epoch; **b)** the mean absolute value of the transition regression coefficients at transition epoch against reward coding regression coefficients at feedback epoch; **c)** the mean value of the transition regression coefficients at transition epoch against absolute value of the reward coding regression coefficients at feedback epoch; **d)** the mean absolute value of the transition regression coefficients at transition epoch against absolute value of the reward coding regression coefficients at feedback epoch. Mean (solid coloured line) and respective SEM (shading) of regression coefficients across recorded neurons are shown.

In line with our observations in the single-cell counting analysis, all but DLPFC regions showed a negative correlation between the regression coefficients for signed transition and signed reward codings, with the most negative value found around 350-500 ms post-secondary reinforcement (Fig. 5.6a). We also found that in both striatal regions the greater the selectivity for overall transition coding, the stronger these neurons encoded reward positively at feedback (Fig. 5.6b), and neurons in caudate were slightly faster to show this positive correlation than in putamen. In contrast, FP cells revealed a rather strong negative correlation between both variables, with greater selectivity for overall transition coding at transition being associated with stronger negative reward coding at feedback. The ACC and DLPFC neurons also presented a tendency for a negative correlation, but this was much weaker than in FP. In regards to how much the strength of rare transition coding at transition could predict any type of reward selectivity, a positive relationship was found in FP and, to a less extent, in ACC (Fig. 5.6c).

We also investigated, across the recorded regions, the number of neurons showing at feedback epoch a concomitant transition and reward selectivity (Fig. 5.7a). The ACC (140/240 neurons), when compared to other regions (91/278 in FP, 52/187 in DLPFC, 48/116 in caudate, 50/120 putamen neurons), had a significantly greater percentage of such

cells ($p < 0.05$ in all pairwise χ^2 tests), if the entire feedback epoch was considered. Interestingly, 21% of these ACC cells encoding significantly both variables did so even when the analysis was rather restricted to the short period in the feedback epoch before the secondary reinforcer presentation. This apparent prescience is possible because the reward structure of the task involved sequences of a few trials with similar reward levels. We consider it further below. At the population level, we also observed that neuronal activity in ACC was much better explained by both variables than the other recorded regions either before or after the secondary reinforcer (Fig. 5.7b).

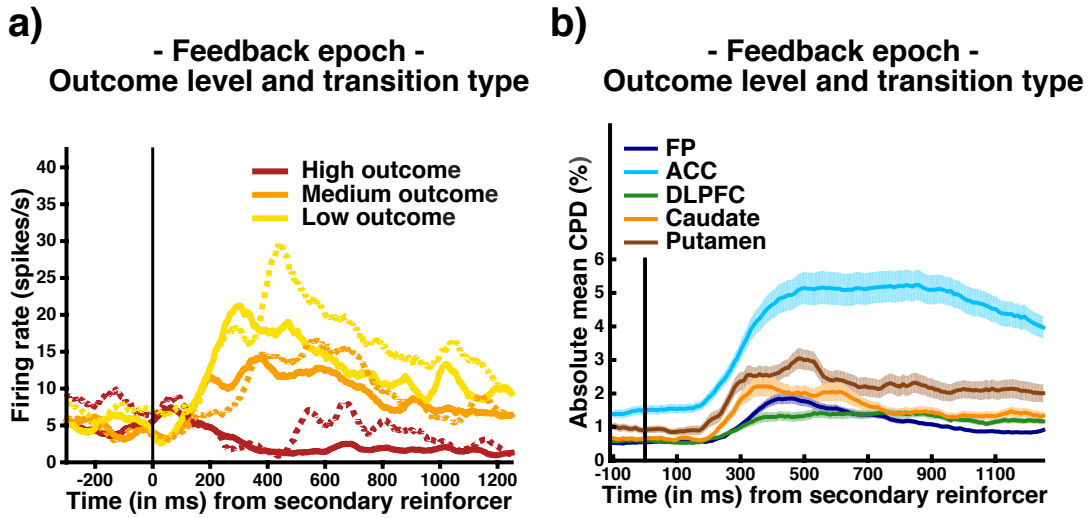


Fig. 5.7 **Additive impact of both reward and transition main effects on neural activity at feedback epoch. a)** A spike density histogram of an ACC single-neuron example of the firing rate at feedback epoch for each outcome level as a function of the transition type on the current trial (solid line=common transition; dotted line=rare transition). **b)** Coding at the population level of information relative to both reward and transition main effects at feedback epoch, as determined by the absolute coefficient of partial determination (CPD) value for both predictors. The 5% trimmed absolute mean (solid coloured line) and respective SEM (shading) across recorded neurons was used. Solid vertical line corresponds to secondary reinforcer presentation.

To understand further this joint modulation of reward and transition at feedback across the population, we examined whether the linear relationship between the neurons' firing rates and reward at feedback was modulated by the transition type in the current trial (Fig. 5.8). Note that, in addition, this analysis also helps to exclude the possibility of a type I error for the observation that some neurons encoded the upcoming reward even before the secondary reinforcer was presented (and also the increased ACC reward coding before secondary reinforcer presentation in Fig. 5.4e). The predictability is greatest when the subject

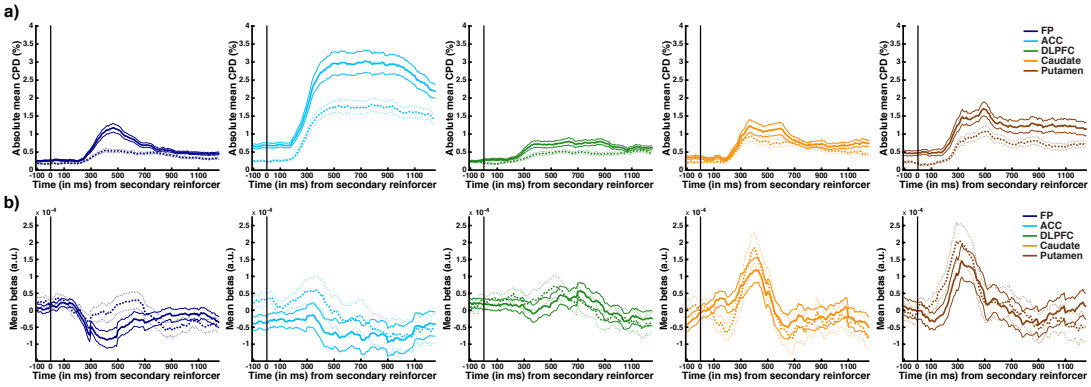


Fig. 5.8 **Population encoding of reward as a function of transition type at feedback epoch.** Reward coding by common (solid coloured line) and rare (dotted coloured line) transition types at the population level of each recorded region, as determined by the absolute coefficient of partial determination (CPD) value (**a**) and the beta regression coefficients (**b**) at the feedback epoch. The 5% trimmed absolute mean (thick line) and respective SE (thin lines) across recorded neurons was used. Solid vertical line corresponds to secondary reinforcer presentation.

discovered a high reward second-stage option and, taking advantage of knowledge about the task's structure, subsequently exploited the appropriate first-stage choice (i.e., the first-stage choice more often associated with the second-stage state where the high reward was found). However, if a rare transition had occurred on that trial, this expectation should reduce significantly. This is because, even if the subject had a long memory trace of the second-stage stimuli rewards, the reward assignments kept changing for the unchosen options, making it hard for the subject to guess the reward associated with an option that is had not experienced recently.

The overall coding of reward, as measured by the CPD, was stronger if the current trial had a common transition than if rare across the recorded regions, but the difference was much stronger in ACC, putamen and FP (Fig. 5.8). Regarding these three regions, the difference between transition types was transient in FP but mantained across the entire feedback epoch for both ACC and putamen. As expected, the anticipatory reward coding (i.e., before the secondary reinforcer presentation), which was most prominent in ACC and putamen, dropped substantially following a rare transition (Fig. 5.8a).

The regression coefficients for the coding of reward (Fig. 5.8b) after the secondary reinforcer are shown and tell a slightly different story. The only region in which these differed as a function of the transition type was the FP, with a stronger negative coding for reward for common versus rare transitions. In other words, FP appears to possess in a more homogeneous population code (given that we considered signed regression coefficients of all

neurons) for information regarding the current state-transition context when coding reward at feedback.

In a final analysis, we used the reward × transition interaction effect (i.e., the product of both terms as regressor) to see if the effect of reward on the neuronal firing rate depended upon the current transition type. Neurons showing a significant reward × transition effect after the secondary reinforcer presentation were present in a proportion above chance level (all binomial tests with $p < 0.05$) in all regions recorded (30% in FP, 46% in ACC, 35% in DLPFC, 39% in caudate and 38% in putamen; Fig. 5.9a), and both ACC and putamen had a significant bias towards coding the effect of the interaction term positively (i.e., the extent to which high outcomes increase firing rate is greater on rare transition trials). Among the prefrontal areas, ACC had significantly more neurons selective for reward × transition ($p < 0.05$ in all pairwise χ^2 tests; Fig. 5.9a), albeit not significantly more so than caudate or putamen ($p > 0.13$ in both pairwise χ^2 tests). At the population level, the coding of reward × transition was generally weak, and did not reflect differences between regions (Fig. 5.9b).

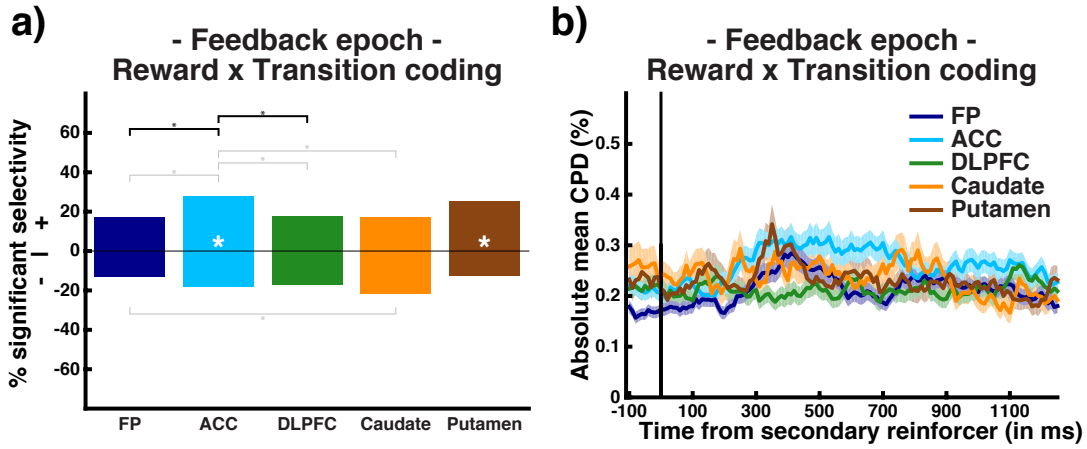


Fig. 5.9 **Population encoding of the reward × transition interaction at feedback epoch. a)** Bar plot with the prevalence of neurons significantly encoding reward × transition interaction, based on the sign of the regression coefficient (+/− if higher rewards have a stronger positive/negative relationship to firing rate in rare trials as compared to common trials, respectively). Single black or grey lines and asterisks, $p < 0.05$ (chi-squared tests), for differences between areas in the number of selective cells overall, and each type of signed coding, respectively; double white asterisks, $p < 0.05$ for the proportion of neurons with positive or negative regression coefficients different from the chance 50%-50% split (binomial test); position of white asterisks indicates the larger population. **b)** Time course of the reward × transition interaction coding at the population level, as determined by the absolute coefficient of partial determination (CPD) value at feedback epoch. The 5% trimmed absolute mean (solid coloured line) and respective SE (shading) across recorded neurons was used; solid vertical line corresponds to secondary reinforcer presentation.

At this stage, it is important to summarise the results from the above neuronal analysis on reward and transition. In regards to reward coding, we found that: 1) the ACC was the region that most prominently encoded the value of the reward at the time of feedback, and this information remained present until the subsequent first-stage choice; 2) the relationship between firing rate and the reward magnitude was predominantly positive for striatal regions and negative for both ACC and FP. When the neural activity related to state-transition information was assessed: 1) the ACC was also the region that most prominently encoded the transition type from when this was revealed until the feedback epoch; 2) at the level of the whole population, FP coding showed a particular bias towards positive encoding of rare transitions. Finally, we also looked at the way both reward and transition was combined and found the following: 1) at the time of feedback, reward and transition information was simultaneously present in ACC; 2) there were specific relationships between the way FP neurons were selective for the transition, when this was revealed, and the way they coded for rewards, at the time of feedback; and finally, 3) for FP neurons, the actual transition experienced in a trial influenced the coding of reward at the time of feedback.

First-stage decision coding

Next we considered whether neurons encoded information about the first-stage decision, including both the selected picture (Fig. 5.10) as well as the motor action required to realise this choice (Fig. 5.11).

In the choice1 epoch, most regions contained neurons that encoded the first-stage stimulus (Pic A or Pic B) about to be chosen (all binomial tests, $p < 0.05$; Fig. 5.10a-b). The proportions of such neurons were similar across regions (most χ^2 test had $p > 0.05$; see for details Fig. 5.10b). Furthermore, no substantial differences were evident at the population level when we analysed at choice1 epoch the impact of first-stage stimulus choice on the spiking activity of the neurons from the five regions (Fig.5.10c), although ACC presented a consistent trend for slightly greater coding. In contrast, FP did not show any encoding for first-stage chosen stimulus throughout the entire epoch.

We also found neurons across the recorded regions encoding the first-stage stimulus chosen on the previous trial (independently of whether it was also chosen on the current trial) and whether the current first-stage choice was a repeat or a switch of the previous trial's one (both with all binomial tests, $p < 0.05$). The proportion of these two latter types of cells was also not significantly different across regions (most pairwise χ^2 test had $p > 0.05$). However, the ACC was the region with the highest likelihood of having any of the three types of the first-stage selective cells ($p < 0.05$ in all pairwise χ^2 tests) and this

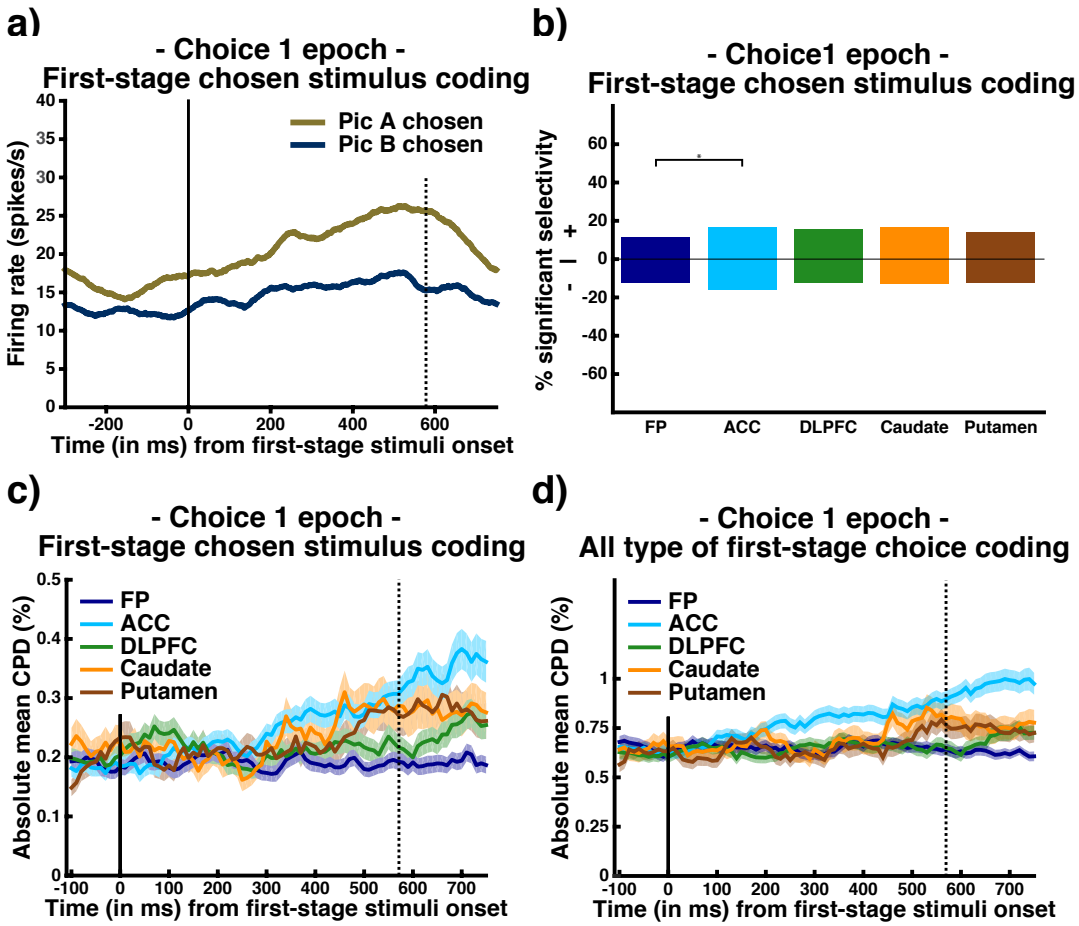


Fig. 5.10 **Population encoding of first-stage stimulus at choice1 epoch. a)** A spike density histogram of an ACC single-neuron example encoding (positively by convention) the chosen first-stage stimulus (car picture or PicA versus watering can or PicB) at choice1 epoch. **b)** Bar plot with the prevalence of neurons significantly encoding the chosen first-stage stimulus, based on the sign of the regression coefficient (+/− for increased firing rate if PicA/PicB, respectively). Single black or grey lines and asterisks, $p < 0.05$ (chi-squared tests), for differences between areas in the selective cells overall number and for each type of signed coding, respectively. **c)** Chosen first-stage stimulus coding at the population level, as determined by the absolute coefficient of partial determination (CPD) value at choice1 epoch. **d)** Coding at the population level of all types of information relative to first-stage stimulus choice, as determined by the absolute CPD value at choice1 epoch for three predictors: currently chosen first-stage stimulus, previous trial's first-stage chosen stimulus and repeat/switch of first-stage choice. The 5% trimmed absolute mean (solid coloured line) and respective SEM (shading) across recorded neurons was used; solid vertical line corresponds to the first-stage stimuli presentation; dotted vertical line represents the mean first-stage reaction time across subjects and sessions.

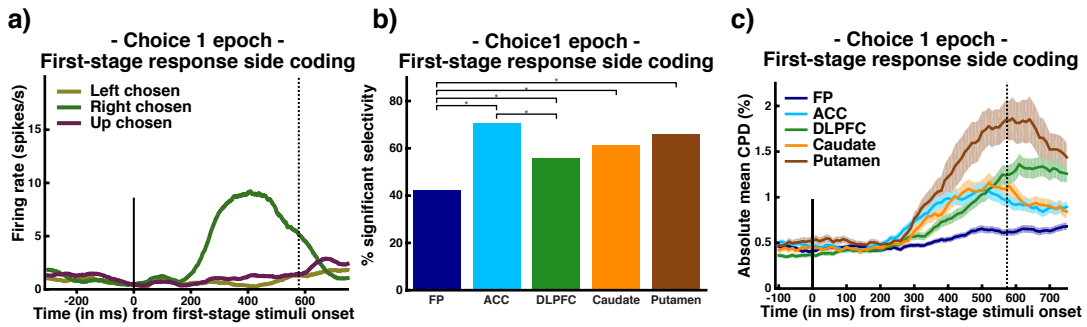


Fig. 5.11 **Population encoding of the chosen first-stage response side at choice1 epoch.** **a)** A spike density histogram of a putamen single-neuron example encoding the chosen first-stage right side at choice1 epoch. **b)** Bar plot with the prevalence of neurons significantly encoding any of the three possible chosen first-stage response sides (left side versus right side versus down/up side for subject C/J respectively). Single black lines and asterisks, $p < 0.05$ (chi-squared tests), for differences between areas in the selective cells overall number. **c)** Chosen first-stage response side coding at the population level, as determined by the absolute coefficient of partial determination (CPD) value at choice1 epoch. The 5% trimmed absolute mean (solid coloured line) and respective SEM (shading) across recorded neurons was used. Solid vertical line corresponds to the first-stage stimuli presentation; dotted vertical line represents the mean first-stage reaction time across subjects and sessions.

was also the case when we examined how much neuronal variance was explained by all three types of first-stage stimulus information (Fig. 5.10d).

Finally, to make a choice between pictures at first-stage, subjects have to perform a joystick movement to select that option. We also found in all regions significantly more neurons than chance that encoded the first-stage side of response (Fig. 5.11a) (all binomial tests, $p < 0.05$). They existed in relatively high proportions (Fig. 5.11b). The FP had the smallest percentage of such cells ($p < 0.05$ in all pairwise χ^2 tests). When we looked at the population explained variance of each region, it was evident that the first-stage response side coding was stronger in putamen (Fig. 5.11c).

Action-value coding

When making choices in sequential decision tasks, animals often make predictions about the value of candidate actions in order to obtain good outcomes and avoid bad ones. Under the RL framework, MF and MB methods both estimate action-values (or Q-values), which specify the overall amount of reward expected in the long-run depending on the actions the agent might take. We used multiple linear regression to assess the extent to which trial-wise MF and MB Q-values for each of the first-stage stimuli, derived from the best fitting

computational model, predicted the firing rate of the recorded neurons. In addition to their participation in choice by encoding action-values of the available options, neurons may also encode the value of the option actually chosen, possibly to contribute to value updating (Lau and Glimcher, 2008; Samejima et al., 2005). Therefore, we not only looked at the selectivity for MF or MB Q-values of any of the two first-stage stimuli (Fig. 5.12 and Table 5.1), but also investigated any additional significant encoding for the chosen first-stage stimulus (Table 5.1). Finally, other first-stage behavioural measures including response side, reaction time (RT), and eye position were taken into account.

Table 5.1 **Regression summary with MF, MB and Hybrid neuronal cell types for action-value coding**

		FP	ACC	DLPFC	Caudate	Putamen
Q_{MF}-value only		**61 (22%)**	**40 (17%)**	**40 (21%)**	**26 (22%)**	**23 (19%)**
	(for only one first-stage stimulus)	53 (87%)	30 (75%)	33 (83%)	22 (85%)	14 (61%)
	+ first-stage choice	13 (21%)	16 (40%)	11 (28%)	7 (27%)	5 (22%)
	+ first-stage side	33 (54%)	30 (75%)	21 (53%)	14 (54%)	18 (78%)
	+ first-stage RT	24 (39%)	18 (45%)	13 (33%)	12 (46%)	16 (70%)
	+ first-stage eye position	37 (61%)	33 (83%)	34 (85%)	17 (65%)	14 (61%)
Q_{MB}-value only		**57 (21%)**	**52 (22%)**	**38 (20%)**	**42 (36%)**	**30 (25%)**
	(for only one first-stage stimulus)	36 (63%)	30 (58%)	25 (66%)	27 (64%)	23 (77%)
	+ first-stage choice	16 (28%)	24 (46%)	19 (50%)	7 (17%)	12 (40%)
	+ first-stage side	25 (44%)	33 (63%)	25 (66%)	21 (50%)	20 (67%)
	+ first-stage RT	22 (39%)	25 (48%)	11 (29%)	20 (48%)	16 (53%)
	+ first-stage eye position	39 (68%)	43 (83%)	30 (79%)	34 (81%)	23 (77%)
Q_{Hybrid}-value		**79 (28%)**	**118 (49%)**	**60 (32%)**	**28 (24%)**	**33 (28%)**
	(for only one first-stage stimulus)	23 (29%)	16 (14%)	19 (31%)	5 (18%)	7 (21%)
	+ first-stage choice	22 (28%)	61 (52%)	22 (37%)	10 (36%)	9 (27%)
	+ first-stage side	41 (52%)	85 (72%)	33 (55%)	22 (79%)	25 (76%)
	+ first-stage RT	31 (39%)	69 (58%)	23 (38%)	13 (46%)	26 (79%)
	+ first-stage eye position	56 (71%)	103 (87%)	54 (90%)	25 (89%)	27 (82%)
No action-value		**81 (29%)**	**30 (12%)**	**49 (26%)**	**20 (17%)**	**34 (28%)**

In bold are the main neuronal cell types and for these the percentages listed in parentheses are relative to the total number of recorded neurons in the region. The other rows list additional selectivity patterns for the three main cell types, with the percentages in parentheses relative to the number of neurons in the respective main neuronal cell type. Abbreviations: FP, frontal pole; ACC, anterior cingulate cortex; DLPFC, dorsolateral prefrontal cortex; MF, model-free; MB, model-based; RT, reaction time.

We found neurons in all brain areas having significant regression coefficients at any time of the choice1 epoch for the first-stage stimuli MF Q-values but not for MB Q-values (22% in FP, 17% in ACC, 21% in DLPFC, 22% in caudate and 19% in putamen; Fig. 5.12a-b). Similarly, we found neurons in all areas which encoded MB Q-values but not MF Q-values (21% in FP, 22% in ACC, 20% in DLPFC, 36% in caudate and 25% in putamen; Fig. 5.12d-e; Table 5.1). No significant difference in the proportion of cells was found across regions for the MF Q-value only neurons ($p > 0.05$ in all pairwise χ^2 tests). On the other

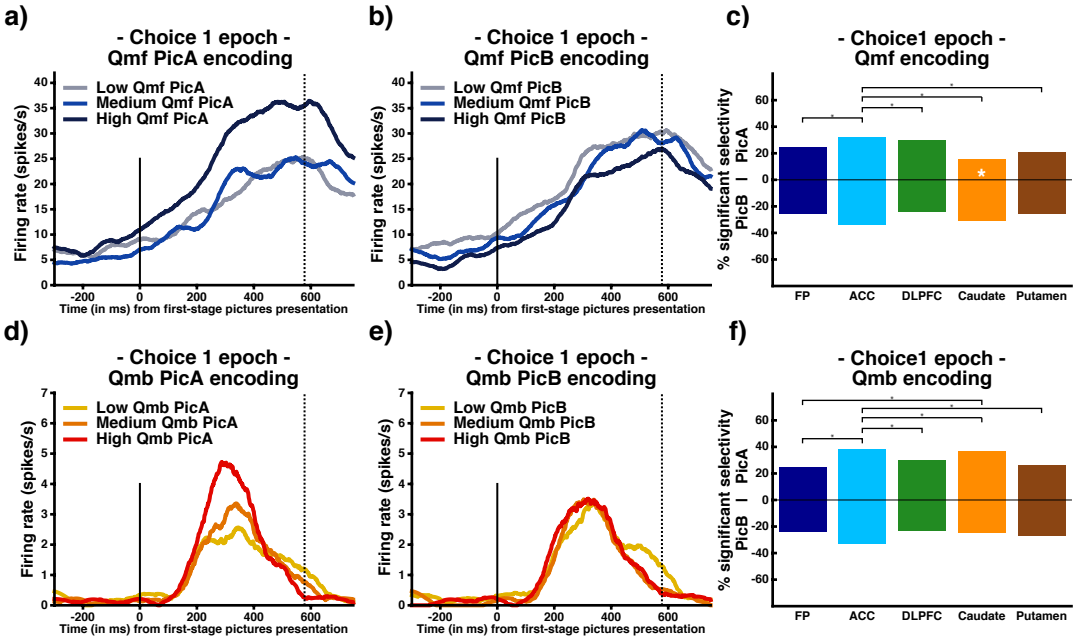


Fig. 5.12 **Model-free and model-based action-value coding at choice1 epoch. a-b)** A spike density histogram of a caudate single-neuron example encoding (positively by convention) the model-free action-value (Qmf) for picture A (PicA) but not for picture B (PicB). **c)** Bar plot with the prevalence of neurons significantly encoding Qmf action-value, based on the sign of the regression coefficient (+/− for Qmf PicA/Qmf PicB selectivity, respectively). Single black or grey lines and asterisks, $p < 0.05$ (chi-squared tests), for differences between areas in the number of selective cells overall, and for each type of signed coding, respectively; double white asterisks, $p < 0.05$ for the proportion of neurons with positive or negative regression coefficients different from the chance 50%-50% split (binomial test); position of white asterisks indicates the larger population. **d-e)** A spike density histogram of an ACC single-neuron example encoding (positively by convention) the model-based action-value (Qmb) for picture A (PicA) but not for picture B (PicB). **f)** Bar plot with the prevalence of neurons significantly encoding Qmb action-value, based on the sign of the regression coefficient (+/− for Qmb PicA/Qmb PicB selectivity, respectively). Single black or grey lines and asterisks, $p < 0.05$ (chi-squared tests), for differences between areas in the number of selective cells overall, and for each type of signed coding, respectively. For **a-b)** and **d-e)**, solid vertical line corresponds to the first-stage stimuli presentation; dotted vertical line represents the mean first-stage reaction time across subjects and sessions.

hand, the caudate contained significantly more neurons encoding MB Q-value only than any prefrontal region ($p < 0.05$ in all pairwise χ^2 tests) and marginally higher percentage than putamen (pairwise $\chi^2 = 9.30$, $p = 0.06$). The large majority of both MF Q-value only and MB Q-value only selective cells showed specific selectivity for only one of the first-stage stimuli (Fig. 5.12a-b, d-e) and not for the other (percentages within MF Q-value only neurons: ranged from 61% in putamen to 87% in FP; percentages within MB Q-value only neurons: ranged from 58% in ACC to 77% in putamen).

In order to realize the hybrid choice behaviour detailed in chapter 3, an integrated value signal needs to be computed in which MF and MB value signals are weighted and combined. An important question is whether this can be observed at the single-neuron level. In other words, whether neurons have access to action-values of both learning strategies during the choice1 epoch. Although all recorded regions included such Hybrid Q-value neurons (28% in FP, 49% in ACC, 32% in DLPFC, 24% in caudate and 28% in putamen), a significantly greater percentage of these cells was found in ACC ($p < 0.05$ in all pairwise χ^2 tests). Despite there being such a high percentage of Hybrid Q-value neurons in ACC, it is interesting to note that only a relatively small percentage (14%) showed specificity for MF and MB Q-values of the exact same first-stage choice (Table 5.1). This is the same as saying that a high percentage of the ACC Hybrid Q-value neurons had, during the entire period of choice1 epoch, access to a mixture of information regarding both MF and MB action-value coding of both first-stage stimulus, which could facilitate the calculation by somewhere else of the Hybrid Q-values.

We also assessed how many of the selective MF and MB action value cells also showed significant regression coefficients to the other behavioural measures analysed (Table 5.1). Overall, there was no significant difference between regions in this characteristic. However, ACC was the region most likely to combine choice with MF coding (40% among the selective MF Q-value only cells), and both DLPFC and ACC also encoded more frequently MB Q-value only and choice (50% and 46% among the selective MB Q-value only cells of DLPFC and ACC, respectively). Note, though, the consistently high percentages of putamen action-value neurons that also encoded first-stage RT and chosen side of response (Table 5.1).

Next, we considered the same information, but now at the level of the whole population. In each area recorded, we measured how much variance in each neuron's firing rate was explained by the MF Q-values only and the MB Q-values only (Fig. 5.13a-b). We found that both MF and MB Q-values consistently explained more of the variance in firing rate in ACC neurons than in neurons from any other region. To investigate if a bias towards

a particular learning strategy was present in each brain area, we calculated the averaged explained variance for the population taking into account all neurons CPD for MF only and MB only action-values (Fig. 5.13c; the convention was positive values for MF and negative values for MB, therefore, an averaged negative value indicates bias towards MB action value coding). ACC neurons showed a strong bias for a MB Q-value coding, leading to a large net negativity in the population response. Caudate and putamen were also biased in favour of a MB strategy, but to a lesser and later extent than the ACC. That there is no net value signal at the population level in FP and DLPFC arises since MB and MF influences are balanced.

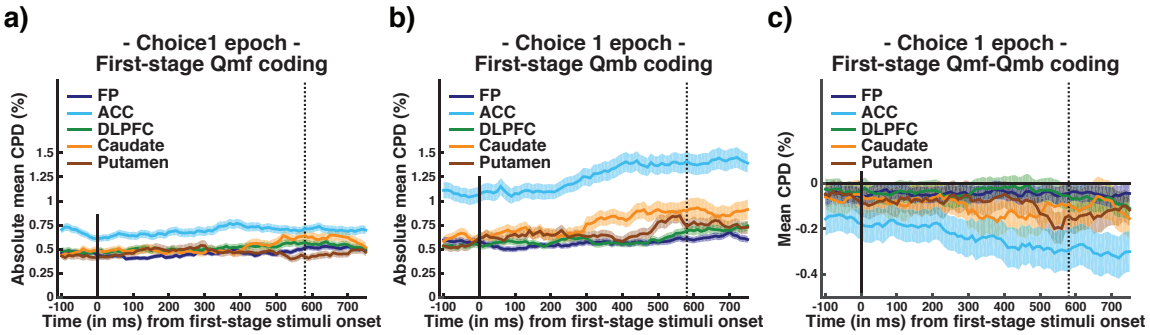


Fig. 5.13 **Population encoding of model-free and model-based action-values.** Coding at the population level, as determined by the absolute coefficient of partial determination (CPD) value at choice1 epoch, for first-stage **a)** model-free (Qmf) and **b)** model-based (Qmb) action-values. **c)** The balance between model-free and model-based action-value coding as determined by the 5% trimmed absolute mean (solid coloured line) and respective SEM (shading) across all neurons taking the CPD values for Qmf (positive sign by convention) and subtracting the CPD values for Qmb (negative sign by convention). Therefore, an averaged negative value indicates bias towards Qmb coding. Solid vertical line corresponds to the first-stage stimuli presentation; dotted vertical line represents the mean first-stage reaction time across subjects and sessions.

Neural representations of prediction error signals

Learning in both RL approaches is mediated by discrepancies in predictions, called prediction errors, which can be used to improve the value estimates and also to choose good actions. In our two-stage Markov decision task, two moments in time are particularly important for detecting prediction error signals: the transition epoch, because the state-transition type (common versus rare) has consequences for the expected value; and the feedback epoch, when the subject knows what level of outcome will be provided, and must update second-stage value estimates.

Prediction error signal at transition epoch

The two possible state-transitions from first to second-stage give rise to a first-stage prediction error signal – the value expected at the time of first-stage choice is compared to a value of the second-stage state at which the subject has just arrived. A good example of this is when subjects experienced on the previous trial a high outcome on a particular second-stage state, then on the following trial they choose the first-stage choice that most likely leads to that same second-stage state and are then faced with either a common or a rare transition. In this example, the common transition would not generate a significant error in value predictions (even if the choice behaviour of our subjects is a hybrid of both strategies, pure MF and pure MB also do not show such differences in this context), given that the subject will be able to pursue his initial chosen value expectation (assuming he chooses the appropriate second-stage stimulus). By contrast, if a rare transition occurred, there is a tendency for a greater error because the chosen expected value is less closely related to the value of the second-stage that the subject predominantly expected. In this case, the error would be stronger if predictions had been derived from a pure MF valuation; whereas MB predictions should not be as greatly affected, since they already include the 30% chance of such rare trials. Neurons that report such first-stage prediction error signals should considerably reduce their firing rates when a rare transition occurs in the above example (Fig. 5.14). This is because the value of the second-stage state that the subject actual transitioned to, is likely lower than the second-stage state more often associated with the chosen first-stage option.

Looking across regions at the average normalised firing rates of all neurons in the case described above, we found that activity in both caudate (Fig. 5.14d) and putamen (Fig. 5.14e) reduced for repeat-rare compared to repeat-common. Note also that the FP cells (Fig. 5.14a) responded in precisely the opposite manner: following a high outcome on the previous trial and a first-stage choice aiming to repeat the same second-stage state experienced on the previous trial, neurons increased their firing rate more for a rare transition than a common one. Given these effects in the average firing rate, we used multiple regression analysis for quantification (Fig. 5.15). We considered how the encoding of the previous outcome at the transition epoch was affected as a function of repeat-common, repeat-rare, switch-common and switch-rare. The results confirmed an abrupt negative correlation between the firing rate at transition epoch and the previous outcome coding for repeat-rare (when compared to repeat-common) in both caudate (Fig. 5.15d) and putamen (Fig. 5.15e), although the decrease was more pronounced in the former. On the other hand, and as expected, FP showed a rather positive correlation for the same condition.

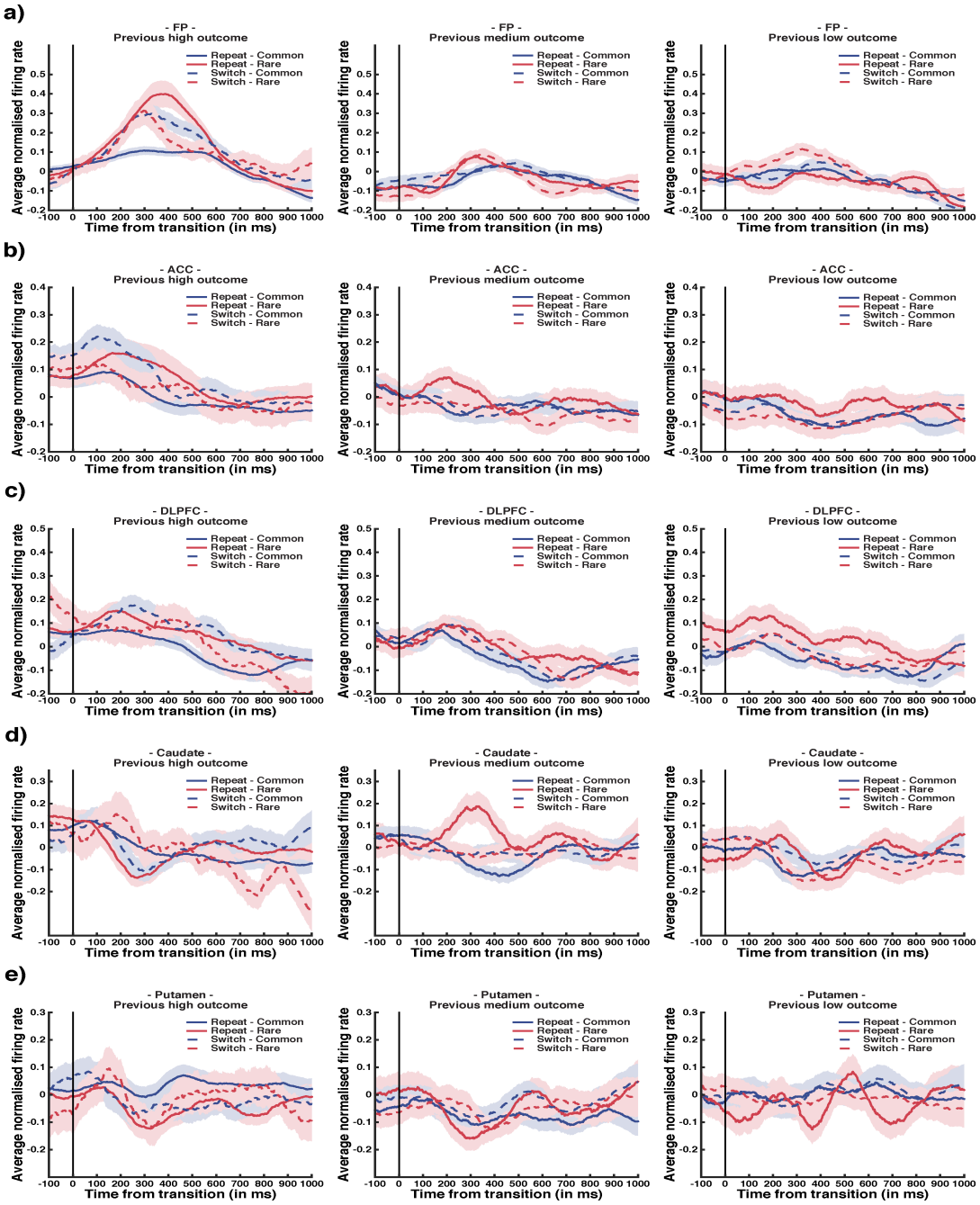


Fig. 5.14 **Impact of transition on neural activity of each region according to the previous outcome and the first-stage choice.** Average normalized neural activity (shading is SEM) at transition epoch across all neurons of each region according to the previous outcome level (left-column for previous high; center-column for previous medium; and right-column for previous low), the current transition (common or rare) and whether the first-stage choice was the one more ('repeat') or less ('switch') likely associated with the previous trial's second-stage state (i.e., 'repeat' or 'switch' the second-stage state where the previous reward was experienced). Solid vertical line corresponds to when the type of transition is revealed.

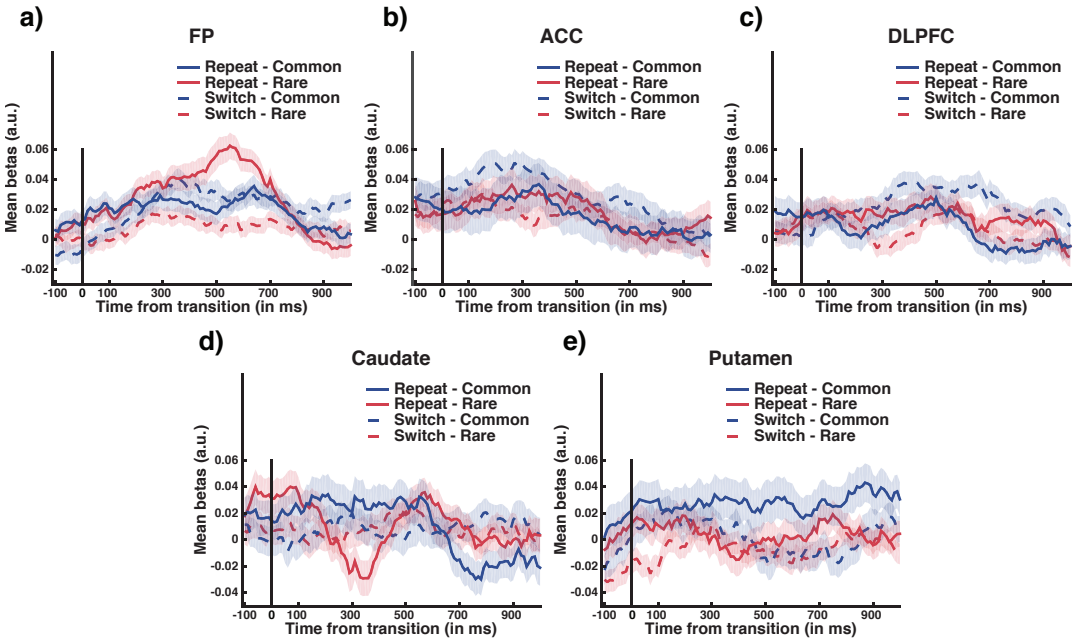


Fig. 5.15 **Population coding of previous outcome at transition epoch as a function of the first-stage choice and the transition type.** Mean value of the beta regression coefficients (shaded area depicts SEM) for reward on the previous trial at transition epoch across all neurons of each region, according to whether the transition was common or rare and whether the first-stage choice was the one more ('repeat') or less ('switch') likely associated with the previous trial's second-stage state (i.e., 'repeat' or 'switch' the second-stage state where the previous reward was experienced). Solid vertical line corresponds to when the type of transition is revealed.

We sought to confirm these results by taking advantage of the trial-by-trial estimates from our best-fit computational model. To link this approach with the results observed in Fig. 5.14, we investigated across regions how the firing rate at transition epoch varied as a function of the first-stage chosen action-value derived from our *HYBRID+* model and the current state-transition (Fig. 5.16).

The results are broadly consistent with our previous analysis. For high chosen first-stage values, a rare transition causes a significant reduction in the firing rate (which would be consistent with a negative prediction error) in both caudate and putamen, and a prominent increase in FP spiking activity. Note that a subtle latency difference was observed between the striatal regions. For the low value condition (right column of Fig. 5.16e) a rare transition increased the neural activity first in the caudate (peak between 200-250ms) earlier than in putamen (500-550ms). The combination of values and choices arises when a relatively low first-stage option is picked (for reasons that could include suboptimal behaviour), but with the rare consequence of going to the less likely second-stage state. This unexpected state could indeed be better in terms of outcome and may generate a response akin to a positive prediction error, hence the rise in firing rate.

In formal terms, the first-stage prediction error at transition epoch is the difference between the value of the second-stage choice about to be made and the first-stage chosen action-value derived by our *HYBRID+* model (as there is no reward associated the state-transition; see chapter 3 - Equation 3.5 of the Methods section). Therefore, a prediction error signal should correlate positively with the value of the second-stage choice and negatively with the expected first-stage chosen action-value in rare trials, once the transition is made known to the subjects. To confirm this, we performed a regression analysis having those two values as predictors of the firing rate at the transition epoch, and being assessed differently according to the transition type (Fig. 5.17).

The neural activity pattern in caudate at transition epoch was, once again, consistent with the features described above for a prediction error signal at rare trials: slight positive coding for second-stage chosen action-value and negative coding for first-stage chosen action-value (see Chosen1Qhyb-Rare and Chosen2Q-Rare in Fig. 5.17d). Interestingly, using the computational variables the same signal was weaker and less sharp in putamen (Fig. 5.17e). On the other hand, the coding scheme observed in FP, as expected, showed the inverse properties: encoded negatively the second-stage choice and positively the expected first-stage chosen action-value, and this was more prominent in rare trials (Fig. 5.17a). It is important to highlight that this prediction error only appears in rare trials, implying that this is not a general TD prediction error signal of dopaminergic neurons (Schultz et al., 1997).

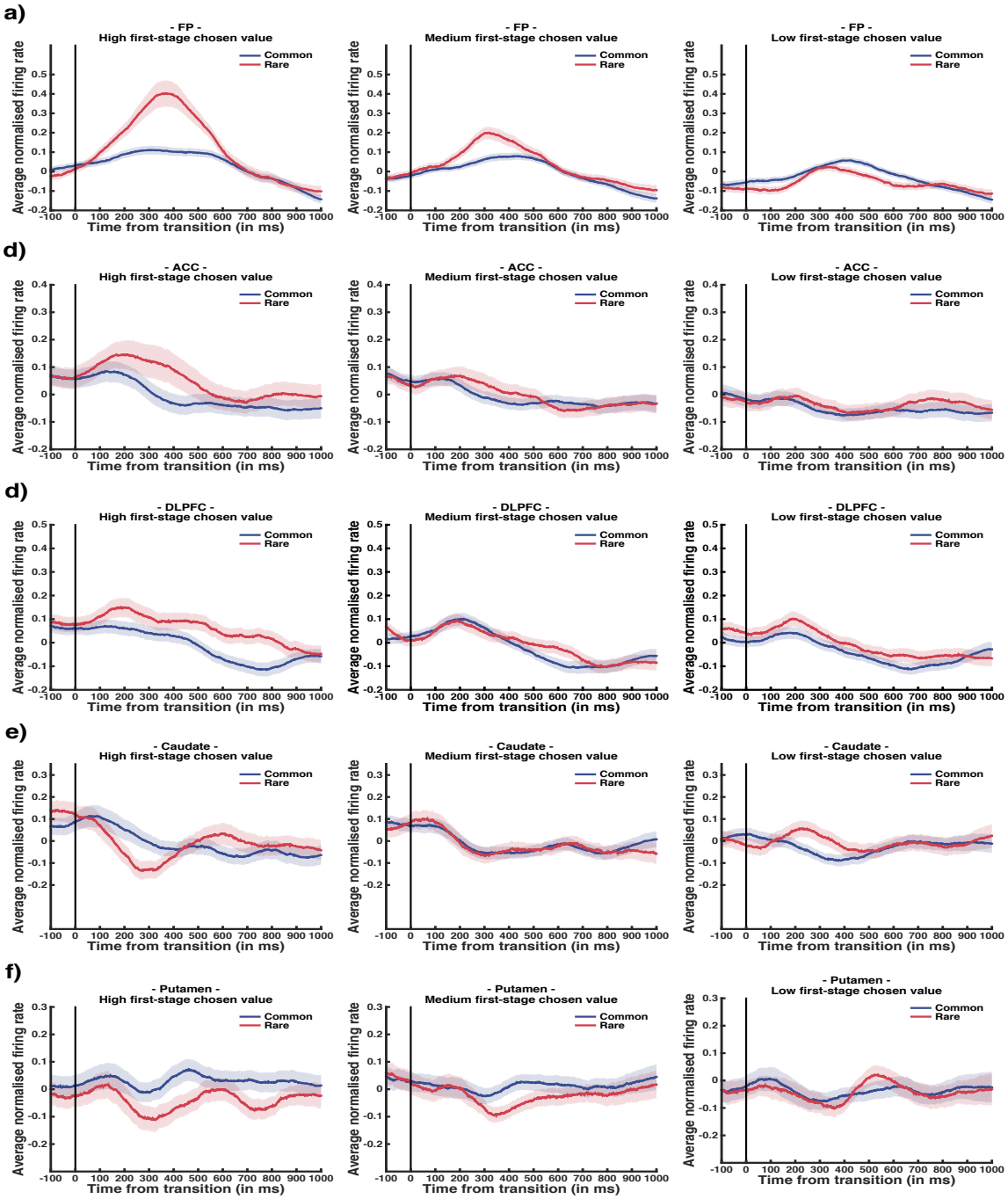


Fig. 5.16 **Impact of transition on neural activity of each region according to the previous outcome and the first-stage chosen action-value.** Average normalized neural activity (shading is SEM) at the transition epoch across all neurons of each region according to terciles of first-stage chosen action-value (left-column for high tercile values; center-column for medium tercile; and right-column for low tercile) and the current transition (common or rare). Solid vertical line corresponds to when the type of transition is revealed.

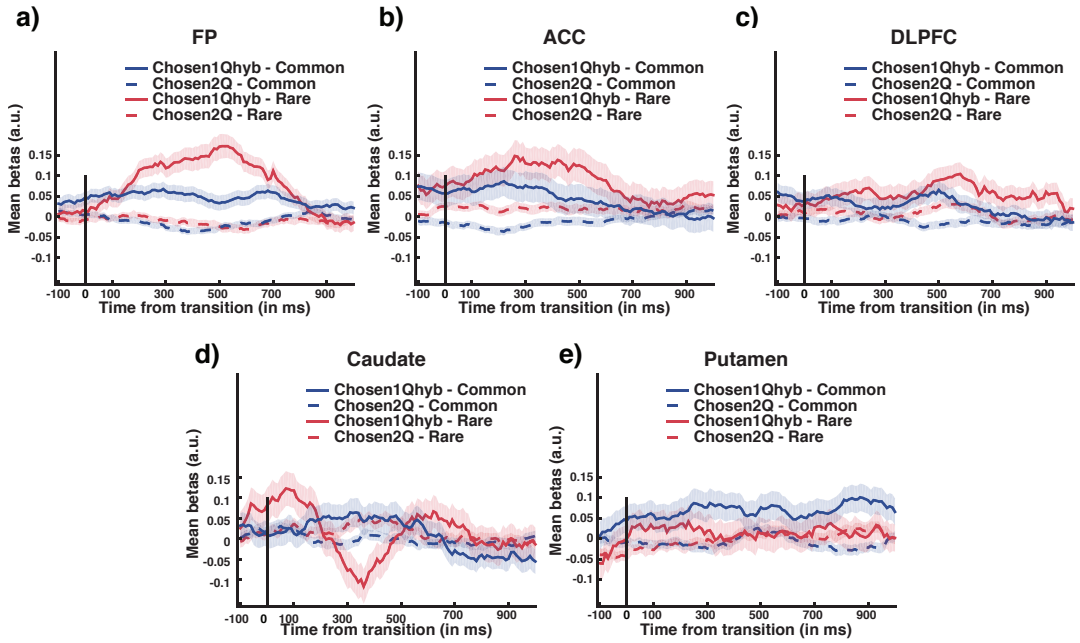


Fig. 5.17 **Population coding of first-stage chosen action-value and second-stage choice value at transition epoch as a function of transition type.** Mean value of the beta regression coefficients (shaded area depicts SEM) at transition epoch across all neurons of each region for first-stage chosen action-value (Chosen1Qhyb) and second-stage chosen action-value (Chosen2Q) derived by our *HYBRID+* model, according to whether the transition was common or rare. Solid vertical line corresponds to when the type of transition is revealed.

Second-stage choice reward prediction error signal in striatum

Second-stage learning is equivalent for both RL methods since there is no further stage to anticipate. In the model, it follows a temporal-difference (TD) learning rule for predicting the immediate reward at feedback. Neuronal activity reporting the necessary reward prediction error should encode the discrepancy between the current and the expected reward (Fig. 5.18).

We found that both caudate and putamen (Fig. 5.18d-e) regions phasically increased their neuronal activity at feedback epoch when outcomes were better than expected (right-column of Fig. 5.18), but reduced their firing rate if the outcome was worse than predicted (left- column of Fig. 5.18). Note some subtle differences between these two areas. First, receiving a medium outcome was encoded differently: in caudate the neural activity for a medium outcome level reflected a profile slightly more comparable to that of the worst outcome level; whereas in putamen the medium outcome level elicited a firing rate pattern closer to that observed after receiving the high outcome outcome. Secondly, when the upcoming reward was the low outcome and following the initial drop in neural activity (i.e., from about 400m post-secondary reinforcer onwards), neurons in putamen showed a marked and persistently increased firing rate throughout the remaining period of the feedback epoch. Such rebound effect was not observed in caudate, where the neural activity returned to baseline. Despite these differences, both patterns of response were qualitatively consistent with a reward prediction error, and were particularly distinct from other brain areas.

The FP cells presented a prominent response if a high reward was expected but the actual outcome was the lowest possible (left-column of Fig. 5.18a). On the other hand, if the expectation medium or low (center- and right-columns in Fig. 5.18a) FP neurons increased their spiking when either the highest or the lowest reward was received. A medium outcome did not elicit firing rate modulation in any expectation scenario. A relatively similar pattern of response, yet less phasic, was observed in ACC – increased neuronal activity if either a high expectation was followed by a low outcome (left-column of Fig. 5.18b) or a high/low reward was received when the expectation was not high (center- and right-columns in Fig. 5.18b). Nonetheless, in this latter scenario ACC encoded negatively the actually received outcome (i.e., more spiking activity for lower outcomes), something not observed in FP. Finally, in DLPFC neurons the only feature to highlight were the differences in latency for each outcome level actually received specifically observed in the low expectation scenario (right-column in Fig. 5.18c), with a firing rate reduction faster if the actual outcome was low and slower if the outcome was high.

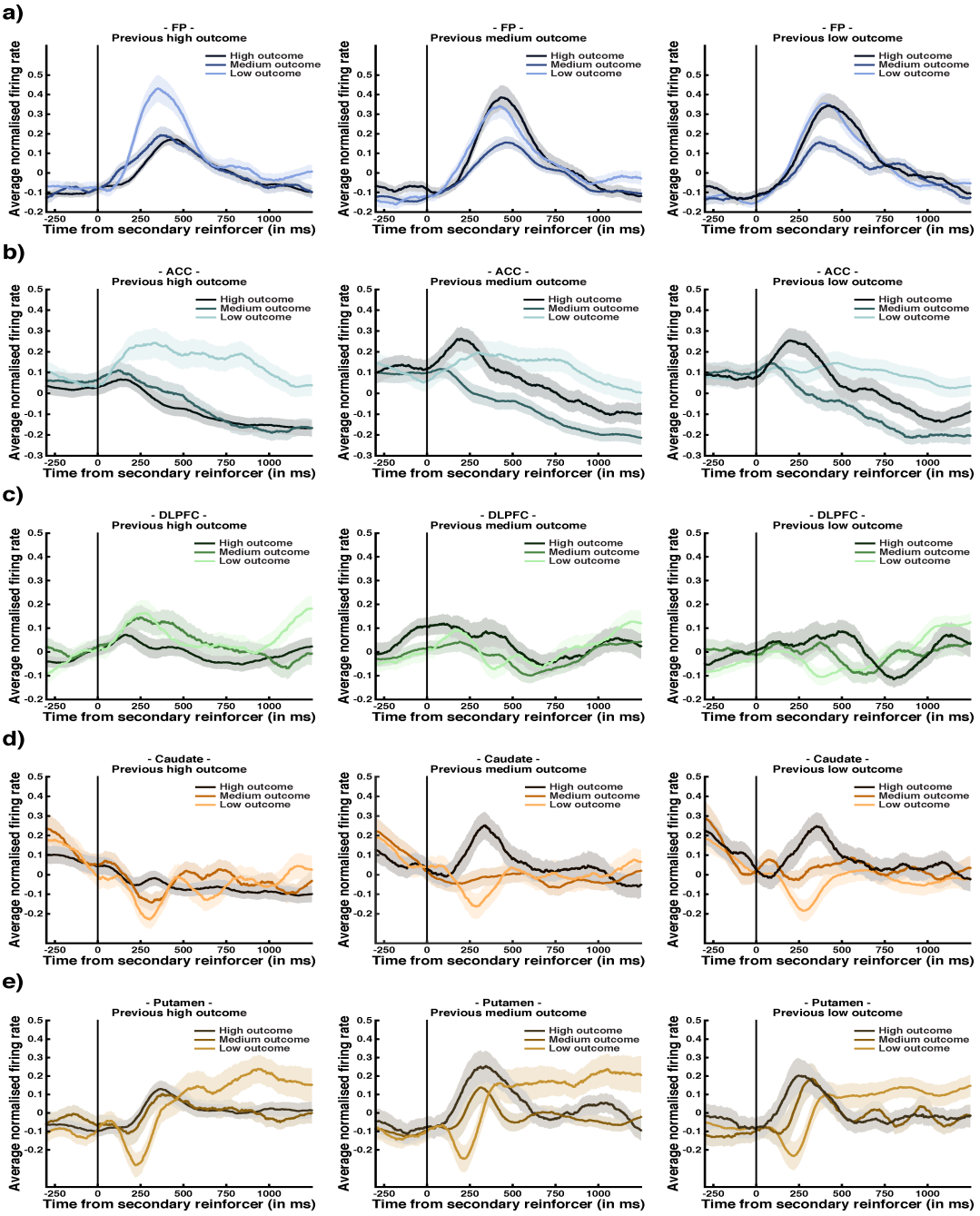


Fig. 5.18 **The neural activity across regions as a function of the expected and actually received outcome with the current second-stage choice.** Average normalized firing rates (shaded area depicts SEM) at feedback epoch across all neurons of each region according to the previous outcome level (left-column for previous high; center-column for previous medium; and right-column for previous low) experienced with the current second-stage choice and the outcome level (high, medium or low) about to be received with that same choice. Solid vertical line corresponds to secondary reinforcer presentation.

Following this, we used multiple linear regression analysis to quantify this involvement of caudate and putamen in second-stage value learning. Our aim was to confirm that the firing rate at feedback epoch of both areas was positively correlated with the outcome revealed by the secondary reinforcer and negatively correlated with the last reward obtained with the same second-stage choice (Fig. 5.19). In addition, if both striatal regions actually reflected a classical TD prediction error signal, the weights of the previous reward information should have decayed exponentially with trials into the past (Fig. 5.20) and resemble the phasic response observed in dopaminergic cells (Bayer and Glimcher, 2005). Indeed, we observed the above quantitative features in caudate and putamen, strongly favouring the encoding of the second-stage reward prediction error in these two regions.

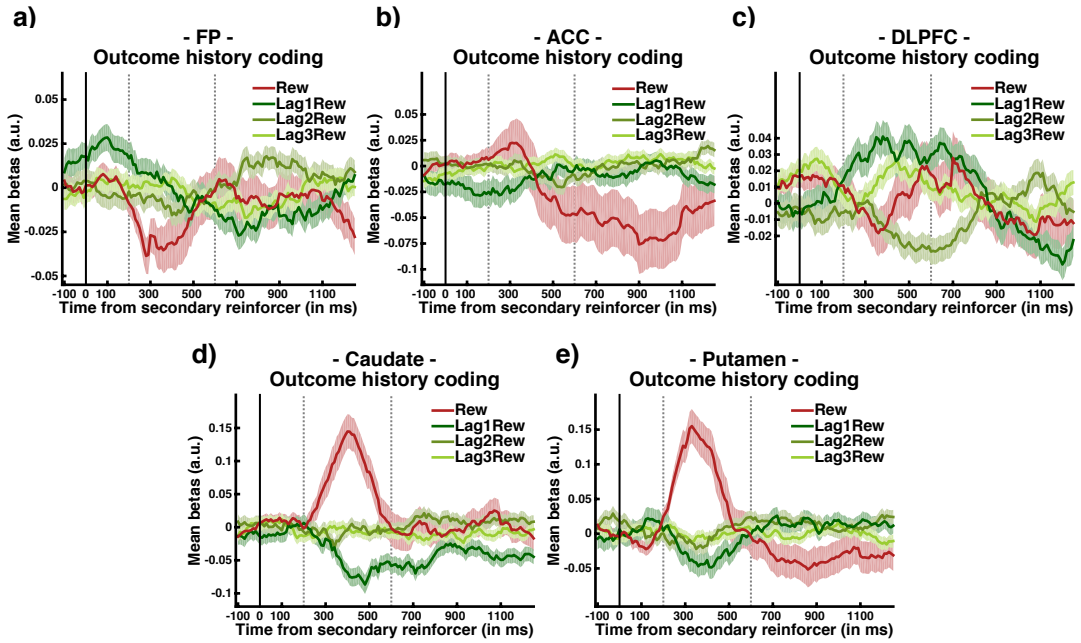


Fig. 5.19 **Population coding of reward history at feedback epoch.** Mean value of the beta regression coefficients (shaded area depicts SEM) at feedback epoch across all neurons of each region for the upcoming reward (Rew), the most-recent (Lag1Rew), the second-most recent (Lag2Rew) and the third-most recent (Lag3Rew) rewards obtained with the current second-stage choice. The Solid vertical line corresponds to secondary reinforcer presentation; dotted grey lines correspond to the period of time from 200 to 600 ms post-secondary reinforcer.

So far, we have used direct, task-based, measures to test the second-stage prediction error signal. To examine these results further, we took advantage of our model fitting and regressed the firing rate at feedback against both the upcoming reward and the computationally derived second-stage chosen value (Fig. 5.21). The results show that the signal in both

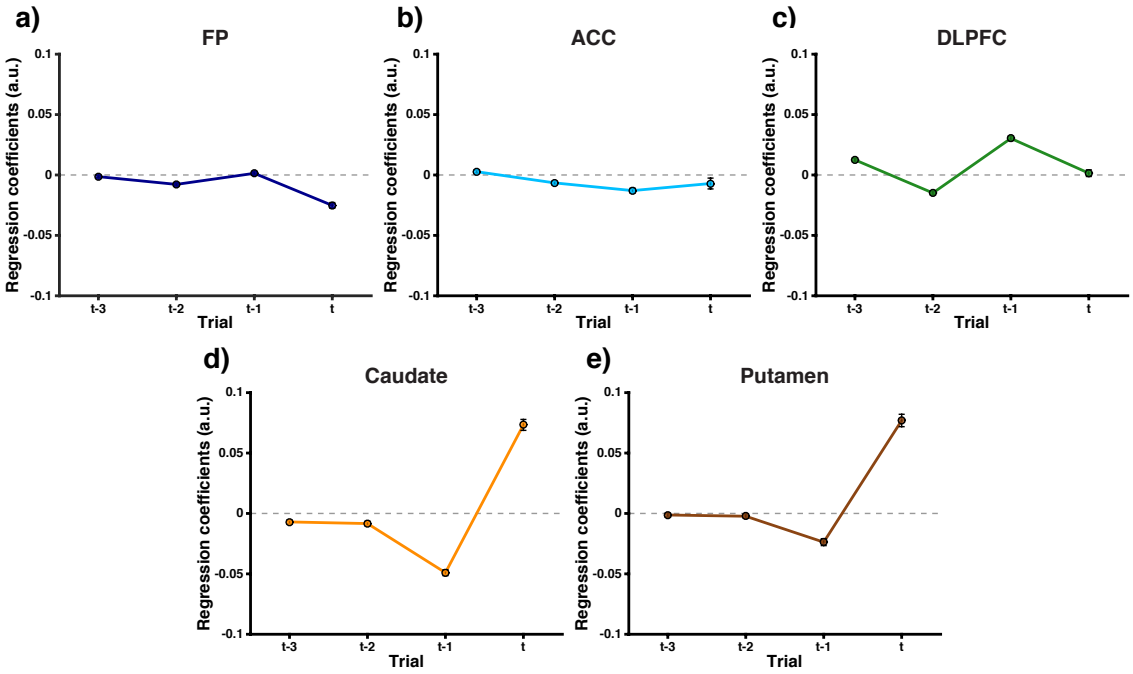


Fig. 5.20 **Reward history impact on the neural activity at feedback epoch across regions.** Mean (and SEM) values of the beta regression coefficients at the period of time of the feedback epoch between the two dotted grey lines in Fig. 5.19 (corresponding to 200-600 ms post-secondary reinforcer) across all neurons of each region, for the upcoming reward (t), the most-recent (t-1), the second-most recent (t-2) and the third-most recent (t-3) rewards obtained with the current second-stage choice.

caudate (Fig. 5.21d) and putamen (Fig. 5.21e) followed the pattern expected for a reward prediction error signal: positive correlation with upcoming reward and negative correlation with second-stage chosen value. Finally, we directly correlated neural activity at feedback against the computationally derived reward prediction error at second-stage choice and also found the expected positive correlation in both caudate and putamen (Fig. 5.22). These results provide more robust evidence linking the neural activity at feedback of both caudate and putamen with the second-stage reward prediction error expected by our RL modelling. Furthermore, they also show that the error signal arises slightly earlier in putamen than in caudate.

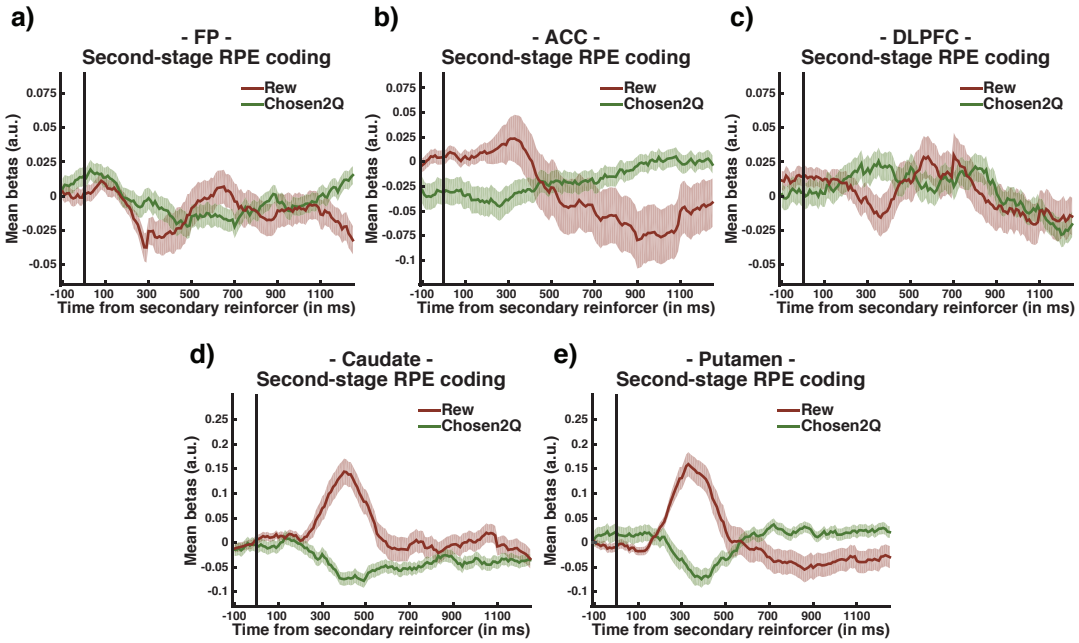


Fig. 5.21 **Population coding of expected second-stage value and the actual reward received at feedback epoch.** Mean value of the beta regression coefficients (shaded area depicts SEM) at feedback epoch across all neurons of each region for the upcoming reward (Rew) and the chosen second-stage action-value (Chosen2Q) derived by the best-fit computational model. Solid vertical line corresponds to secondary reinforcer presentation.

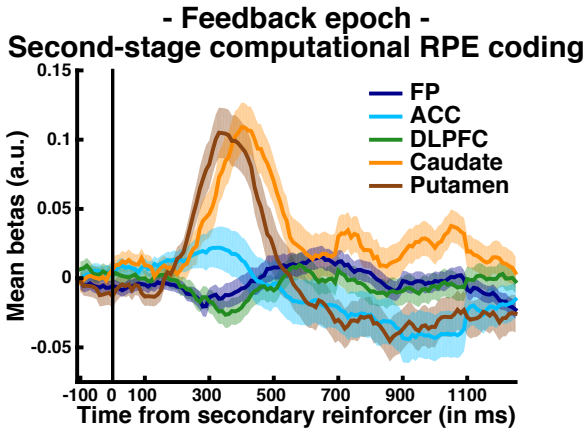


Fig. 5.22 **Simple linear regression on neural activity at feedback across regions for the second-stage reward prediction error.** Mean value of the beta regression coefficients (shaded area depicts SEM) at feedback epoch across all neurons of each region for the trial-by-trial second-stage reward prediction error values derived by the best-fit computational model. Solid vertical line corresponds to secondary reinforcer presentation.

5.5 Discussion

We recorded single-neuron activity from regions in the prefrontal cortex (FP, ACC and DLPFC) and in the dorsal striatum (caudate and putamen) of two non-human primates while they performed a sequential decision task which induced trial-by-trial adjustments in choice that combine both MF and MB-RL control. The striatum is the main input structure of the basal ganglia for both cortical as well as dopaminergic projections, and is believed to participate in the selection of rewarded actions by encoding the value of the available choices. On the other hand, lesion studies in humans and animals imply that PFC areas make fundamental and specialized contributions to optimal reward learning and decision making. Using an analysis approach that took into account both observable behavioural variables and computational measures derived from RL models of the task, we have reported experimental data which demonstrates a robust link between neural signals and key components of RL at different moments in the task.

We follow the organization of the results section, separating the discussion according to the three major epochs of the task (reward/transition/choice; the transition from first- to second-stage; the observation of the secondary reinforcer), and then according the key brain regions: the ACC, the FP, the caudate and putamen and the DLPFC. We consider both behaviourally- and model-defined correlates.

Reward, transition and choice coding in the prefrontal cortex and basal ganglia

A key characteristic in RL is the ability to use past experience to select the best course of action from among competing alternatives. More than just simply having representations based on reward history (as in MF valuation), the knowledge of how actions influence the transitions among different states (as in MB valuation) may be critical to flexible decision making when faced with environmental changes. With this in mind, we highlight our main findings about the single-neuron and population correlates of the most recent reward, state-transition and first-stage choice. We then assess the implications for the structures that are thereby highlighted.

First, despite the ubiquitous coding of the upcoming reward (in our case, the value of the attended secondary reinforcer stimulus) throughout the brain (Apicella et al., 1991; Kennerley and Walton, 2011; Kennerley et al., 2009; Paton et al., 2006; Platt and Glimcher, 1999; Roesch and Olson, 2003), correlates of reward magnitude were found to be significantly more common and stronger in ACC, but faster to emerge in putamen and caudate. Regional differences in the overall population code were also observed, with caudate and putamen neurons firing more to higher rewards, but ACC and FP neurons responding more for lower outcomes. The information about the most recent reward remained persistently high in ACC from feedback until the next trial's first-stage choice. Second, more ACC neurons encoded the type of transition; with the population selectivity remaining high from the moment the state-transition occurred until the moment the feedback was revealed. Moreover, individual neuron's selectivity for the transition type spanned both transition and feedback epochs. Finally, we looked at factors related to first-stage choice and found that among areas, it was in the ACC that the explanation of neuronal activity at choice epoch by all types of first-stage stimulus information: previous choice, current choice and whether previous and current were the same (i.e., repeat or stay first-stage strategy) was most proficient. On the other hand, the putamen more prominently encoded the first-stage side of response, consistent with its well known involvement in movement preparation and action execution (Romo et al., 1992; Schultz and Romo, 1992).

In addition to these additive effects, we also presented neural evidence of combined effects of these different behavioural factors. This analysis provided a more detailed description of the computations performed by single-neurons as well as of the various populations. First, we found that the strength of the transition coding at the transition epoch was positively correlated with the (positive) reward coding in striatum and negatively correlated with

the (negative) reward coding in the FP at feedback. Here, strength includes increases or decreases of neural activity with the rare transition. Concomitantly, the strength of reward coding (again ignoring the direction of firing rate changes) was positively correlated with the (directional) transition coding at the transition epoch in the FP. These findings suggest that in FP the greater the encoding of the transition the more negative is the encoding of reward later in the trial, and the greater the increase in firing rate for rare transition (compared to common) the more reward information is coded at feedback. Second, ACC revealed a particularly strong additive effect of both reward and transition at feedback. And finally, the way reward was coded at the population level at feedback epoch was different according to the transition type in FP.

Probably the most noticeable feature of the above results was the multiple encoding of relevant learning variables by ACC compared to other regions. The link between reward and transition coding is particularly critical for MB valuation, which exploits a model of the structure of the environment to compute the cumulative reward. A subpopulation of neurons in ACC may support this link by multiplexing both reward and state-transition computations at feedback. The involvement of ACC in reward-guided learning and decision-making has been extensively documented using either lesion, neurophysiological or functional neuroimaging work (Alexander and Brown, 2011; Hayden et al., 2011b; Kennerley and Walton, 2011; Kennerley et al., 2006; Matsumoto et al., 2007; Quilodran et al., 2008; Rushworth et al., 2011; Rushworth and Behrens, 2008; Shenhav et al., 2013). The key computations more consistently associated with ACC across the different types of studies seem to be reward monitoring and the coding of choice-reward associations in non-stationary or foraging environments (Behrens et al., 2007; Hayden et al., 2011b). Furthermore, the ACC feature of multiplexing at the single neuron level different types of decision variables, for example value and choice, has been also shown by others in a context of a pure decision making task (Hayden and Platt, 2010; Kennerley et al., 2011, 2009).

The bias towards a negative net reward coding (i.e., increased firing rate for lower outcomes) in ACC corroborates previous studies, which reported more ACC neurons responding to errors than to rewarding feedback (Quilodran et al., 2008). It is also in line with several event-related potential studies that have consistently localized an ACC component called the error-related negativity, as it is typically more negative after participants make incorrect responses, compared to correct choices (Gehring et al., 1993). However various other factors could be important. One is that this error-related activity is confounded by the fact that errors happen less frequently than rewards (given that subjects are often overtrained) and, indeed, when both are counterbalanced the neural sensitivity to reward

and unrewarded (error) outcomes seem to equalise (Jessup et al., 2010; Kennerley et al., 2011). Given that in our task there were no errors per se, but instead reward reassignments, a second hypothesis is to relate our ACC feedback-related activity with a potential role in directing behaviour towards goals or in foraging guidance (Rushworth et al., 2012). In fact, the error-related activity observed in ACC is modulated by the amount of predicted reward and by external signals indicating the necessity to shift response, which challenges the idea of a mere error detection signal and rather suggests a role in predicting the likely outcomes of actions (Alexander and Brown, 2011; Amiez et al., 2005). Third, a wealth of evidence has related the ACC activity with the ability for internally maintained goals to overcome prepotent or stimulus-driven responses – or cognitive control, as it happens in Stroop test (see Shenhav et al. 2013 review). Such control is more often required in difficult situations when automatic responses could lead to bad outcomes. In our case, a low outcome is often associated with subsequent demanding decisions involving a cost-effectiveness evaluation in order to maintain the goal of searching for the best long term reward possible. Fourth, a high outcome is slightly more instructive in the task than a low outcome level (i.e., with a high outcome subjects aim to repeat the same second-stage choice, whereas low outcomes may not necessarily mean to avoid a particular second-stage state, as there is another option where a good outcome could be found). Thus low outcomes are also associated with more uncertainty about the outcomes of his future choices.

Our interpretation of the present outcome-related ACC results is also closely related to recent recordings from rat ACC (Karlsson et al., 2012), where a radical shift in the pattern of ACC activity occurred at the start of a period of exploratory behaviour (i.e., when an exploitative period ended in favour of a *knowing nothing* period) rather than during the acquisition of new information per se, suggesting that the ACC was active at the point at which estimation uncertainty increased. It also fits well with human neuroimaging observations that ACC is more active in response to new observations when there is higher uncertainty of the outcomes provided by the environment (Behrens et al., 2007; Rushworth and Behrens, 2008).

The psychological concept of cognitive control linked to ACC function mentioned above, entails leveraging higher-order representations or contextual information to overcome what could be viewed as habitual actions (Braver, 2012). This yields, therefore, compelling similarity with the MB-RL state-transition knowledge and could be related with the prominent transition coding observed here in ACC. In fact, human subjects performing a similar two-stage decision task with higher MB weight in their choice, also had higher values of goal-directed behaviour across different cognitive control tasks (Otto et al., 2014), supporting

this correspondence. Some computational cognitive control theories for ACC use the RL framework and, although not clearly stated, their implementation has an implicit MB-RL role for ACC (Alexander and Brown, 2011; Holroyd and Yeung, 2011). The Predicted Response Outcome (PRO) model proposed by Alexander and Brown (2011) is one theory that has managed to reproduce most ACC results found in cognitive control tasks. Broadly, it detects if an expected outcome fails to occur at the expected time (a negative surprise reported by an unmet prediction error signal), and does this by monitoring the error likelihoods of all possible outcomes. This latter knowledge of all possible outcomes, very much resembles the MB-RL awareness of the entire decision-tree. Moreover, in a recent human neuroimaging study with a saccadic planning task directly assessing information theoretical measures, the activity in ACC was specifically correlated with trials on which an internal model was updated (O'Reilly et al., 2013), in line with our observations of an ACC involvement in state-transition coding.

Our findings on the FP are particularly novel. We found a FP population bias towards a transient negative relationship between reward magnitude and firing rate at feedback as well as an increased activity for rare transitions at transition epoch. Although this coding was generally weak, such a pattern of responding in the population of FP neurons seemed relatively homogeneous and specific to factors that are key for updating the behavioural strategy. Low outcomes are, as mentioned earlier, associated with a subsequent exploratory period where subjects aim to find the highest outcome, and rare transitions are particularly relevant when interpreting the context of the received outcome. Consistent with this, FP not only combined both reward and transition information from different task epochs, but it also showed different encoding of reward at feedback depending on whether the current transition was common or rare. The overall conclusions from recent FP lesion studies highlight its particular involvement in rapid learning and in context-dependent allocation of cognitive resources, which is overall supportive of the short-lived FP activity profile we observed. The only other study to record from single-cells in the primate FP found surprisingly simple neuronal responses (Tsujimoto et al., 2010, 2012). In that study the monkey was cued to repeat or switch the choice of the previous trial and a response signal in FP was found around feedback time to encode the response that was correct according to the cued strategy. The authors considered that this could promote learning by associating outcomes with responses based on some memory, rule or stored representations. Others further emphasised this result as a forward implementation of internal models relevant for the task (Koechlin, 2011), with particular advantages in exploratory contexts (Daw et al., 2006).

Concerning the findings in other regions, the earlier coding in striatum of the reward

information revealed by the secondary reinforcer is not surprising given the robust corticostriatal projections from extrastriate visual cortex (Saint-Cyr et al., 1990). The population bias towards a positive reward coding is also consistent with previous findings in other reward learning contexts (Schultz et al., 2003), which also reported more striatal response for increasing reward expectations. Of interest, a recent study reporting neural responses to objects frequently changing their value (as in our task), also found a positive value for the difference between striatal neurons responding more strongly to high-valued objects and to low-valued objects (Kim and Hikosaka, 2013).

We were surprised by the apparent absence of coding in the DLPFC for specific reward, transition or choice. The disruption of dorsolateral prefrontal cortex with transcranial magnetic stimulation (TMS) in humans performing a similar two-stage task impaired MB behaviour in favour of behaviour driven by MF control (Smittenaar et al., 2013). Furthermore, previous recordings in primate DLPFC, while subjects sought an optimal choice, found that single neurons' firing rates changed with both choice and reward history (Barraclough et al., 2004; Seo et al., 2007). We believe that the incongruity between our findings and these other studies could be related to particularities of our experimental design. First, note that the single-neuron results from others have used tasks requiring saccade-contingent choices, whereas the current study required a joystick movement to indicate the decision. Therefore some discrepancy might be expected given the DLPFC's strong connection to supplementary eye fields (Huerta and Kaas, 1990). Unpublished results from several studies in our lab indicate that neuronal selectivity in ACC and DLPFC increase when choices are made by forelimb (joystick) or eyes, respectively (c.f., Kennerley et al., 2009). In any case, this cannot explain the TMS result. For this, we believe that the subjects in the human study may have been more reliant on working memory than our subjects, who had extensive exposure to the task, and that the TMS disrupted the working memory, rather than MB function of DLPFC. Our best best-fitting behavioural MB-RL model for both subjects involved the state-transition probability distributions being known from start of the sessions (see chapter 3 for details). By knowing the task structure well, our subjects could reduce the requirement on working memory associated with state-transition learning, two processes known to elicit DLPFC activity (Funahashi et al., 1993; Gläscher et al., 2010; Otto et al., 2013; Watanabe, 1996). In fact, as humans increase familiarity with a similar two-stage task as the one used here, model-based reasoning seem to depend less on such executive processing (Economides et al., 2015). At the computational level, any such enhanced efficiency in the implementation of MB reasoning could involve a progressively MF implementation of MB reasoning via representational change (Akam et al., 2015). At the

implementational level, it may recruit long-term or episodic memory systems, such as the hippocampus, which have also been involved in MB learning by keeping active sequential representations or paths of the environment (Bornstein and Daw, 2012; Johnson and Redish, 2007).

Model-free and model-based value representations for action selection

Learning action values, i.e. the mapping from state-action pairs to expected return, is key to action selection in RL. This mapping can be MF or MB; the resulting action values represent the long-run desirability of actions, and thus are the basis of appropriate choice (Sutton and Barto, 1998). Complementing and extending previous reports of MF action-value encoding in the primate brain (Lau and Glimcher, 2008; Samejima et al., 2005), we found single-neuron representations of MB action values. Despite the multiple views on MB (or goal-directed) behaviour (Daw and Dayan, 2014; Doll et al., 2012), this is, to our knowledge, the first evidence of single-neuron representations of MB action values, as formally defined in RL. Another of our novel results was the discovery in both prefrontal and striatal regions of neurons that covaried with both MF and MB action-values during the time of first-stage choice. Further, cells selective only for MF or MB value predictions were found to be widespread and richly interdigitated across the different neuronal populations. All these findings sit comfortably with the hybrid approach found to fit the choice behaviour the best (see chapter 3), albeit without using the overall hybrid weight that we identified behaviourally. They also shed light on contemporary controversies surrounding the neural mechanisms of the action selection process (Rushworth et al., 2012).

Two divergent ideas have recently attained prominence about the potential mechanisms and pathways involved in value-based decision making circuitry, suggesting either a parallel or serial operation (Cisek and Kalaska, 2010; Padoa-Schioppa, 2011). One view proposes that prefrontal and subcortical regions may compare options in different frames of reference, and that these computations can take place in a simultaneous and parallel manner during decision making (Cisek and Kalaska, 2010). According to this hypothesis, a choice emerges through a distributed consensus between regions, although the decision signal may emerge first in one area depending on its advantageous specialisation for the particular task. In the alternative model, decisions are made solely through a serial pathway where different valuations of the options are represented in distinct parts of the brain leading to a final single utility measure capturing the subjective value of each option (Padoa-Schioppa, 2011). The fact that in our study prefrontal and striatal areas showed relatively similar representations

at choice time of both MF and MB valuations of the available options, lends support to a parallel organisation for the action-selection circuitry. It is important to note that some parallel views (Cisek and Kalaska, 2010) intend for prefrontal cortex and striatum to compute different things, not the same thing, and we also do not know if these are not just part of a single calculation. Nevertheless, the higher prevalence of hybrid value neurons in ACC together with its early latency in MF and MB population coding, specifically implicate this structure in integrating MF and MB values in order to guide optimal behaviour.

The proposal that the ACC is involved in competition-related processes of monitoring as well as in resolving response error and conflict, fits well with its apparent involvement in numerous cognitive processes including behavioural flexibility (Hayden and Platt, 2010; Quilodran et al., 2008), action-outcome prediction (Alexander and Brown, 2011), environmental volatility (Behrens et al., 2007; Kolling et al., 2012) or more broad concepts of conflict (van Veen et al., 2001) and cognitive control (Shenhav et al., 2013). In a recent human study in which the task design favoured either MF or MB control, a comparison signal was also found in ACC reflecting the difference in the uncertainty of the estimates derived by each valuation strategy (Lee et al., 2014). In fact, the connectivity profile in resting-state functional MRI of the primate ACC region in which our neurons have been recorded, closely resembles the profile of the human anterior rostral cingulate zone implicated in all these studies (Neubert et al., 2015). Furthermore, the observed mixed action-value coding of ACC at choice time goes well with its access to both reward history (of MF and MB importance) and transition information (of unique MB relevance) discussed earlier, and with its well known multiplexing qualities (Hayden and Platt, 2010; Kennerley et al., 2009). In conclusion, our ACC results suggest that several ideas about this brain region could potentially be unified into a more normative framework of study based on clear computational features of RL theory.

Our task was *stimulus*-based, in that values were associated with stimuli, and actions were directed at them too. Thus what we have called action values might instead have been called stimulus values. Various other tasks are *action*-based, with values being associated with such things as the side of a joystick movement Samejima et al. 2005 or a saccade Lau and Glimcher 2008. Outcome-related values associated with specific stimuli prior to decisions, and independent of the nature of the movement performed (referred by some authors as ”offer-value” neurons) have previously been described in the orbitofrontal cortex (Padoa-Schioppa and Assad, 2006). These neurons were found not in a learning context, but while monkeys chose between two overlearned stimuli associated with different types of reward offered in different amounts. Although the values used in such economic decision-

making task could be viewed as cached values (and in this sense could be considered MF action-values), both MF and MB valuation coincide for that specific choice and do not allow any clear distinction between the learning strategies used. Finally, note that some evidence suggests that ACC is specialised to represent value according to the *action* required, whereas orbitofrontal cortex does so in according to the *stimuli* concerned (Kennerley et al., 2006; Rudebeck et al., 2008). However, other authors have also documented that disrupting the ACC interferes with learning about stimulus-outcome associations Amiez et al. (2006), in line with our findings showing representations of action-values for stimuli.

Our observations were in general supportive of the idea that some striatal neurons encode action-values derived purely from a MF-RL algorithm (Lau and Glimcher, 2008; Samejima et al., 2005). The proportion of such neurons we found (22%/19% in caudate/putamen) was relatively smaller than those previously reported (43% in Samejima et al. 2005, 62% in Lau and Glimcher 2008). Furthermore, we also found significantly fewer neurons (7%/5% in caudate/putamen) whose activity covaried with MF action-values and also with the chosen option (in our case the chosen first-stage stimulus), compared with other studies (22% in Samejima et al. 2005, 31% in Lau and Glimcher 2008). Finally, a relatively high percentage of putamen MF action-value cells also correlated with response side and reaction time – suggesting that the involvement of this region in selecting and executing the adequate motor plan (as it encoded strongly the first-stage response side) could be potentially modulated by the MF action value of the available options.

When comparing our data with these two striatal studies, several methodological considerations may be important. First, our task was stimulus-based and, as a consequence, it requires a more allocentric representation of the choice. This may be more demanding than a more natural hand movement or saccade. This might help explain the smaller number of MF action-value cells that we observed. Second the location of recordings varied: Samejima et al. 2005 recorded in the more posterior putamen and in dorsal parts of the caudate nucleus; Lau and Glimcher 2008 recorded in the caudate nucleus only; and our recordings were restricted to more anterior regions of dorsal caudate and putamen. Third, those studies restricted their analysis to neurons that were active during the task (Samejima et al., 2005) or selected cells based on firing activity (phasically active neurons) and responsiveness in a prior saccade task (Lau and Glimcher, 2008). We made no distinction between tonic or phasic striatal neurons and did not select neurons based on their responsiveness, in both cases to try and collect a representative sample of the overall activity across the primate dorsal striatum. Finally, some of the cells reported in those studies could also be MB action-value cells because both learning strategies coincide in their tasks. It is interesting to see that if

one takes into account MF only and MB only cells (59% in caudate and 44% in putamen), the total proportions are quite similar to those observed in the previous studies. Hybrid cells were excluded from this, since those previous studies did not engender conflict between MB and MF values.

However, MF action value neurons were not restricted to striatum. Many were also found in prefrontal regions, although the population representation of the MF value was not strong. The ACC offered the strongest code for MF values, and indeed exhibited this even before presentation of the first-stage choice. This involvement of ACC in MF-RL has received some support from previous literature. From both lesion and electrophysiological data, the modulation found in this area by the history of actions and outcomes together with its involvement in error monitoring has conveyed a functional link with the dopaminergic prediction error signal and with MF behaviour (Amiez et al., 2005; Kennerley and Walton, 2011; Kennerley et al., 2011; Matsumoto et al., 2007). Given the strong anatomical connections between ACC and VTA/SNpc it is no surprise that ACC and dopamine computations share some similarities (Williams and Goldman-Rakic, 1998). There are, however, some caveats. First, most previous studies used either simple learning tasks or focused on the action selection process with values learnt from direct experience, which makes it impossible to distinguish MF from MB valuations. Second, the value- related signals observed in ACC include complex feedback information (Quilodran et al., 2008), as it encodes information about many aspects of decisions (Kennerley et al., 2011, 2009) and outcomes that have been observed but not directly experienced (Hayden et al., 2009). Such elaborated reward signals are less likely to be part of a conventional MF calculation.

Neurons selective for MB values were also widespread in the prefrontal cortex and striatum. It was notable that the area with the highest relative percentage of such cells was the caudate; however, at the level of the whole population, the ACC dominated. The part of the caudate in which most of our recordings were performed is anatomically the region with highest afferent projections from ACC (Haber and Knutson, 2009). In addition, the more anterior regions of caudate seem to be not only necessary for (if lesioned) but also capable of modulating (if stimulated) the acquisition of new associations in instrumental tasks (Miyachi et al., 1997; Williams and Eskandar, 2006), in line with theoretical proposals suggesting greater MB-RL involvement at the start of the learning process (Daw et al., 2005). Previous neurophysiological studies also documented greater activity in anterior caudate for the initial phases of learning, as well as more flexible value coding, whereas the more posterior primate striatum has shown greater responses for over-learned motor sequences and more stable value coding (Belin et al., 2009; Kim and Hikosaka, 2013; Miyachi et al., 2002;

Pasupathy and Miller, 2005). Furthermore, other authors have highlighted a role in sequence representations or in integrated information comparison (Cai et al., 2011; Seo et al., 2012), processes more often linked with MB valuation.

The apparently limited selectivity of neurons in the DLPFC for action values might seem unexpected. Previous work has reported that DLPFC encodes the value of chosen items during early phases of decision-making prior to the transformation to the actual response (Cai and Padoa-Schioppa, 2014; Kim et al., 2008). It has also been demonstrated that DLPFC neurons encode strategies and rules in various types of tasks (Asaad et al., 2000; Genovesio et al., 2005; Wallis et al., 2001; White and Wise, 1999). However, in all these other studies a saccade was used to indicate choice, by contrast with the hand-movement required by our subjects. Perhaps more fundamental is the observation that lesions to the DLPFC do not seem to cause severe action selection deficits, unless the task requires a behaviour determined by a rule (Baxter et al., 2008; Buckley et al., 2009). Furthermore from a functional point of view, the role of DLPFC seems to extend beyond the maintenance of short-term memories to be crucial in guiding attention towards behaviourally relevant information (Buckley et al., 2009; Kadohisa et al., 2013; Lebedev et al., 2004), rather than a specific role in decision making per se.

First-stage reward prediction error encoding

A key characteristic of the design of the task is that the prediction about the future reward is almost certain to change as a result of the stochastic transition from first- to second-stage state. This engenders a temporal difference prediction error, which is used by the MF system to update the prediction of the first-stage stimulus that was chosen. This is the case even if the choice itself depended on both MB and MF values. In principle, it is possible to calculate the prediction error using either MB or MF assessments of value. Note, though, in the task, that these two assessments coincide at the second stage, rendering part of the difference moot. Further, the MB system does not use the prediction error for learning, basing its valuation instead on the state-action-state transition probabilities, and the history of outcomes (Gläscher et al., 2010).

It has been previously shown that this first-stage prediction error signal exists in humans performing the same two-stage task (Daw et al., 2011). Unexpectedly, the BOLD signal in both ventral striatum and medial prefrontal cortex reported a prediction error based on a mixture of MF and MB evaluations, in proportions matching those that determined choice behaviour for the individuals concerned. Such a hybrid was unexpected in ventral striatum,

since this is the favoured location for BOLD correlates of MF prediction errors (McClure et al., 2003; O'Doherty et al., 2003; Seymour et al., 2004). It was also unexpected for medial (and ventromedial) prefrontal cortex, as these have been more closely associated with MB evaluation, given the sensitivity to task contingencies shown by their value-related signals (Hampton et al., 2008, 2006; Valentin et al., 2007). Our data also suggest the concomitant existence of MB and MF prediction errors during performance. Neural activity in both striatal regions, although slightly more consistent and faster in caudate than putamen, showed features of an error signal useful for the learning purposes. On the other hand, the unique pattern of response from FP neurons strongly favour the involvement of this region in reporting planning prediction errors. Daw et al. (2011) speculated that their results implied a more interactive relationship between the learning strategies, with temporal difference prediction errors calculated using MB values coexisting with more conventional MF ones, and being used to optimise behaviour either online or through learning. This could operate through the actor-critic algorithm (Barto et al., 1983, 1995), with an integrated MF and MB critic training a combined actor.

One of our most novel findings was the single-neuron FP signal observed at the time of transition. The FP firing rate increased to a greater extent for higher first-stage chosen values given a rare transition than given a common one. In other words, the FP neural activity encoded a quantity akin to the foregone expected value or *regret*, because the rare transition would prevent the subject from pursuing the value its first-stage choice might have been expected to enable. These findings extend previous single-neuron reports of FP activity in a non-feedback related epoch, and are also supportive of a role for the FP in counterfactual choice and sequential decision making, as has previously been proposed (Boorman et al., 2011, 2009; Charron and Koechlin, 2010; Koechlin et al., 1999). The existence of such a correlate in the non-human primate is ground-breaking, as it challenges ideas of a unique anthropoid reasoning and planning ability supported by lateral FP (Koechlin, 2011). Although this idea of FP (in particular, the lateral FP part where our recordings were performed) being a specialised human brain structure also has some anatomical support with the mismatch in functional MRI connectivity patterns of lateral FP between humans and macaques (Neubert et al., 2014), it is important to note that such patterns of connectivity were obtained in different cognitive states (i.e., anaesthetised animals versus restive awake humans), making it hard to directly relate findings from both species.

One might object that the positive correlation with the expected first-stage chosen value and the slight negative correlation with the upcoming chosen second-stage value shown in the FP suggests that this is just a negatively-signed prediction error signal (consistent with

the negative outcome encoding we observed in this area). However, this ignores the dependence on the nature of the transition – it is this that implicates counterfactual reasoning.

Both lesion and functional neuroimaging studies in humans have suggested that FP monitors alternative courses of action. This role is better exemplified in sequential decision-making tasks, with their extra richness and complexity (Boorman et al., 2011, 2009; Burgess et al., 2007, 2000; Charron and Koechlin, 2010; Koechlin et al., 1999). Serial order behaviour and the ability to hold different levels of *schemata for action* in mind simultaneously, has long been of great relevance for psychology (Lashley, 1951). The "regret theory" of Loomes and Sugden (1982) challenged more standard theories of decision-making under uncertainty (Kahneman and Tversky, 1979), by proposing that subjects are influenced by regret, which arises when they discover that the option they took is worse than one they could have taken instead; and 'rejoice', when the option turns out better. The idea is that in making choices, subjects try to anticipate and take account of future regret and rejoice. This is rather in line with the FP encoding of foregone value that we observed, and the time point at which this is seen.

Two very recent studies reported subtle differences compared with controls in the behaviour of monkeys with lesions that included the area of FP from which we recorded. One of these studies highlighted the role of FP in rapid learning of the relative value of wide-ranging novel alternatives (Boschin et al., 2015). The other study found that FP lesioned animals were less prone to distraction than controls while maintaining task-relevant information (Mansouri et al., 2015). This has led the authors to propose that FP redistributes cognitive resources away from the task at hand, in line with proposals from human studies involving FP in coordinating attention between externally-presented and internally-represented information (Burgess et al., 2007). This latter possibility could fit with our findings in that a rare transition that occurs when the subject's first-stage choice should have given access to a valuable second-stage option, could lead to disengagement. The subject could just choose randomly and hope for a better first- to second-stage transition next time. Equally, it could be the case that the choice in this unchosen second-stage state could be highly rewarding, and in this case the FP signal might help the subject to focus on this exploratory benefit. Indeed, tracking alternative task strategies is of particular importance during exploration, and could explain this apparent functional correlate of the FP (Daw et al., 2006).

Second-stage reward prediction error in striatum

In the final section, we focused on the second-stage TD reward prediction error (i.e., the difference between received and expected reward), a key teaching signal for second-stage reward-based learning (Sutton and Barto, 1998). Midbrain dopamine neurons, which send many projections to the striatum and prefrontal cortex, are known to encode reward prediction errors (Schultz et al., 1997) and the release of dopamine promotes corticostriatal plasticity in target neurons (Barto et al., 1995; Calabresi et al., 2007; Suri and Schultz, 1999). Non-dopaminergic cells have also been found to report elements of temporal difference prediction errors at this time point. Similar (but not identical) feedback-related signals, revealing differences between received and expected outcome, have been reported in striatum (Apicella et al., 2009; Asaad and Eskandar, 2011; Kim et al., 2009; Oyama et al., 2010) as well as in prefrontal cortex, particularly in ACC (Kennerley et al., 2011; Matsumoto et al., 2007). However, in most of these studies the definition of reward prediction error was not as crisp as in RL theory, the analysis did not address the quantitative features of the signal, or, more critically the neurons did not exhibit the negative aspect of the prediction error, when outcomes were worse than expected.

We found that the firing rates of both caudate and putamen neurons at the time of feedback were very similar to what would be expected of dopaminergic neurons: a phasic response, with a short-onset latency and, most importantly, with parametric features of a reward prediction error (Bayer and Glimcher, 2005; Schultz et al., 1997). Hence, the firing rate in these striatal neurons correlated positively with the upcoming reward revealed by the secondary reinforcer, and negatively with what was expected given the previous outcome history of the second-stage choice. In fact, more recent outcomes exerted greater influence on firing than outcomes in more distant trials, resembling the exponential decay also observed on first-stage choice behaviour with the reward history effect (see chapter 3). Further evidence confirmed the quantitative properties of this second-stage reward prediction error signal, taking advantage of the trial-by-trial estimates derived from our best-fit computational model.

One interesting difference between the second-stage reward prediction error signals of caudate and putamen should be highlighted. The neural activity of these two regions showed distinct encodings of the subjective value of medium outcome level: in the caudate, it was treated either neutrally or slightly more like a low outcome; in the putamen, the response was closer to that expected for a high outcome. One possibility is a link of this difference with the similar divergence found in our behavioural analysis for choice and reaction times at the first stage. The medium outcome was treated more like the low outcome in terms

of choice 3.3, thereby being putatively more closely related to the caudate, but the high outcome in terms of reaction time 3.12, thereby being more closely associated with the putamen. The former is consistent with the slightly earlier and stronger population code observed in caudate when compared to putamen, the latter is consistent with the selectivity exhibited by the putamen with respect to the hand movement during the choice epoch.

The difference could also reflect distinct processing or sensitivity to delayed rewards, similar to the segregation found between signals in the dorsal and ventral primate striatum related to temporally discounted values (Cai et al., 2011). Dorsal striatum had a more important role in choosing a particular action based on temporally discounted values than the ventral striatum, in line with what we observed in our caudate data. Whereas ventral striatum was found to be more involved in the state-value, i.e., it encoded the sum of the temporally discounted values of the available options. This latter possibility hints at some hierarchical structure within striatum for reward processing (Bornstein and Daw, 2011; Kim and Hikosaka, 2013), where caudate (particularly more anterior parts; see Kim and Hikosaka 2013; Miyachi et al. 1997; Williams and Eskandar 2006) incorporates more choice-outcome information and it is closer to a MB strategy; putamen (particularly more posterior parts; see Kim and Hikosaka 2013; Miyachi et al. 1997) will be more focused on the links between the actual motor response and the outcome; and finally the ventral striatum associated with the role of the 'critic', which learns predictions of long term future reward (O'Doherty et al. 2004; see also Dayan and Berridge 2014 for MF and MB Pavlovian reward learning).

The ACC has been implicated in monitoring behavioural errors, with studies showing that ACC neurons can encode either a positive or a negative difference between actual and expected outcomes (Critchley et al., 2005; Holroyd et al., 2004; Kennerley et al., 2011; Matsumoto et al., 2007; Paulus et al., 2002; Silvetti et al., 2013b). An important view is that these ACC error- or feedback-related responses to both positive or negative outcomes do not encode prediction errors as such, but rather that they signal outcomes that are relevant for behavioural adaptation (Hayden et al., 2011b; Quilodran et al., 2008). Two good examples of this are the ACC activity that signals the end of an exploratory period and thus the shift towards exploitation (Karlsson et al., 2012; Quilodran et al., 2008), and the coding of the relative value of foraging compared to exploiting a resource (Hayden et al., 2011b).

This interpretation is consistent with our finding that the overall response in the ACC as a function of the expectation context (i.e., the value of second-stage choice) was particularly high for situations that were more instructive about what to do on the next trial. To illustrate this, if the second-stage choice had previously been rewarded, but was then associated with a low outcome, ACC neurons increased their firing rates, potentially signalling the need

to search for a different second-stage option. On the other hand, if the previous outcome had been medium or low, ACC neurons fire if the new outcome is the best one, which indicates the need to exploit that option. It was also notable that the spiking activity for low reward was stronger than for medium in this latter case. Low outcomes are relatively more instructive than medium ones, because the associated action should definitely not be repeated. Medium outcomes lead to an exploration-exploitation dilemma. It is likely that the specificity of the ACC relates in part to its position within the reward circuit and to the use of outcome information for action value adjustments and behavioural regulation or global changes in goal-directed policy.

The response pattern observed in the FP rather closely resembled that in the ACC, except that it was more phasic. This outcome-related activity is consistent with the only other primate neurophysiology study in the FP. There, neurons encoded the response that was correct (or behaviourally relevant) according to the cued strategy around feedback time (Tsujimoto et al., 2010, 2012). Human fMRI studies have also reported significant FP feedback activity consequent on the absence of an expected reward, as well as for the unexpected occurrence of a reward, particularly in an instrumental context (Ramnani et al., 2004). Furthermore, some authors have observed clear FP surprise responses that could not be accounted for as a classic reward prediction error or state-prediction error (Chumbley et al., 2014). This reinforces the account we gave above of the feedback-related signal in the ACC. The resemblance of the observed signals between ACC and FP also fits well the specific, dense, and direct neuroanatomical crosstalk between these two structures, which generally suggest strengthening of FP excitation by ACC projections (Medalla and Barbas, 2010).

Outcome-related activity in DLPFC has more consistently been found to be affected by previous actions (several trials into the past) than by prediction errors (Asaad and Eskandar, 2011; Barraclough et al., 2004; Tsujimoto et al., 2010). Indeed, we also failed to find evidence suggestive of an error signal in DLPFC. Others have also reported much less prediction error coding in DLPFC when compared to other prefrontal regions (Kennerley et al., 2011; Matsumoto et al., 2007). Note a particular feature of the design of our task, namely that the secondary reinforcer was presented in the center of the screen. This was to avoid a potential confound associated with the observation that DLPFC neurons are known to encode both value and spatial position (Kennerley et al., 2009; Kennerley and Wallis, 2009b; Rao et al., 1997). Consider a task in which visual feedback informing subjects as to whether or not they are being given reward is presented on the chosen side (as in Asaad and Eskandar, 2011). DLPFC neurons, by responding according to whether or not reward is presented on their preferred side of the screen, could erroneously be classified as reflecting

prediction errors.

Conclusion

In conclusion, using a decision task which admits formally distinguishable MF and MB values, we showed that key RL elements were encoded in different brain regions at different time points. The observed widespread and simultaneous representations of MF and MB value computations are consistent with the view that these controllers operate in parallel. However, we found that their associated signals were more richly intertwined than had originally been expected. We found that the ACC lay at the heart of the arbitration process between MF and MB control, being crucial both in valuation and in the promotion of optimal behaviour. Our data also confirmed the well known role of striatum in reinforcement and for guiding actions based on past rewards. Finally, we provided novel neurophysiological evidence in favour of the role of the FP in representing or processing counterfactuals. By extending such a sophisticated concept into identifiable activity in the FP of nonhuman primates, we offer further buttressing of the utility of this animal model for the understanding of some of the most complex cognitive behaviours exhibited by humans.

Chapter 6

Concluding remarks

Abstract

This thesis used behavioural measures, computational modelling and single-neuron physiology to investigate the role of the prefrontal cortex and the basal ganglia in model-free and model-based reinforcement learning approaches. Here, we bring together some of the major points made in the experimental chapters (chapter 3, chapter 4 and chapter 5), highlighting the reasons why they constitute advances for our understanding of MF and MB computations in the brain.

Introduction

The study set out to explore contemporary computational approaches that suggest two major competing and cooperating systems for reinforcement learning (RL) and behavioural control: model-free (MF) or habitual, and model-based (MB) or goal-directed. Although a wealth of studies in various species, has revealed regions of the brain that are particularly implicated in these forms of control – notably prefrontal cortex and the basal ganglia – much less is known about their neural realisations. Furthermore, although various studies have traditionally favoured their computational, behavioural and neural segregation, recent findings in humans (Daw et al., 2011) have challenged this idea. Instead, they have suggested a more promiscuous computational architecture. In this thesis, we aimed to contribute further data and understanding about the implementation and interaction of the two systems.

To achieve these goals, the three levels of computational modelling proposed by Marr (1982) have served as a constant inspiration:

- **Computational level**: our main goal focused on studying control policies of a two-stage, non-stationary, decision process such that the acquisition of reward is maximised. As ways of tackling the computational problem, we focused on RL methods. This was introduced in chapter 1, but also discussed in the remaining chapters.

- **Algorithmic level**: the algorithms used involved MF-RL methods, for which estimates are calculated from direct, retrospective, experience of rewards, using a temporally sophisticated form of prediction error; and MB-RL methods, for which estimates are prospective, based on a learned characterization of the environment and its affordances. The algorithms were formally detailed in chapter 1; however, further refinements proved necessary in the light of our behavioural data (chapter 3).

- **Implementational level**: the spiralling, richly connected, midbrain-striatal-prefrontal network known to be part of the neural architecture for affective decision-making was dissected in chapter 2. The signals derived from the algorithmic solutions to the computational problem were correlated with several behavioural measures, such as reaction time (in chapter 3) and pupil response (in chapter 4), as well as with single-neuron activity of three prefrontal cortex regions and two striatal areas (in chapter 5).

The experimental work presented in chapter 3, chapter 4 and chapter 5 was therefore strongly hypothesis-driven. In the next section, we evaluate the fate of these hypotheses in summarising and discussing the main findings from each experimental chapter.

Summary of contributions

Behavioural results

The behavioural analysis in chapter 3 suggested:

- Choice behaviour in non-human primates revealed combined MF and MB influences, with the influence of the latter approaching 90%

- The best-fit hybrid computational model used $SARSA$ as the MF algorithm, and a MB approach in which the state-transition probabilities were known from the start

- A credit assignment weighting procedure was required in the model so that the influence of the immediately previous trial information was strengthened

- The reaction time (RT) analysis found that decisions taking into account both outcome and transition structure took longer, whence the main effect of outcome was stronger overall

The fact that both subjects showed clear evidence of combined MF and MB-RL choice behaviour was essential for the main goal of the project. At the same time, this work was the first to replicate in animals other than humans concomitant use of both learning strategies. The way both descriptive and computational types of analysis were related ended up being very fruitful, as discrepancies prompted refinements to the initial, more conventional, hybrid model. The key discrepancy was the apparently excessive influence of information from the immediately previous trial. The required additional credit assignment was based on both MB and MF principles: MF, as the implementation was retrospective; and MB, as, algorithmically, it integrated both reward and transition information. Although this innovation had not previously been considered for variants of the two-step task, it does relate to other suggestions in the literature (Akam et al., 2015). One possibility is that it arises as a form of counterfactual thinking associated regret/rejoice theories. Support for this was also found in our pupil and neural data. Pupil dilation was influenced by disappointment; frontal pole (FP) single-neuron findings were also suggestive of value arising from the action that was not chosen. Unfortunately, the present work did not directly test the impact of this neural counterfactual evidence on behaviour and further analysis may be needed. Nevertheless, given that little is yet known about the computational and neural foundations of counterfactual reasoning these results and further analyses, our result could shed some light onto the field and foster further theoretical and experimental research.

The RT analysis corroborated the main behavioural findings, but it also provided interesting evidence regarding different influences. Theoretical accounts have suggested a speed accuracy trade-off between MF and MB computations, with the former being fast and the latter relatively slow. Indeed, the RT analysis confirmed that decisions taking into account both outcome and transition structure took longer. This RT effect followed a similar exponential decay with trials into the past as in the choice data. It would be harder to square with the suggestion that faster responses arise from the chunking of sequential actions, something that our design deters, with randomized positions for second-stage stimuli. It also militates against MB proposals emphasizing pre-computations at the time of outcome, where the re-evaluation of the utility of states given the received outcome helps future choice. Overall,

the RT evidence is supportive of a forward looking MB valuation process happening at the time of choice. It was notable that the RTs, particularly at the time of fixation, were more strongly influenced by the main effect of outcome, than any effect of transition or outcome $\times$ transition. This may be consistent with the observation that the average outcome rate, estimated in a MF way from recent past trials, and putatively reported through tonic activity of dopamine neurons, is a main mediator of the vigour of actions.

Finally, I would like to share two findings from the challenging and laborious training protocol. Unfortunately, the constant changes required to tackle the adversities encountered during this training process made any formal analysis of the task training approach unreliable or almost impossible. First, in initial phases of learning the task the outcome levels only differed in the magnitude of the rewards. The training proceeded well, with significant differences in first-stage choice behaviour between common and rare trials for the highest outcome level (repeat if common-high and switch if rare-high). This was a signature of the presence of MB behaviour. However, the expected difference between transition types for the lowest outcome level was hard to elicit (i.e., we did not see common-low being different than rare-low, although the overall probability of repeat was low). Two possible reasons were considered: the lowest outcome level was not as instructive as the highest outcome; and the overall subjective value of the low outcome level might not have been sufficiently low as to promote further attentional effort for further optimisation of the task. We therefore tried aversive outcomes (by delivering a diluted quinine solution) for the low outcome level. This rapidly had the expected consequence. Unfortunately, though, the change was probably too dramatic and, with time, led to some stereotypical behaviour on the next first-stage choice (the subject always chose a particular response side) if a low outcome level was received. As a consequence, not only did we introduce delays, but also we tried having three, rather than two, possible motor responses. This was very successful in achieving the results presented in this thesis.

Second, in another experimental phase of the training protocol, the stimuli used as well as the background colours for each state varied every day, but keeping the 70%/30% state-transition structure remained constant throughout the session. These changes elicited new transition matrix learning on every session. As a whole, the results were consistent with a MF response profile in these earlier phases of learning. This anecdotal evidence is not consistent with the idea that goal-directed or MB learning is prominent in initial phases of learning (Dickinson and Balleine, 2002), given the statistical inefficiency (and consequently higher uncertainty) of the MF-RL, as proposed by some theoretical views (Daw et al., 2005). This finding deserves further exploration in future experimental work.

Pupil results

The analysis of pupil dilation in chapter 4 suggested:

- Pupil size encoded positively the expected chosen value both pre- and post-choice
- The choice valuations reflected in the pupil were overwhelmingly MB
- The effect of what can be construed as a disappointment signal elicited pupil constriction
- Pupil dilation around the time of feedback reported the reward prediction error of the second-stage choice

A conventional suggestion is that the locus coeruleus-noradrenaline system mediates pupil dilation and cognitive states. Thus, a large proportion of pupil studies have focused, often with success, on testing theoretical proposals related to this neuromodulatory system. Hence, pupil dilation has been linked to high uncertainty or contexts where this happens, such as when a change point occurs or in exploratory behaviour. Although our experimental work did not aim to disprove these findings, it suggests that expectations of value might also be important. In any case, the value-based nature of the pupil modulation is not necessarily unpredicted or specific, as it an ubiquitous signal throughout the brain.

The observation that MF values made no contribution to the value expectation signals observed in pupil dilation, i.e., that they were purely MB, was more unexpected. In addition, the pupil diameter showed an independent effect of the state-transition information, another MB feature, and in a way inconsistent with just surprise, because its diameter decreased in rare trials (whereas others have reported pupil dilation for surprising events). Despite the reports of cognition-related pupil responses, dilation is more consistently regarded as a measure of emotional and autonomic activation, and furthermore to be controlled by evolutionarily primitive brainstem centres. With this in mind, when applying a task designed to detect simultaneous signals of both MF and MB-RL approaches, one might have thought that pupil would report the less sophisticated MF valuations. By revealing the contrary, the present work unmasks our ignorance regarding the neural substrates mediating the influence of cognitive state on pupil diameter. It also emphasises the role of top-down influences that could also arise, for instance, from prefrontal and basal ganglia structures.

Pupil diameter also reported the discrepancy between the expected value and either an external factor which bore on the final outcome (i.e., the type of transition) or the actual

received outcome. These prediction error signals are known to drive learning in both economics and psychology studies, and they have also been observed to control pupil diameter. To our knowledge, this finding is novel, and so needs further confirmation. However, disappointment is indeed just an evaluative process that takes other signals claimed to be present in the pupil into account: the value expectation and the uncertainty.

A similarly new interesting finding was the robust correlation between feedback-related pupil activity and a reward prediction error. The fact that this occurs at second-stage of the task, and both MF and MB valuations are the same prevents us from concluding its status as a signal of one or other controller.

A final speculation concerns the neural substrates of these pupil signals. If one had been asked to select the brain region that could best correlate with the above pupil findings, various lines of evidence point to the anterior cingulate cortex (ACC). Anatomically, the ACC has strong projections to the locus ceruleus, as mentioned in chapter 2. Functionally, the ACC has been shown to be involved in many relevant processes discussed in chapter 2 and chapter 5, which include value-based learning and foraging, error monitoring and conditions of high estimation uncertainty. Physiologically, our own results in chapter 5 speak directly to this link, although no actual mediation analysis was performed. Single-neuron activity in ACC was strongly correlated with expected chosen value, chosen MB estimates and transition type. Nevertheless, the disappointment effect is more closely related to the FP counterfactual evidence, and the second-stage reward prediction error to striatal activity. In conclusion, if the eyes are really *the window onto the soul,* it is crucial that future research dissect the neural mechanisms underlying all these influences in pupil in order to take full advantage of this simple and non-invasive measure.

Neuronal results

At a neural level, the present study found rich representations of key elements of RL theory. Highlights presented in chapter 5 include:

- The ACC was the region that most prominently encoded the value of the reward at the time of feedback. This information remained present until the subsequent first-stage choice

- The relationship between firing rate and the reward magnitude was predominantly positive for striatal regions and negative for both ACC and FP

- The ACC was also the region that most prominently encoded the transition type from when this was revealed until the feedback epoch
- At the level of the whole population, FP coding showed a particular bias towards positive encoding of rare transitions
- At the time of feedback, reward and transition information was simultaneously present in ACC
- There were specific relationships between the way FP neurons were selective for the transition, when this was revealed, and the way they coded for rewards, at the time of feedback
- For FP neurons, the actual transition experienced in a trial influenced the coding of reward at the time of feedback
- The identity of the first-stage stimulus that was picked weakly coded across all regions at the time of choice, but the side of the chosen first-stage response was strongly coded in the putamen
- Exclusively MF and exclusively MB action-value coding neurons were discovered in all brain areas recorded and in relatively similar proportions
- The same broad distribution was observed for neurons showing both MF and MB computations; the ACC showed significantly greater proportion of these hybrid RL cells than other regions
- Neural activity in the caudate and, to a lesser extent, the putamen, at the time of the first- to second-stage transition was consistent with a first-stage reward prediction error
- FP neurons increased their firing rate more when a rare transition was revealed for higher expected values of the outcome that was thereby out of reach
- Both caudate and putamen encoded the second-stage reward prediction error, albeit exhibiting slight qualitative differences

Just a note on probably one of the most prevalent questions among readers of this thesis – why neural activity from orbitofrontal cortex (or even ventromedial prefrontal cortex) was not performed? –, given the extensive literature suggesting links with MB behaviour.

Unfortunately, for both health and welfare reasons, there is a limit to the amount of neurophysiological data collection that can be obtained from a given subject, and a limit to how long they will remain motivated to work on the task. We had hoped to record from OFC and vmPFC, but because of the depth of these areas, we reserved these areas for last and unfortunately were not able to explore these areas in the end. Nonetheless, this should not diminish the results obtained from 5 different brain regions, all of which provided unique functional selectivity types other than perhaps DLPFC. We hope to examine OFC, vmPFC and dopamine responses in a MB task in the future.

Given the extensive collection of neuronal findings and the thorough discussion in chapter 5, here we provide a brief summary as to how each region's signals might contribute to MF and MB learning.

The ACC was the region whose activity was most closely related to all aspects of the task. In terms of its role: first, neurons there most consistently covaried with aspects of both MF-RL (reward coding) and MB-RL (reward and transition coding). This does not imply that the region is particularly focused on performing MF or MB updates, or that it is involved in transition learning. Instead, it means that the key ingredients for both value computations are present at the same time, across different time periods in ACC.

Note that although the neural implementation of MF predictions has been much investigated recently, much less is known about the mechanisms supporting MB control in the brain. Therefore, our single-neuron activity showing ACC representation of the state-transition information, its evolution across the trial and the concomitant coding with reward have important implications, and might perhaps prompt new theories about the algorithmic underpinnings of MB control.

Second, the above findings together with the disproportionately larger proportion of cells with access to both MF and MB values at the time of choice, implicate ACC in the arbitration process between the two learning strategies that has been considered crucial to guide optimal behaviour. This latter view could provide a new theoretical framework to unify the different roles attributed to ACC, in particular performance monitoring, error learning, conflict resolution and cognitive control.

The basal ganglia is believed to lie at the heart of the machinery for reinforcement learning in the brain, with the striatum being regarded as a particularly important area for action selection. The clearest and most singular characteristic we found for both the caudate and the putamen was their encoding of the second-stage reward prediction error. Such a teaching signal is well known to be encoded in midbrain dopamine neurons, which send strong projections to the striatum as well as prefrontal cortex, but has scarcely been reported in the past

in striatal cells themselves. Comparing this activity to prior reports of the phasic responses of dopaminergic neurons to reward-predicting stimuli, caudate and putamen modulation occurs at much longer latency (dopaminergic cells: latencies of <100 ms and durations <200 ms; here, caudate and putamen with latencies of ranging from 250-350 ms). It is therefore tempting to speculate that such striatal response is consequence of the dopamine release in the region. Having said that, this is quite a complicated task and some neuronal variance is likely associated with other functions. Most dopaminergic cells studies used relatively simple intrumental tasks or even just pavlovian ones. Although, striatal responses were not as fast as the ones reported to dopamine, one should bear in mind task complexity differences when evaluating latencies differences between studies.

The same caudate and putamen regions showed neurons involved in MF and MB action valuation. Thus, the ingredients for both action selection and action reinforcement are both present in the same regions. However, there were different response patterns in caudate and putamen suggesting that the two regions are not exact mirrors of each another: activity in the caudate matched better the subjects' choices; putamen responses was more in line with the profile observed in the reaction time analysis. This may suggest that activities in each region are updated in a segregated manner, supporting a more choice-based influence in caudate and a response vigour-based influence in putamen. The more prominent negative prediction error signal in caudate relative to putamen observed in rare transitions relative to first-stage chosen value expectation further supports this view.

In humans, the FP occupies the most anterior position at the proposed rostro-caudal axis processing gradient of the frontal cortex. It also has a much larger volume in the human brain than in any other animal. It is thus a target of immense debate in the human neuroscience literature, as it only seems necessary in complex multi-task scenarios, where engagement in branching processes (Koechlin et al., 1999), exploratory decisions (Daw et al., 2006) or monitoring of alternative choices values (Boorman et al., 2011, 2009) are required. One notion is that it supports some unique anthropoid reasoning and planning ability.

Our results suggest that this putatively unique function cannot be counterfactual reasoning. Indeed, it was the FP encoding of foregone value, at the particular time point at which this became apparent, that most clearly suggested a role of this region in counterfactual thinking. This result is neurophysiologically novel. It was specially compelling, given the overall weak selectivity in the FP.

Very little can be said about a specific DLPFC role in our task as the only most prominent observation was the relatively long and maintained coding of reward from feedback into the inter-trial period. This might be seen as a contrast to other studies suggesting its involvement

in value computations or choice. This might result from the facts that choice in the present task was indicated by a hand movement instead of a saccade, and that working memory related processes were not particularly addressed in this task.

6.1 Future directions

Finally, the work presented here paves the way for future theoretical and experimental investigations to address a number of new questions, some of which include:

- Following our modelling findings suggesting refinements to more conventional hybrid approaches with rich credit assignment features, the next question is: what are the neural substrates supporting credit assignment processes? Does this have a widespread distribution as well or is it governed by specific interactions between certain regions? Finer-grained analyses exploring correlations between stimulus specific and feedback activity may reveal insights into this important learning signature.

- In our task, the behavioural analysis suggested that the transition matrix was known from start. However, several questions remain unanswered: how is the state-transition structure learned? What structures are involved? How do neural signals allow the creation of the internal state-space that the subjects use?

- Throughout our experimental work, several independent processes identified through behavioural analysis, reaction time, pupil and single-neuron activity, could be indirectly linked. A good example of this is the speculation about an influence of putamen outcome-related activity on reaction time. However, our analysis lacked depth into more direct assessments of such relationships (e.g., mediation analysis). A more detailed description of how the different elements are linked should provide further insights into the actual mechanism as well as its influences.

- In our analysis, we only considered the reward magnitude given that the delays were fixed. However, the contribution of delay to choice could have important implications and deserve further analysis in the future.

- To extend our understanding of how this task is implemented and the signals that other structures could provide, it would be very interesting if other areas were also recorded. As main targets, it would be interesting to see recordings in dopamine midbrain neurons, orbitofrontal or ventromedial prefrontal cortex, hippocampus, entorhinal cortex

and amygdala. Previous work suggests involvement of these structures in MF and MB, but most studies did not offer the advantage of simultaneous single-neuron activity.

- Given the preponderance of ACC in encoding most information relevant for the task, interference manipulations (lesion, inactivation or stimulation) during performance of the two-stage decision task implemented here could further confirm our hypothesis or generate more questions.

- To further address the possibility of an ACC involvement in the arbitration process of MF and MB RL, one could try to model uncertainties in the estimates of each learning strategy and correlate on a trial-by-trial basis with the single-neuron activity. This way, not only proposals that uncertainty is crucial in such arbitration process are tested, but it would also provide more direct evidence.

6.2 Conclusion

This project started with a relatively straight forward aim to test theoretical proposals of RL principles that could govern animal learning and decision-making. More importantly, it tested whether some of the postulated computational signals are used by the brain to solve the problems it faces. Such reverse engineering process has the advantage of providing more quantitative information regarding the neural mechanisms and the underlying model, which could have critical contributions in either detecting or fixing encountered anomalies in the system. It turned out to be an initially ambitious approach but with successful outcomes.

While training non-human primates on a complex sequential decision task, we were able to unravel by their choices and their response vigour, features of a combined MF and MB RL control. More importantly, the thorough analysis employed discovered incongruence with more conventional models of such hybrid behaviour. This motivated new ideas of credit assignment where reward, contextual and temporal information could be integrated to boost the choice either performed or not. Interestingly, we found that other behavioural measures such as pupil diameter, reflected prospective MB valuation and the consequent disappointment if the expected reward can no longer be achieved. All these correlates were relatively novel and prompt further confirmatory evidence and dissection of the underlying mechanisms.

Regarding the actual implementation of the MF and MB computations in the brain, the simultaneous recordings in several different prefrontal and basal ganglia regions was for sure appropriate. Not only it was found that the signals involved in both learning methods are more complex and richly intertwined across the prefrontal-striatal circuitry, but a variety of specific signals were also found in support of functional segregation. There was no surprise to see the ACC very participative in a demanding learning task, but it was unexpected to see evidence that offers hope in reuniting dispersed ACC theories in a well established computational framework. Although the evidence is not causal enough, the reported findings push forward the idea that the role of ACC is the resolution of the competition or the need for cooperation between MF and MB control, with the final intent to promote adaptive behaviour. The striatal feedback-activity elegantly reflected a clear involvement in learning by reporting a reward prediction errors. Finally, one is only reassured that it is doing science if it experiences that feeling of the unexpected discovery. The FP results presented in this thesis, are novel in the sense that it implicates neurophysiologically this region in counterfactual reasoning and relates nicely with the modelling refinement proposed here as well as the pupil signal. By challenging some views of unique human cognitive capacities, the data presented here reassures future research that the doors are still open to dive into the complexities of animal reasoning.

References

Adams, C. D. (1982). Variations in the sensitivity of instrumental responding to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B*, 34(2):77–98.

Adams, C. D. and Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B*, 33(2):109–121.

Akam, T., Costa, R., and Dayan, P. (2015). Simple plans or sophisticated habits? state, transition and learning interactions in the two-step task. *bioRxiv*.

Alexander, G. E., DeLong, M. R., and Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, 9:357–381.

Alexander, W. H. and Brown, J. W. (2011). Medial prefrontal cortex as an action-outcome predictor. *Nat Neurosci*, 14(10):1338–1344.

Amemori, K.-i. and Sawaguchi, T. (2006). Contrasting effects of reward expectation on sensory and motor memories in primate prefrontal neurons. *Cerebral Cortex*, 16(7):1002–1015.

Amiez, C., Joseph, J., and Procyk, E. (2006). Reward encoding in the monkey anterior cingulate cortex. *Cerebral Cortex*, 16(7):1040–1055.

Amiez, C., Joseph, J.-P., and Procyk, E. (2005). Anterior cingulate error-related activity is modulated by predicted reward. *European Journal of Neuroscience*, 21(12):3447–3452.

An, X., Bandler, R., Öngür, D., and Price, J. (1998). Prefrontal cortical projections to longitudinal columns in the midbrain periaqueductal gray in macaque monkeys. *The Journal of Comparative Neurology*, 401(4):455–479.

Aosaki, T., Graybiel, A., and Kimura, M. (1994a). Effect of the nigrostriatal dopamine system on acquired neural responses in the striatum of behaving monkeys. *Science*, 265(5170):412–415.

Aosaki, T., Tsubokawa, H., Ishida, A., Watanabe, K., Graybiel, A., and Kimura, M. (1994b). Responses of tonically active neurons in the primate's striatum undergo systematic changes during behavioral sensorimotor conditioning. *The Journal of Neuroscience*, 14(6):3969–3984.

Apicella, P., Deffains, M., Ravel, S., and Legallet, E. (2009). Tonically active neurons in the striatum differentiate between delivery and omission of expected reward in a probabilistic task context. *European Journal of Neuroscience*, 30(3):515–526.

Apicella, P., Ljungberg, T., Scarnati, E., and Schultz, W. (1991). Responses to reward in monkey dorsal and ventral striatum. *Experimental Brain Research*, 85(3):491–500.

Arikuni, T. and Kubota, K. (1986). The organization of prefrontocaudate projections and their laminar origin in the macaque monkey: A retrograde study using hrp-gel. *The Journal of Comparative Neurology*, 244(4):492–510.

Arikuni, T., Sako, H., and Murata, A. (1994). Ipsilateral connections of the anterior cingulate cortex with the frontal and medial temporal cortices in the macaque monkey. *Neuroscience Research*, 21(1):19 – 39.

Asaad, W. F. and Eskandar, E. N. (2011). Encoding of both positive and negative reward prediction errors by neurons of the primate lateral prefrontal cortex and caudate nucleus. *The Journal of Neuroscience*, 31(49):17772–17787.

Asaad, W. F., Rainer, G., and Miller, E. K. (2000). Task-specific neural activity in the primate prefrontal cortex. *Journal of Neurophysiology*, 84(1):451–459.

Aston-Jones, G. and Cohen, J. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*, 28:403–450.

Averbeck, B. B. and Seo, M. (2008). The statistical neuroanatomy of frontal networks in the macaque. *PLoS Comput Biol*, 4(4):e1000050.

Averbeck, B. B., Sohn, J.-W. W., and Lee, D. (2006). Activity in prefrontal cortex during dynamic selection of action sequences. *Nature neuroscience*, 9(2):276–282.

Baddeley, A. (1992). Working memory. *Science*, 255(5044):556–559.

Badre, D. and D'Esposito, M. (2009). Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Reviews Neuroscience*, 10(9):659–669.

Badre, D. and Wagner, A. D. (2004). Selection, integration, and conflict monitoring: Assessing the nature and generality of prefrontal cognitive control mechanisms. *Neuron*, 41(3):473 – 487.

Balleine, B. and Dickinson, A. (1991). Instrumental performance following reinforcer devaluation depends upon incentive learning. *The Quarterly Journal of Experimental Psychology Section B*, 43(3):279–296.

Balleine, B. and Dickinson, A. (1998a). The role of incentive learning in instrumental outcome revaluation by sensory-specific satiety. *Animal Learning & Behavior*, 26(1):46–59.

Balleine, B. W. (2005). Neural bases of food-seeking: Affect, arousal and reward in corticostriatolimbic circuits. *Physiology & Behavior*, 86(5):717 – 730. Purdue University Ingestive Behavior Research Center Symposium. Dietary Influences on Obesity: Environment, Behavior and Biology.

Balleine, B. W., Delgado, M. R., and Hikosaka, O. (2007). The role of the dorsal striatum in reward and decision-making. *The Journal of Neuroscience*, 27(31):8161–8165.

Balleine, B. W. and Dickinson, A. (1998b). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37(4–5):407 – 419.

Balleine, B. W., Killcross, A. S., and Dickinson, A. (2003). The effect of lesions of the basolateral amygdala on instrumental conditioning. *The Journal of Neuroscience*, 23(2):666–675.

Balleine, B. W. and O'Doherty, J. P. (2009). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, 35(1):48–69.

Barbas, H., Ghashghaei, H., Dombrowski, S., and Rempel-Clower, N. (1999). Medial prefrontal cortices are unified by common connections with superior temporal cortices and distinguished by input from memory-related areas in the rhesus monkey. *The Journal of Comparative Neurology*, 410(3):343–367.

Barbas, H. and Pandya, D. N. (1989). Architecture and intrinsic connections of the prefrontal cortex in the rhesus monkey. *The Journal of Comparative Neurology*, 286(3):353–375.

Barker, J. M., Torregrossa, M. M., and Taylor, J. R. (2013). Bidirectional modulation of infralimbic dopamine d1 and d2 receptor activity regulates flexible reward seeking. *Frontiers in Neuroscience*, 7(126).

Barraclough, D. J., Conroy, M. L., and Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, 7(4):404–410.

Barto, A., Sutton, R., and Anderson, C. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *Systems, Man and Cybernetics, IEEE Transactions on*, SMC-13(5):834–846.

Barto, A. G., Houk, J. C., and Adams, J. L. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In Houk, J. C. and Davis, J. L., editors, *Models of information processing in the basal ganglia*, pages 249–270. Mit Press, Cambridge, MA, USA.

Barto, A. G. and Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, 13(4):341–379.

Bates, J. F. and Goldman-Rakic, P. S. (1993). Prefrontal connections of medial motor areas in the rhesus monkey. *The Journal of Comparative Neurology*, 336(2):211–228.

Baxter, M. G., Gaffan, D., Kyriazis, D. A., and Mitchell, A. S. (2008). Dorsolateral prefrontal lesions do not impair tests of scene learning and decision-making that require frontal–temporal interaction. *European Journal of Neuroscience*, 28(3):491–499.

Baxter, M. G. and Murray, E. A. (2002). The amygdala and reward. *Nat Rev Neurosci*, 3(7):563–573.

Bayer, H. M. and Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1):129 – 141.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91(2):276–292.

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., and Rushworth, M. F. S. (2008). Associative learning of social value. *Nature*, 456(7219):245–249.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., and Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9):1214–1221.

Belin, D., Jonkman, S., Dickinson, A., Robbins, T. W., and Everitt, B. J. (2009). Parallel and interactive learning processes within the basal ganglia: Relevance for the understanding of addiction. *Behavioural Brain Research*, 199(1):89 – 102. Special issue on the role of the basal ganglia in learning and memory.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition.

Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*. Athena Scientific, 1st edition.

Bolles, R. C., Holtz, R., Dunn, T., and Hill, W. (1980). Comparisons of stimulus learning and response learning in a punishment situation. *Learning and Motivation*, 11(1):78 – 96.

Boorman, E. D., Behrens, T. E., and Rushworth, M. F. (2011). Counterfactual choice and learning in a neural network centered on human lateral frontopolar cortex. *PLoS Biol*, 9:e1001093.

Boorman, E. D., Behrens, T. E., Woolrich, M. W., and Rushworth, M. F. (2009). How green is the grass on the other side? frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron*, 62(5):733 – 743.

Bornstein, A. M. and Daw, N. D. (2011). Multiplicity of control in the basal ganglia: computational roles of striatal subregions. *Current Opinion in Neurobiology*, 21(3):374 – 380.

Bornstein, A. M. and Daw, N. D. (2012). Dissociating hippocampal and striatal contributions to sequential prediction learning. *European Journal of Neuroscience*, 35(7):1011–1023.

Boschin, E. A., Piekema, C., and Buckley, M. J. (2015). Essential functions of primate frontopolar cortex in cognition. *Proceedings of the National Academy of Sciences*, 112(9):E1020–E1027.

Botvinick, M. and Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369(1655).

Botvinick, M. M., Niv, Y., and Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, 113(3):262 – 280. Reinforcement learning and higher cognition.

Bouret, S. and Richmond, B. J. (2010). Ventromedial and orbital prefrontal neurons differentially encode internally and externally driven motivational values in monkeys. *The Journal of Neuroscience*, 30(25):8591–8601.

Bouret, S. and Sara, S. J. (2005). Network reset: a simplified overarching theory of locus coeruleus noradrenaline function. *Trends in Neurosciences*, 28(11):574 – 582.

Bradshaw, J. (1967). Pupil size as a measure of arousal during information processing. *Nature*, 216(5114):515–516.

Brafman, R. I. and Tennenholtz, M. (2003). R-max - a general polynomial time algorithm for near-optimal reinforcement learning. *J. Mach. Learn. Res.*, 3:213–231.

Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in Cognitive Sciences*, 16(2):106 – 113.

Bromberg-Martin, E. S., Matsumoto, M., and Hikosaka, O. (2010). Dopamine in motivational control: Rewarding, aversive, and alerting. *Neuron*, 68(5):815 – 834.

Browne, C., Powley, E., Whitehouse, D., Lucas, S., Cowling, P., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S. (2012). A survey of monte carlo tree search methods. *Computational Intelligence and AI in Games, IEEE Transactions on*, 4(1):1–43.

Brozoski, T., Brown, R., Rosvold, H., and Goldman, P. (1979). Cognitive deficit caused by regional depletion of dopamine in prefrontal cortex of rhesus monkey. *Science*, 205(4409):929–932.

Buckley, M. J., Mansouri, F. A., Hoda, H., Mahboubi, M., Browning, P. G. F., Kwok, S. C., Phillips, A., and Tanaka, K. (2009). Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. *Science*, 325(5936):52–58.

Burgess, P. W., Dumontheil, I., and Gilbert, S. J. (2007). The gateway hypothesis of rostral prefrontal cortex (area 10) function. *Trends in Cognitive Sciences*, 11(7):290 – 298.

Burgess, P. W., Veitch, E., de Lacy Costello, A., and Shallice, T. (2000). The cognitive and neuroanatomical correlates of multitasking. *Neuropsychologia*, 38(6):848 – 863.

Bush, R. and Mosteller, F. (2006). A mathematical model for simple learning. In Fienberg, S. and Hoaglin, D., editors, *Selected Papers of Frederick Mosteller*, Springer Series in Statistics, pages 221–234. Springer New York.

Butler, A. B. and Hodos, W. (2005). *Overview of the Forebrain*, pages 341–372. John Wiley & Sons, Inc.

Cai, X., Kim, S., and Lee, D. (2011). Heterogeneous coding of temporally discounted values in the dorsal and ventral striatum during intertemporal choice. *Neuron*, 69(1):170 – 182.

Cai, X. and Padoa-Schioppa, C. (2014). Contributions of orbitofrontal and lateral prefrontal cortices to economic choice and the good-to-action transformation. *Neuron*, 81(5):1140 – 1151.

Calabresi, P., Picconi, B., Tozzi, A., and Filippo, M. D. (2007). Dopamine-mediated regulation of corticostriatal synaptic plasticity. *Trends in Neurosciences*, 30(5):211 – 219. Fifty years of dopamine research.

Calzavara, R., Mailly, P., and Haber, S. N. (2007). Relationship between the corticostriatal terminals from areas 9 and 46, and those from area 8a, dorsal and rostral premotor cortex and area 24c: an anatomical substrate for cognition to action. *European Journal of Neuroscience*, 26(7):2005–2024.

Campbell, M., Jr., A. H., and hsiung Hsu, F. (2002). Deep blue. *Artificial Intelligence*, 134(1–2):57 – 83.

Carmichael, S. and Price, J. (1996). Connectional networks within the orbital and medial prefrontal cortex of macaque monkeys. *The Journal of Comparative Neurology*, 371(2):179–207.

Carmichael, S. T. and Price, J. L. (1995a). Limbic connections of the orbital and medial prefrontal cortex in macaque monkeys. *The Journal of Comparative Neurology*, 363(4):615–641.

Carmichael, S. T. and Price, J. L. (1995b). Limbic connections of the orbital and medial prefrontal cortex in macaque monkeys. *The Journal of Comparative Neurology*, 363(4):615–641.

Carmichael, S. T. and Price, J. L. (1995c). Limbic connections of the orbital and medial prefrontal cortex in macaque monkeys. *The Journal of Comparative Neurology*, 363(4):615–641.

Carpenter, M. B. and Peter, P. (1972). Nigrostriatal and nigrothalamic fibers in the rhesus monkey. *The Journal of Comparative Neurology*, 144(1):93–115.

Charron, S. and Koechlin, E. (2010). Divided representation of concurrent goals in the human frontal lobes. *Science*, 328(5976):360–363.

Chiew, K. S. and Braver, T. S. (2013). Temporal dynamics of motivation-cognitive control interactions revealed by high-resolution pupillometry. *Frontiers in Psychology*, 4(15).

Chumbley, J. R., Burke, C. J., Stephan, K. E., Friston, K. J., Tobler, P. N., and Fehr, E. (2014). Surprise beyond prediction error. *Human Brain Mapping*, 35(9):4805–4814.

Cisek, P. and Kalaska, J. F. (2010). Neural Mechanisms for Interacting with a World Full of Action Choices. In Hyman, SE, editor, *ANNUAL REVIEW OF NEUROSCIENCE, VOL 33*, volume 33 of *Annual Review of Neuroscience*, pages 269–298.

Colwill, R. M. and Rescorla, R. A. (1985). Postconditioning devaluation of a reinforcer affects instrumental responding. *Journal of Experimental Psychology: Animal Behavior Processes*, 11(1):120–132.

Corbit, L. H. and Balleine, B. W. (2003). Instrumental and pavlovian incentive processes have dissociable effects on components of a heterogeneous instrumental chain. *J Exp Psychol Anim Behav Process*, 29(2):99–106.

Corrado, G. S., Sugrue, L. P., Seung, H. S., and Newsome, W. T. (2005). Linear-nonlinear-poisson models of primate choice dynamics. *Journal of the Experimental Analysis of Behavior*, 84(3):581–617.

Coutureau, E. and Killcross, S. (2003). Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. *Behavioural Brain Research*, 146(1–2):167 – 174. The Rodent Prefrontal Cortex.

Critchley, H. D., Tang, J., Glaser, D., Butterworth, B., and Dolan, R. J. (2005). Anterior cingulate activity during error and autonomic response. *NeuroImage*, 27(4):885 – 895.

Cromwell, H. C. and Schultz, W. (2003). Effects of expectations for different reward magnitudes on neuronal activity in primate striatum. *Journal of Neurophysiology*, 89(5):2823–2838.

D'Ardenne, K., McClure, S. M., Nystrom, L. E., and Cohen, J. D. (2008). Bold responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, 319(5867):1264–1267.

Daw, N. D. and Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 369(1655).

Daw, N. D. and Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, 16(2):199 – 204.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204 – 1215.

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., and Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879.

Daw, N. D. and O'Doherty, J. P. (2014). Chapter 21 - multiple systems for value learning. In Fehr, P. W. G., editor, *Neuroeconomics (Second Edition)*, pages 393 – 410. Academic Press, San Diego, second edition edition.

Day, J. J., Wheeler, R. A., Roitman, M. F., and Carelli, R. M. (2006). Nucleus accumbens neurons encode pavlovian approach behaviors: evidence from an autoshaping paradigm. *European Journal of Neuroscience*, 23(5):1341–1351.

Dayan, P. (2009). Dopamine, reinforcement learning, and addiction. *Pharmacopsychiatry*, 42:S56–65.

Dayan, P. (2012a). How to set the switches on this thing. *Current Opinion in Neurobiology*, 22(6):1068 – 1074. Decision making.

Dayan, P. (2012b). Twenty-five lessons from computational neuromodulation. *Neuron*, 76(1):240 – 256.

Dayan, P. and Berridge, K. (2014). Model-based and model-free pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience*, 14(2):473–492.

Dayan, P. and Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, 18(2):185 – 196.

Dayan, P., Niv, Y., Seymour, B., and Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural Networks*, 19(8):1153 – 1160. Neurobiology of Decision Making Neurobiology of Decision Making.

de Gee, J. W., Knapen, T., and Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences*, 111(5):E618–E625.

de Wit, S. and Dickinson, A. (2009). Associative theories of goal-directed behaviour: a case for animal–human translational models. *Psychological Research PRPF*, 73(4):463–476.

de Wit, S., Standing, H., DeVito, E., Robinson, O., Ridderinkhof, K., Robbins, T., and Sahakian, B. (2012a). Reliance on habits at the expense of goal-directed control following dopamine precursor depletion. *Psychopharmacology*, 219(2):621–631.

de Wit, S., Watson, P., Harsay, H. A., Cohen, M. X., van de Vijver, I., and Ridderinkhof, K. R. (2012b). Corticostriatal connectivity underlies individual differences in the balance between habitual and goal-directed action control. *The Journal of Neuroscience*, 32(35):12066–12075.

Debener, S., Ullsperger, M., Siegel, M., Fiehler, K., von Cramon, D. Y., and Engel, A. K. (2005). Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. *The Journal of Neuroscience*, 25(50):11730–11737.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38.

Deserno, L., Huys, Q. J. M., Boehme, R., Buchert, R., Heinze, H.-J., Grace, A. A., Dolan, R. J., Heinz, A., and Schlagenhauf, F. (2015). Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences*, 112(5):1595–1600.

Dezfouli, A. and Balleine, B. W. (2013). Actions, action sequences and habits: Evidence that goal-directed and habitual action control are hierarchically organized. *PLoS Comput Biol*, 9(12):e1003364.

Dias-Ferreira, E., Sousa, J. C., Melo, I., Morgado, P., Mesquita, A. R., Cerqueira, J. J., Costa, R. M., and Sousa, N. (2009). Chronic stress causes frontostriatal reorganization and affects decision-making. *Science*, 325(5940):621–625.

Dickinson, A. (1985). Actions and Habits: The Development of Behavioural Autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 308(1135):67–78.

Dickinson, A. (1996). Bidirectional instrumental conditioning. *The Quarterly Journal of Experimental Psychology Section B*, 49(4):289–306. PMID: 8962537.

Dickinson, A. (1998). Omission learning after instrumental pretraining. *The Quarterly Journal of Experimental Psychology Section B*, 51(3):271–286.

Dickinson, A. and Balleine, B. (1994). Motivational control of goal-directed action. *Animal Learning & Behavior*, 22(1):1–18.

Dickinson, A. and Balleine, B. W. (2002). The Role of Learning in the Operation of Motivational Systems. In Pashler, H. and Gallistel, R., editors, *Stevens' Handbook of Experimental Psychology*, volume 3: Learning, Motivation and Emotion, pages 497–533. John Wiley & Sons, New York, 3rd edition.

Ding, L. and Gold, J. I. (2013). The basal ganglia's contributions to perceptual decision making. *Neuron*, 79(4):640 – 649.

Dolan, R. J. and Dayan, P. (2013). Goals and habits in the brain. *Neuron*, 80(2):312 – 325.

Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., and Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nat Neurosci*, 18(5):767–772.

Doll, B. B., Simon, D. A., and Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, 22(6):1075 – 1081. Decision making.

Doya, K., Samejima, K., Katagiri, K.-i., and Kawato, M. (2002). Multiple Model-Based Reinforcement Learning. *Neural Computation*, 14(6):1347–1369.

Durstewitz, D., Seamans, J. K., and Sejnowski, T. J. (2000). Neurocomputational models of working memory. *Nature Neuroscience*, pages 1184–1191.

Ebitz, R. B. and Platt, M. L. (2015). Neuronal activity in primate dorsal anterior cingulate cortex signals task conflict and predicts adjustments in pupil-linked arousal. *Neuron*, 85(3):628 – 640.

Economides, M., Kurth-Nelson, Z., Lübbert, A., Guitart-Masip, M., and Dolan, R. J. (2015). Model-based reasoning in humans becomes automatic with training. *PLoS Comput Biol*, 11(9):e1004463.

Einhauser, W., Koch, C., and Carter, O. (2010). Pupil dilation betrays the timing of decisions. *Frontiers in Human Neuroscience*, 4(18).

Einhäuser, W., Stout, J., Koch, C., and Carter, O. (2008). Pupil dilation reflects perceptual selection and predicts subsequent stability in perceptual rivalry. *Proceedings of the National Academy of Sciences*, 105(5):1704–1709.

Faure, A., Haberland, U., Condé, F., and Massioui, N. E. (2005). Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation. *The Journal of Neuroscience*, 25(11):2771–2780.

Ferry, A. T., Öngür, D., An, X., and Price, J. L. (2000). Prefrontal cortical projections to the striatum in macaque monkeys: Evidence for an organization related to prefrontal networks. *The Journal of Comparative Neurology*, 425(3):447–470.

Fiedler, S. and Glöckner, A. (2012). The dynamics of decision making in risky choice: An eye-tracking analysis. *Frontiers in Psychology*, 3(335).

Fiorillo, C. D., Tobler, P. N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science (New York, N.Y.)*, 299(5614):1898–1902.

Frankle, W. G., Laruelle, M., , and Haber, S. N. (2006). Prefrontal cortical projections to the midbrain in primates: evidence for a sparse connection. *Neuropsychopharmacology*, 31(8):1627 – 1636.

Funahashi, S., Bruce, C., and Goldman-Rakic, P. (1993). Dorsolateral prefrontal lesions and oculomotor delayed-response performance: evidence for mnemonic "scotomas". *The Journal of Neuroscience*, 13(4):1479–1497.

Funahashi, S., Bruce, C. J., and Goldman-Rakic, P. S. (1989). Mnemonic coding of visual space in the monkeys dorsolateral prefrontal cortex. *Journal of Neurophysiology*, 61(2):331–349.

Fuster, J. M. (2001). The prefrontal cortex—an update: Time is of the essence. *Neuron*, 30(2):319 – 333.

Gaspar, P., Berger, B., Febvret, A., Vigny, A., and Henry, J. P. (1989). Catecholamine innervation of the human cerebral cortex as revealed by comparative immunohistochemistry of tyrosine hydroxylase and dopamine-beta-hydroxylase. *The Journal of Comparative Neurology*, 279(2):249–271.

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., and Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, 4(6):385–390.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15):2865–2873.

Genovesio, A., Brasted, P. J., Mitz, A. R., and Wise, S. P. (2005). Prefrontal cortex activity related to abstract response strategies. *Neuron*, 47(2):307 – 320.

Gershman, S. J., Markman, A. B., and Otto, A. R. (2012). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, Publish Ahead of Print.

Gershman, S. J., Markman, A. B., and Otto, A. R. (2014). Retrospective revaluation in sequential decision making: a tale of two systems. *Journal of Experimental Psychology: General*, 143(1):182–194.

Ghering, W. J. and Knight, R. T. (2000). Prefrontal-cingulate interactions in action monitoring. *Nature Neuroscience*, 3:516.

Gilzenrat, M., Nieuwenhuis, S., Jepma, M., and Cohen, J. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience*, 10(2):252–269.

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585 – 595.

Gläscher, J., Hampton, A. N., and O'Doherty, J. P. (2009). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral Cortex*, 19(2):483–495.

Gottlieb, J. and Balan, P. (2010). Attention as a decision in information space. *Trends in Cognitive Sciences*, 14(6):240 – 248.

Grindley, G. C. (1932). The formation of a simple habit in guinea-pigs. *British Journal of Psychology. General Section*, 23(2):127–147.

Haber, S., Kunishio, K., Mizobuchi, M., and Lynd-Balta, E. (1995). The orbital and medial prefrontal circuit through the primate basal ganglia. *The Journal of Neuroscience*, 15(7):4851–4867.

Haber, S. N., Fudge, J. L., and McFarland, N. R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *The Journal of Neuroscience*, 20(6):2369–2382.

Haber, S. N. and Gdowski, M. J. (2004). Chapter 21 - the basal ganglia. In Mai, G. P. K., editor, *The Human Nervous System (Second Edition)*, pages 676 – 738. Academic Press, San Diego, second edition edition.

Haber, S. N. and Knutson, B. (2009). The reward circuit: Linking primate anatomy and human imaging. *Neuropsychopharmacology*, 35(1):4–26.

Haber, S. N. and McFarland, N. R. (1999). The concept of the ventral striatum in nonhuman primates. *Annals of the New York Academy of Sciences*, 877(1):33–48.

Hadland, K. A., Rushworth, M., Gaffan, D., and Passingham, R. E. (2003). The anterior cingulate and reward-guided selection of actions. *Journal of Neurophysiology*, 89(2):1161–1164.

Hammond, L. J. (1980). The effect of contingency upon the appetitive conditioning of free-operant behavior. *Journal of the Experimental Analysis of Behavior*, 34(3):297–304.

Hampton, A. N., Bossaerts, P., and O'Doherty, J. P. (2008). Neural correlates of mentalizing-related computations during strategic interactions in humans. *Proceedings of the National Academy of Sciences*, 105(18):6741–6746.

Hampton, A. N., Bossaerts, P., and O'Doherty, J. P. (2006). The role of the ventromedial prefrontal cortex in abstract state-based inference during decision making in humans. *The Journal of Neuroscience*, 26(32):8360–8367.

Hayden, B. Y., Heilbronner, S. R., Pearson, J. M., and Platt, M. L. (2011a). Surprise signals in anterior cingulate cortex: Neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. *The Journal of Neuroscience*, 31(11):4178–4187.

Hayden, B. Y., Pearson, J. M., and Platt, M. L. (2009). Fictive reward signals in the anterior cingulate cortex. *Science*, 324(5929):948–950.

Hayden, B. Y., Pearson, J. M., and Platt, M. L. (2011b). Neuronal basis of sequential foraging decisions in a patchy environment. *Nat Neurosci*, 14(7):933–939.

Hayden, B. Y. and Platt, M. L. (2010). Neurons in anterior cingulate cortex multiplex information about reward and action. *The Journal of Neuroscience*, 30(9):3339–3346.

Hengst, B. (2012). Hierarchical approaches. In Wiering, M. and van Otterlo, M., editors, *Reinforcement Learning*, volume 12 of *Adaptation, Learning, and Optimization*, pages 293–323. Springer Berlin Heidelberg.

Hess, E. H. (1972). Pupillometrics: A method of studying mental, emotional and sensory processes. In Greenfield, N. S. and Sternbach, R. A., editors, *Handbook of Psychophysiology*, pages 491 – 531. Holt, Rinehart & Winston, New York.

Hess, E. H. and Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, 143(3611):1190–1192.

Hikosaka, K. and Watanabe, M. (2000). Delay activity of orbital and lateral prefrontal neurons of the monkey varying with different rewards. *Cerebral Cortex*, 10(3):263–271.

Hikosaka, O., Bromberg-Martin, E., Hong, S., and Matsumoto, M. (2008). New insights on the subcortical representation of reward. *Current Opinion in Neurobiology*, 18(2):203 – 208.

Histed, M. H., Pasupathy, A., and Miller, E. K. (2009). Learning substrates in the primate prefrontal cortex and striatum: Sustained activity related to successful actions. *Neuron*, 63(2):244 – 253.

Hitchcott, P. K., Quinn, J. J., and Taylor, J. R. (2007). Bidirectional modulation of goal-directed actions by prefrontal cortical dopamine. *Cerebral Cortex*, 17(12):2820–2827.

Holroyd, C. and Yeung, N. (2011). An integrative theory of anterior cingulate cortex function: Option selection in hierarchical reinforcement learning. In R.B. Mars, J. Sallet, M. R. and Yeung, N., editors, *Neural Basis of Motivational and Cognitive Control*. The MIT Press, Cambridge, MA.

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R. B., Coles, M. G. H., and Cohen, J. D. (2004). Dorsal anterior cingulate cortex shows fmri response to internal and external error signals. *Nat Neurosci*, 7(5):497–498.

Huerta, M. F. and Kaas, J. H. (1990). Supplementary eye field as defined by intracortical microstimulation: Connections in macaques. *The Journal of Comparative Neurology*, 293(2):299–330.

Hull, C. (1943). *Principles of Behavior: An Introduction to Behavior Theory*. The Century psychology series. D. Appleton-Century Company, Incorporated.

Huys, Q. J. M., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., and Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS Comput Biol*, 7(4):e1002028.

Huys, Q. J. M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., and Roiser, J. P. (2012). Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Comput Biol*, 8(3):e1002410.

Ito, M. and Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *The Journal of Neuroscience*, 29(31):9861–9874.

Ito, M. and Doya, K. (2011). Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current Opinion in Neurobiology*, 21(3):368 – 373.

Jacobsen, C. (1935). Functions of frontal association area in primates. *Archives of Neurology & Psychiatry*, 33(3):558–569.

Jessup, R. K., Busemeyer, J. R., and Brown, J. W. (2010). Error effects in anterior cingulate cortex reverse when error likelihood is high. *The Journal of Neuroscience*, 30(9):3467–3472.

Joel, D. and Weiner, I. (2000). The connections of the dopaminergic system with the striatum in rats and primates: an analysis with respect to the functional and compartmental organization of the striatum. *Neuroscience*, 96(3):451 – 474.

Johnson, A. and Redish, A. D. (2005). Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Networks*, 18(9):1163 – 1171.

Johnson, A. and Redish, A. D. (2007). Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point. *The Journal of Neuroscience*, 27(45):12176–12189.

Jones, J. L., Esber, G. R., McDannald, M. A., Gruber, A. J., Hernandez, A., Mirenzi, A., and Schoenbaum, G. (2012). Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science*, 338(6109):953–956.

Jueptner, M., Frith, C. D., Brooks, D. J., Frackowiak, R., and Passingham, R. E. (1997). Anatomy of motor learning. ii. subcortical structures and learning by trial and error. *Journal of Neurophysiology*, 77(3):1325–1337.

J.Yu, A. and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, 46(19):681 – 692.

Kable, J. W. and Glimcher, P. W. (2009). The neurobiology of decision: Consensus and controversy. *Neuron*, 63(6):733 – 745.

Kadohisa, M., Petrov, P., Stokes, M., Sigala, N., Buckley, M., Gaffan, D., Kusunoki, M., and Duncan, J. (2013). Dynamic construction of a coherent attentional state in a prefrontal cell population. *Neuron*, 80(1):235 – 246.

Kahneman, D. and Beatty, J. (1966). Pupil diameter and load on memory. *Science*, 154(3756):1583–1585.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–91.

Karlsson, M. P., Tervo, D. G. R., and Karpova, A. Y. (2012). Network resets in medial prefrontal cortex mark the onset of behavioral uncertainty. *Science*, 338(6103):135–139.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Katahira, K. (2015). The relation between reinforcement learning parameters and the influence of reinforcement history on choice behavior. *Journal of Mathematical Psychology*, 66(0):59 – 69.

Keay, K., Clement, C., Owler, B., Depaulis, A., and Bandler, R. (1994). Convergence of deep somatic and visceral nociceptive information onto a discrete ventrolateral midbrain periaqueductal gray region. *Neuroscience*, 61(4):727 – 732.

Kemp, J. M. and Powell, T. P. S. (1970). The corico-striate projection in the monkey. *Brain*, 93(3):525–546.

Kennerley, S. and Walton, M. (2011). Decision making and reward in frontal cortex: Complementary evidence from neurophysiological and neuropsychological studies. *Behavioral Neuroscience*, 125(3):297–317.

Kennerley, S. W., Behrens, T. E. J., and Wallis, J. D. (2011). Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. *Nat Neurosci*, 14(12):1581–1589.

Kennerley, S. W., Dahmubed, A. F., Lara, A. H., and Wallis, J. D. (2009). Neurons in the frontal lobe encode the value of multiple decision variables. *Journal of cognitive neuroscience*, 21(6):1162–1178.

Kennerley, S. W. and Wallis, J. D. (2009a). Evaluating choices by single neurons in the frontal lobe: outcome value encoded across multiple decision variables. *European Journal of Neuroscience*, 29(10):2061–2073.

Kennerley, S. W. and Wallis, J. D. (2009b). Reward-dependent modulation of working memory in lateral prefrontal cortex. *The Journal of Neuroscience*, 29(10):3259–3270.

Kennerley, S. W., Walton, M. E., Behrens, T. E. J., Buckley, M. J., and Rushworth, M. F. S. (2006). Optimal decision making and the anterior cingulate cortex. *Nature Neuroscience*, 9(7):940–947.

Keramati, M., Dezfouli, A., and Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput Biol*, 7(5):e1002055.

Killcross, S. and Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, 13(4):400–408.

Kim, H., Sul, J. H., Huh, N., Lee, D., and Jung, M. W. (2009). Role of striatum in updating values of chosen actions. *The Journal of Neuroscience*, 29(47):14701–14712.

Kim, H. F. and Hikosaka, O. (2013). Distinct basal ganglia circuits controlling behaviors guided by flexible and stable values. *Neuron*, 79(5):1001 – 1010.

Kim, S., Hwang, J., and Lee, D. (2008). Prefrontal coding of temporally discounted values during intertemporal choice. *Neuron*, 59(1):161 – 172.

Klossek, U. M. H., Russell, J., and Dickinson, A. (2008). The control of instrumental action following outcome devaluation in young children aged between 1 and 4 years. *Journal of Experimental Psychology: General*, 137(1):39–51.

Kobayashi, S., Pinto de Carvalho, O., and Schultz, W. (2010). Adaptation of reward sensitivity in orbitofrontal neurons. *The Journal of Neuroscience*, 30(2):534–544.

Koechlin, E. (2011). Frontal pole function: what is specifically human? *Trends in Cognitive Sciences*, 15(6):241 –.

Koechlin, E., Basso, G., Pietrini, P., Panzer, S., and Grafman, J. (1999). The role of the anterior prefrontal cortex in human cognition. *Nature*, 399(6732):148–151.

Koechlin, E., Corrado, G., Pietrini, P., and Grafman, J. (2000). Dissociating the role of the medial and lateral anterior prefrontal cortex in human planning. *Proceedings of the National Academy of Sciences*, 97(13):7651–7656.

Kolling, N., Behrens, T. E. J., Mars, R. B., and Rushworth, M. F. S. (2012). Neural mechanisms of foraging. *Science*, 336(6077):95–98.

Kunishio, K. and Haber, S. N. (1994). Primate cingulostriatal projection: Limbic striatal versus sensorimotor striatal input. *The Journal of Comparative Neurology*, 350(3):337–356.

Künzle, H. (1977). Projections from the primary somatosensory cortex to basal ganglia and thalamus in the monkey. *Experimental Brain Research*, 30(4):481–492.

Lashley, K. (1951). The problem of serial order in behavior. In Jeffress, L. A., editor, *Cerebral mechanisms in behavior: The Hixon Symposium*. John Wiley, New York.

Lau, B. and Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, 84(3):555–579.

Lau, B. and Glimcher, P. W. (2008). Value representations in the primate striatum during matching behavior. *Neuron*, 58(3):451 – 463.

Lavin, C., San Martín, R., and Rosales Jubal, E. (2014). Pupil dilation signals uncertainty and surprise in a learning gambling task. *Frontiers in Behavioral Neuroscience*, 7(218).

Lebedev, M. A., Messinger, A., Kralik, J. D., and Wise, S. P. (2004). Representation of attended versus remembered locations in prefrontal cortex. *PLoS Biol*, 2(11):e365.

Lee, D. and Seo, H. (2007). Mechanisms of reinforcement learning and decision making in the primate dorsolateral prefrontal cortex. *Annals of the New York Academy of Sciences*, 1104(1):108–122.

Lee, D., Seo, H., and Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience*, 35(1):287–308. PMID: 22462543.

Lee, S. W., Shimojo, S., and O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, 81(3):687 – 699.

Lehéricy, S., Benali, H., Van de Moortele, P.-F., Pélégrini-Issac, M., Waechter, T., Ugurbil, K., and Doyon, J. (2005). Distinct basal ganglia territories are engaged in early and advanced motor sequence learning. *Proceedings of the National Academy of Sciences of the United States of America*, 102(35):12566–12571.

Leon, M. I. and Shadlen, M. N. (1999). Effect of expected reward magnitude on the response of neurons in the dorsolateral prefrontal cortex of the macaque. *Neuron*, 24(2):415 – 425.

Levy, R. and Goldman-Rakic, P. (2000). Segregation of working memory functions within the dorsolateral prefrontal cortex. In Schneider, W., Owen, A., and Duncan, J., editors, *Executive Control and the Frontal Lobe: Current Issues*, pages 23–32. Springer Berlin Heidelberg.

Li, L., Walsh, T. J., and Littman, M. L. (2006). Towards a unified theory of state abstraction for mdps. In *In Proceedings of the Ninth International Symposium on Artificial Intelligence and Mathematics*, pages 531–539.

Lingawi, N. W. and Balleine, B. W. (2012). Amygdala central nucleus interacts with dorsolateral striatum to regulate the acquisition of habits. *The Journal of Neuroscience*, 32(3):1073–1081.

Little, J. T., Johnson, D. N., Johnson, D. N., Weingartner, H., and Sunderland, T. (1998). Combined nicotinic and muscarinic blockade in elderly normal volunteers: Cognitive, behavioral, and physiologic responses. *Neuropsychopharmacology*, 19(1):574 – 582.

Ljungberg, T., Apicella, P., and Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, 67(1):145–163.

Loomes, G. and Sugden, R. (1982). Regret theory: An alternative theory of rational choice under uncertainty. *The Economic Journal*, 92(368):pp. 805–824.

Lynd-Balta, E. and Haber, S. (1994). The organization of midbrain projections to the ventral striatum in the primate. *Neuroscience*, 59(3):609 – 623.

Maia, T. V. and Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nat Neurosci*, 14(2):154–162.

Mansouri, F. A., Buckley, M. J., Mahboubi, M., and Tanaka, K. (2015). Behavioral consequences of selective damage to frontal pole and posterior cingulate cortices. *Proceedings of the National Academy of Sciences*, 112(29):E3940–E3949.

Mansouri, F. A., Buckley, M. J., and Tanaka, K. (2007). Mnemonic function of the dorsolateral prefrontal cortex in conflict-induced behavioral adjustment. *Science*, 318(5852):987–990.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA.

Matsumoto, K., Suzuki, W., and Tanaka, K. (2003). Neuronal correlates of goal-based motor selection in the prefrontal cortex. *Science*, 301(5630):229–232.

Matsumoto, M., Matsumoto, K., Abe, H., and Tanaka, K. (2007). Medial prefrontal cell activity signaling prediction errors of action values. *Nature Neuroscience*, 10(5):647–656.

McClure, S. M., Berns, G. S., and Montague, P. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, 38(2):339 – 346.

McFarland, N. R. and Haber, S. N. (2000). Convergent inputs from thalamic motor nuclei and frontal cortical areas to the dorsal striatum in the primate. *The Journal of Neuroscience*, 20(10):3798–3813.

Medalla, M. and Barbas, H. (2010). Anterior cingulate synapses in prefrontal areas 10 and 46 suggest differential influence in cognitive control. *The Journal of Neuroscience*, 30(48):16068–16081.

Menzel, R. and Fischer, J. (2011). *Animal Thinking: Contemporary Issues in Comparative Cognition ; [Eighth Ernst Strüngmann Forum Held Sep. 26 - Oct. 1, 2010, Frankfurt Am Main]*. Strüngmann Forum reports. Mit Press.

Miller, E. K. and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24(1):167–202. PMID: 11283309.

Mishkin, M., Suzuki, W. A., Gadian, D. G., and Vargha–Khadem, F. (1997). Hierarchical organization of cognitive memory. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 352(1360):1461–1467.

Miyachi, S., Hikosaka, O., and Lu, X. (2002). Differential activation of monkey striatal neurons in the early and late stages of procedural learning. *Experimental Brain Research*, 146(1):122–126.

Miyachi, S., Hikosaka, O., Miyashita, K., Kárádi, Z., and Rand, M. K. (1997). Differential roles of monkey striatum in learning of sequential hand movement. *Experimental Brain Research*, 115(1):1–5.

Monosov, I. E. and Hikosaka, O. (2012). Regionally distinct processing of rewards and punishments by the primate ventromedial prefrontal cortex. *The Journal of Neuroscience*, 32(30):10318–10330.

Montague, P., Dayan, P., and Sejnowski, T. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, 16:1936–1947.

Montague, P. R., Dayan, P., Person, C., and Sejnowski, T. J. (1995). Bee foraging in uncertain environments using predictive hebbian learning. *Nature*, 377(October):725–728.

Moore, A. and Atkeson, C. (1993). Prioritized sweeping: Reinforcement learning with less data and less time. *Machine Learning*, 13(1):103–130.

Morecraft, R. J. and Hoesen, G. W. V. (1998). Convergence of limbic input to the cingulate motor cortex in the rhesus monkey. *Brain Research Bulletin*, 45(2):209 – 232.

Morecraft, R. J. and van Hoesen, G. W. (1993). Frontal granular cortex input to the cingulate (m3), supplementary (m2) and primary (m1) motor cortices in the rhesus monkey. *The Journal of Comparative Neurology*, 337(4):669–689.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., and Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience*, 9(8):1057–1063.

Morrison, S. E., Saez, A., Lau, B., and Salzman, C. D. (2011). Different time courses for learning-related changes in amygdala and orbitofrontal cortex. *Neuron*, 71(6):1127 – 1140.

Muenzinger, K. F. (1938). Vicarious trial and error at a point of choice: I. a general survey of its relation to learning efficiency. *The Pedagogical Seminary and Journal of Genetic Psychology*, 53(1):75–86.

Murray, E. A., Bussey, T. J., and Wise, S. P. (2000). Role of prefrontal cortex in a network for arbitrary visuomotor mapping. *Experimental Brain Research*, 133(1):114–129.

Nakano, K., Kayahara, T., and Chiba, T. (1999). Afferent connections to the ventral striatum from the medial prefrontal cortex (area 25) and the thalamic nuclei in the macaque monkey. *Annals of the New York Academy of Sciences*, 877(1):667–670.

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., and Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature neuroscience*, 15(7):1040–1046.

Neubert, F.-X., Mars, R. B., Sallet, J., and Rushworth, M. F. S. (2015). Connectivity reveals relationship of brain areas for reward-guided learning and decision making in human and monkey frontal cortex. *Proceedings of the National Academy of Sciences*, 112(20):E2695–E2704.

Neubert, F.-X., Mars, R. B., Thomas, A. G., Sallet, J., and Rushworth, M. F. (2014). Comparison of human ventral frontal cortex areas for cognitive control and language with areas in monkey frontal cortex. *Neuron*, 81(3):700 – 713.

Niv, Y., Daw, N., Joel, D., and Dayan, P. (2007). Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology*, 191(3):507–520.

Niv, Y., Daw, N. D., and Dayan, P. (2006). Choice values. *Nature Neuroscience*, 9(8):987–988.

Noonan, M. P., Walton, M. E., Behrens, T. E. J., Sallet, J., Buckley, M. J., and Rushworth, M. F. S. (2010). Separate value comparison and learning mechanisms in macaque medial and lateral orbitofrontal cortex. *Proceedings of the National Academy of Sciences*, 107(47):20547–20552.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., and Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, 304(5669):452–454.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., and Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2):329 – 337.

Ostlund, S. B. and Balleine, B. W. (2005). Lesions of medial prefrontal cortex disrupt the acquisition but not the expression of goal-directed learning. *The Journal of Neuroscience*, 25(34):7763–7770.

Otto, A. R., Gershman, S. J., Markman, A. B., and Daw, N. D. (2013). The curse of planning: Dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological Science*.

Otto, A. R., Skatova, A., Madlon-Kay, S., and Daw, N. D. (2014). Cognitive control predicts use of model-based reinforcement learning. *Journal of Cognitive Neuroscience*, 27(2):319–333.

Oyama, K., Hernádi, I., Iijima, T., and Tsutsui, K.-I. (2010). Reward prediction error coding in dorsal striatal neurons. *The Journal of Neuroscience*, 30(34):11447–11457.

O'Reilly, J. X., Schüffelgen, U., Cuell, S. F., Behrens, T. E. J., Mars, R. B., and Rushworth, M. F. S. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences*, 110(38):E3660–E3669.

Padoa-Schioppa, C. (2011). Neurobiology of Economic Choice: A Good-Based Model. In Hyman, SE and Jessell, TM and Shatz, CJ and Stevens, CF and Zoghbi, HY, editor, *ANNUAL REVIEW OF NEUROSCIENCE, VOL 34*, volume 34 of *Annual Review of Neuroscience*, pages 333–359.

Padoa-Schioppa, C. and Assad, J. A. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature*, 441(7090):223–226.

Pandya, D., Van Hoesen, G., and Mesulam, M.-M. (1981). Efferent connections of the cingulate gyrus in the rhesus monkey. *Experimental Brain Research*, 42(3-4):319–330.

Parikh, V., Kozak, R., Martinez, V., and Sarter, M. (2007). Prefrontal acetylcholine release controls cue detection on multiple timescales. *Neuron*, 56(1):141 – 154.

Pasupathy, A. and Miller, E. K. (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*, 433(7028):873–876.

Paton, J. J., Belova, M. A., Morrison, S. E., and Salzman, C. D. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature*, 439(7078):865–870.

Paulus, M. P., Hozack, N., Frank, L., and Brown, G. G. (2002). Error rate and outcome predictability affect neural activation in prefrontal cortex and anterior cingulate during decision-making. *NeuroImage*, 15(4):836 – 846.

Percheron, G. and Filion, M. (1991). Parallel processing in the basal ganglia: up to a point. *Trends in Neurosciences*, 14:55–56.

Petrides, M. (1995). Impairments on nonspatial self-ordered and externally ordered working memory tasks after lesions of the mid-dorsal part of the lateral frontal cortex in the monkey. *The Journal of Neuroscience*, 15(1):359–375.

Petrides, M. (2005a). Lateral prefrontal cortex: architectonic and functional organization. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1456):781–795.

Petrides, M. (2005b). Lateral prefrontal cortex: architectonic and functional organization. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1456):781–795.

Petrides, M. and Pandya, D. N. (2004). Chapter 25 - the frontal cortex. In Mai, G. P. K., editor, *The Human Nervous System (Second Edition)*, pages 950 – 972. Academic Press, San Diego, second edition edition.

Petrides, M. and Pandya, D. N. (2007). Efferent association pathways from the rostral prefrontal cortex in the macaque monkey. *The Journal of Neuroscience*, 27(43):11573–11586.

Petrides, M. and Pandya, D. N. (2012). Chapter 26 - the frontal cortex. In Paxinos, J. K. M., editor, *The Human Nervous System (Third Edition)*, pages 988 – 1011. Academic Press, San Diego, third edition edition.

Pinto, L., Goard, M. J., Estandian, D., Xu, M., Kwan, A. C., Lee, S.-H., Harrison, T. C., Feng, G., and Dan, Y. (2013). Fast modulation of visual perception by basal forebrain cholinergic neurons. *Nat Neurosci*, 16(12):1857–1863.

Platt, M. L. and Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, 400(6741):233–238.

Porrino, L. J. and Goldman-Rakic, P. S. (1982). Brainstem innervation of prefrontal and anterior cingulate cortex in the rhesus monkey revealed by retrograde transport of hrp. *The Journal of Comparative Neurology*, 205(1):63–76.

Preuschoff, K., 't Hart, B. M., and Einhauser, W. (2011). Pupil dilation signals surprise: evidence for noradrenaline's role in decision making. *Frontiers in Neuroscience*, 5(115).

Preuss, T. M. (1995). Do rats have prefrontal cortex? the rose-woolsey-akert program reconsidered. *Journal of Cognitive Neuroscience*, 1:1–26.

Price, J. L. (2007). Definition of the orbital cortex in relation to specific connections with limbic and visceral structures and other cortical regions. *Annals of the New York Academy of Sciences*, 1121(1):54–71.

Quilodran, R., Rothé, M., and Procyk, E. (2008). Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron*, 57(2):314 – 325.

Raghanti, M., Stimpson, C., Marcinkiewicz, J., Erwin, J., Hof, P., and Sherwood, C. (2008). Cortical dopaminergic innervation among humans, chimpanzees, and macaque monkeys: A comparative study. *Neuroscience*, 155(1):203 – 220.

Ramnani, N., Elliott, R., Athwal, B., and Passingham, R. (2004). Prediction error for free monetary reward in the human prefrontal cortex. *NeuroImage*, 23(3):777 – 786.

Rao, S. C., Rainer, G., and Miller, E. K. (1997). Integration of what and where in the primate prefrontal cortex. *Science*, 276(5313):821–824.

Rescorla, R. A. and Wagner, A. W. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical Conditioning II: Current Research and Theory*, pages 64–99. Appleton-Century-Crofts, New York.

Roesch, M. R., Calu, D. J., and Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience*, 10(12):1615–1624.

Roesch, M. R. and Olson, C. R. (2003). Impact of expected reward on neuronal activity in prefrontal cortex, frontal and supplementary eye fields and premotor cortex. *Journal of Neurophysiology*, 90(3):1766–1789.

Romo, R., Scarnati, E., and Schultz, W. (1992). Role of primate basal ganglia and frontal cortex in the internal generation of movements. *Experimental Brain Research*, 91(3):385–395.

Romo, R. and Schultz, W. (1990). Dopamine neurons of the monkey midbrain: contingencies of responses to active touch during self-initiated arm movements. *Journal of Neurophysiology*, 63(3):592–606.

Rosene, D. and Van Hoesen, G. (1977). Hippocampal efferents reach widespread areas of cerebral cortex and amygdala in the rhesus monkey. *Science*, 198(4314):315–317.

Rudebeck, P. H., Behrens, T. E., Kennerley, S. W., Baxter, M. G., Buckley, M. J., Walton, M. E., and Rushworth, M. F. S. (2008). Frontal cortex subregions play distinct roles in choices between actions and stimuli. *The Journal of Neuroscience*, 28(51):13775–13785.

Rudebeck, P. H. and Murray, E. A. (2008). Amygdala and orbitofrontal cortex lesions differentially influence choices during object reversal learning. *The Journal of Neuroscience*, 28(33):8338–8343.

Rudebeck, P. H. and Murray, E. A. (2011). Dissociable effects of subtotal lesions within the macaque orbital prefrontal cortex on reward-guided behavior. *The Journal of Neuroscience*, 31(29):10569–10578.

Rudebeck, P. H., Saunders, R. C., Prescott, A. T., Chau, L. S., and Murray, E. A. (2013). Prefrontal mechanisms of behavioral flexibility, emotion regulation and value updating. *Nature Neuroscience*, 16(8):1140–1145.

Rummery, G. A. and Niranjan, M. (1994). On-line Q-learning using connectionist systems. Technical Report TR 166, Cambridge University Engineering Department, Cambridge, England.

Rushworth, M. F., Kolling, N., Sallet, J., and Mars, R. B. (2012). Valuation and decision-making in frontal cortex: one or many serial or parallel systems? *Current Opinion in Neurobiology*, 22(6):946 – 955. Decision making.

Rushworth, M. F., Noonan, M. P., Boorman, E. D., Walton, M. E., and Behrens, T. E. (2011). Frontal cortex and reward-guided learning and decision-making. *Neuron*, 70(6):1054 – 1069.

Rushworth, M. F. S. and Behrens, T. E. J. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, 11(4):389–397.

Saint-Cyr, J. A., Ungerleider, L. G., and Desimone, R. (1990). Organization of visual cortical inputs to the striatum and subsequent outputs to the pallido-nigral complex in the monkey. *The Journal of Comparative Neurology*, 298(2):129–156.

Saleem, K. S., Miller, B., and Price, J. L. (2014). Subdivisions and connectional networks of the lateral prefrontal cortex in the macaque monkey. *Journal of Comparative Neurology*, 522(7):1641–1690.

Sallet, J., Mars, R. B., Quilodran, R., Procyk, E., Petrides, M., and Rushworth, M. F. S. (2011). Neuroanatomical basis of motivational and cognitive control: A focus on the medial and lateral prefrontal cortex. In Mars, R. B., Sallet, J., Rushworth, M. F. S., and Yeung, N., editors, *Neural Basis of Motivational and Cognitive Control*, pages 5–20. The MIT Press.

Samejima, K. and Doya, K. (2007). Multiple representations of belief states and action values in corticobasal ganglia loops. *Annals of the New York Academy of Sciences*, 1104(1):213–228.

Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, 310(5752):1337–1340.

Samuels, E. R. and Szabadi, E. (2008). Functional neuroanatomy of the noradrenergic locus coeruleus: its roles in the regulation of arousal and autonomic function part ii: physiological and pharmacological manipulations and pathological alterations of locus coeruleus activity in humans. *Curr Neuropharmacol*, 6(3):254–285.

Sara, S. J. (2009). The locus coeruleus and noradrenergic modulation of cognition. *Nat Rev Neurosci*, 10(3):211–223.

Satterthwaite, T. D., Green, L., Myerson, J., Parker, J., Ramaratnam, M., and Buckner, R. L. (2007). Dissociable but inter-related systems of cognitive control and reward during decision making: Evidence from pupillometry and event-related fmri. *NeuroImage*, 37(3):1017 – 1031.

Sawaguchi, T. and Goldman-Rakic, P. (1991). D1 dopamine receptors in prefrontal cortex: involvement in working memory. *Science*, 251(4996):947–950.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, 36(2):241 – 263.

Schultz, W. (2007). Multiple dopamine functions at different time courses. *ANNUAL REVIEW OF NEUROSCIENCE*, 30:259–288.

Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, 13:900–913.

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

Schultz, W. and Romo, R. (1990). Dopamine neurons of the monkey midbrain: contingencies of responses to stimuli eliciting immediate behavioral reactions. *Journal of Neurophysiology*, 63(3):607–624.

Schultz, W. and Romo, R. (1992). Role of primate basal ganglia and frontal cortex in the internal generation of movements. *Experimental Brain Research*, 91(3):363–384.

Schultz, W., Tremblay, L., and Hollerman, J. R. (2003). Changes in behavior-related neuronal activity in the striatum during learning. *Trends in Neurosciences*, 26(6):321 – 328.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Seamans, J. K. and Yang, C. R. (2004). The principal features and mechanisms of dopamine modulation in the prefrontal cortex. *Progress in Neurobiology*, 74(1):1 – 58.

Selemon, L. and Goldman-Rakic, P. (1985). Longitudinal topography and interdigitation of corticostriatal projections in the rhesus monkey. *The Journal of Neuroscience*, 5(3):776–794.

Semendeferi, K., Armstrong, E., Schleicher, A., Zilles, K., and Van Hoesen, G. W. (2001). Prefrontal cortex in humans and apes: A comparative study of area 10. *American Journal of Physical Anthropology*, 114(3):224–241.

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

Seo, H. and Lee, D. (2007). Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. *The Journal of Neuroscience*, 27(31):8366–8377.

Seo, M., Lee, E., and Averbeck, B. B. (2012). Action selection and action value in frontal-striatal circuits. *Neuron*, 74(5):947 – 960.

Sesack, S. R. and Grace, A. A. (2009). Cortico-basal ganglia reward network: Microcircuitry. *Neuropsychopharmacology*, 35(1):27–47.

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., Friston, K. J., and Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans. *Nature*, 429(6992):664–667.

Shenhav, A., Botvinick, M. M., and Cohen, J. D. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, 79(2):217 – 240.

Shidara, M., Aigner, T. G., and Richmond, B. J. (1998). Neuronal signals in the monkey ventral striatum related to progress through a predictable series of trials. *The Journal of Neuroscience*, 18(7):2613–2625.

Shima, K. and Tanji, J. (1998). Role for cingulate motor area cells in voluntary movement selection based on reward. *Science*, 282(5392):1335–1338.

Silvetti, M., Seurinck, R., van Bochove, M., and Verguts, T. (2013a). The influence of the noradrenergic system on optimal control of neural plasticity. *Frontiers in Behavioral Neuroscience*, 7(160).

Silvetti, M., Seurinck, R., and Verguts, T. (2013b). Value and prediction error estimation account for volatility effects in acc: A model-based fmri study. *Cortex*, 49(6):1627 – 1635.

Simon, D. A. and Daw, N. D. (2011). Environmental statistics and the trade-off between model-based and td learning in humans. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 24*, pages 127–135. Curran Associates, Inc.

Skinner, B. F. (1948). 'Superstition' in the pigeon. *Journal of Experimental Psychology*, 38(2):168–172.

Smith, K. S. and Graybiel, A. M. (2014). Investigating habits: Strategies, technologies, and models. *Frontiers in Behavioral Neuroscience*, 8(39).

Smith, K. S., Virkud, A., Deisseroth, K., and Graybiel, A. M. (2012). Reversible online control of habitual behavior by optogenetic perturbation of medial prefrontal cortex. *Proceedings of the National Academy of Sciences*, 109(46):18932–18937.

Smittenaar, P., FitzGerald, T. H., Romei, V., Wright, N. D., and Dolan, R. J. (2013). Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron*, 80(4):914 – 919.

Sowell, E. R., Thompson, P. M., Leonard, C. M., Welcome, S. E., Kan, E., and Toga, A. W. (2004). Longitudinal mapping of cortical thickness and brain growth in normal children. *The Journal of Neuroscience*, 24(38):8223–8231.

Squire, L. R., Stark, C. E., and Clark, R. E. (2004). The medial temporal lobe*. *Annual Review of Neuroscience*, 27(1):279–306. PMID: 15217334.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., and Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, 46(4):1004 – 1017.

Suri, R. and Schultz, W. (1999). A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, 91(3):871 – 890.

Sutton, R. S. (1990). Integrated architecture for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the seventh international conference (1990) on Machine learning*, pages 216–224, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Sutton, R. S. and Barto, A. G. (1998). *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition.

Sutton, R. S., Precup, D., and Singh, S. (1999). Between {MDPs} and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1–2):181 – 211.

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

Tamar, A., Di Castro, D., and Meir, R. (2012). Integrating a partial model into model free reinforcement learning. *J. Mach. Learn. Res.*, 13:1927–1966.

Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., and Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience*, 7(8):887–893.

Thomas, W. (1664). *Cerebri Anatome: Cui Accessit Nervorum Descriptio Et Usus*. Nabu Press.

Thorndike, E. L. (1911.). *Animal intelligence;*. New York,The Macmillan company,. http://www.biodiversitylibrary.org/bibliography/1201.

Thorpe, S., Rolls, E., and Maddison, S. (1983). The orbitofrontal cortex: Neuronal activity in the behaving monkey. *Experimental Brain Research*, 49(1):93–115.

Tobler, P. N., Fiorillo, C. D., and Schultz, W. (2005). Adaptive Coding of Reward Value by Dopamine Neurons. *Science*, 307(5715):1642–1645.

Tolman, E. C. (1938). The determiners of behavior at a choice point. *Psychological review*, 45(1):1–41. NR: 153 reference(s) present, 153 reference(s) displayed RX: 107 (on Nov 16, 2007).

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological review*, 55(4):189–208.

Tolman, E. C., Ritchie, B. F., and Kalish, D. (1946). Studies in spatial learning. i. orientation and the short-cut. *Journal of Experimental Psychology*, 36(1):13–24.

Tran-Tu-Yen, D. A. S., Marchand, A. R., Pape, J.-R., Di Scala, G., and Coutureau, E. (2009). Transient role of the rat prelimbic cortex in goal-directed behaviour. *European Journal of Neuroscience*, 30(3):464–471.

Tremblay, L. and Schultz, W. (1999). Relative reward preference in primate orbitofrontal cortex. *Nature*, 398(6729):704–708.

Tricomi, E., Balleine, B. W., and O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience*, 29(11):2225–2232.

Tsujimoto, S., Genovesio, A., and Wise, S. P. (2010). Evaluating self-generated decisions in frontal pole cortex of monkeys. *Nat Neurosci*, 13(1):120–126.

Tsujimoto, S., Genovesio, A., and Wise, S. P. (2012). Neuronal activity during a cued strategy task: Comparison of dorsolateral, orbital, and polar prefrontal cortex. *The Journal of Neuroscience*, 32(32):11017–11031.

Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., and Aston-Jones, G. (1999). The role of locus coeruleus in the regulation of cognitive performance. *Science*, 283(5401):549–554.

Valentin, V. V., Dickinson, A., and O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *The Journal of Neuroscience*, 27(15):4019–4026.

Van Hoesen, G. W., Yeterian, E. H., and Lavizzo-Mourey, R. (1981). Widespread corticostriate projections from temporal cortex of the rhesus monkey. *The Journal of Comparative Neurology*, 199(2):205–219.

van Veen, V., Cohen, J. D., Botvinick, M. M., Stenger, V., and Carter, C. S. (2001). Anterior cingulate cortex, conflict monitoring, and levels of processing. *NeuroImage*, 14(6):1302 – 1308.

Varazzani, C., San-Galli, A., Gilardeau, S., and Bouret, S. (2015). Noradrenaline and dopamine neurons in the reward/effort trade-off: A direct electrophysiological comparison in behaving monkeys. *The Journal of Neuroscience*, 35(20):7866–7877.

Vijayraghavan, S., Wang, M., Birnbaum, S. G., Williams, G. V., and Arnsten, A. F. T. (2007). Inverted-u dopamine d1 receptor actions on prefrontal neurons engaged in working memory. *Nat Neurosci*, 10(3):376–384.

Vogt, B. A., Finch, D. M., and Olson, C. R. (1992). Functional heterogeneity in cingulate cortex: The anterior executive and posterior evaluative regions. *Cerebral Cortex*, 2(6):435–443.

Vogt, B. A. and Pandya, D. N. (1987). Cingulate cortex of the rhesus monkey: Ii. cortical afferents. *The Journal of Comparative Neurology*, 262(2):271–289.

Vogt, B. A., Vogt, L., Farber, N. B., and Bush, G. (2005). Architecture and neurocytology of monkey cingulate gyrus. *The Journal of Comparative Neurology*, 485(3):218–239.

Vogt, C. and Vogt, O. (1941). Thalamusstudien i-iii. *J. Psychol. Neurol.*, 50:33–154.

Wallis, J. D., Anderson, K. C., and Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature*, 411(6840):953–956.

Walton, M. E., Behrens, T. E., Buckley, M. J., Rudebeck, P. H., and 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.

Wang, L. P., Li, F., Wang, D., Xie, K., Wang, D., Shen, X., and Tsien, J. Z. (2011). {NMDA} receptors in dopaminergic neurons are crucial for habit learning. *Neuron*, 72(6):1055 – 1066.

Watanabe, M. (1996). Reward expectancy in primate prefrontal neurons. *Nature*, 382:629–632.

Watkins, C. J. C. H. (1989). *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, UK.

Weinshenker, D. and Schroeder, J. P. (2007). There and back again: a tale of norepinephrine and drug addiction. *Neuropsychopharmacology*, 32:1433 – 1451. 7.

White, I. M. and Wise, S. P. (1999). Rule-dependent neuronal activity in the prefrontal cortex. *Experimental Brain Research*, 126(3):315–335.

Williams, S. M. and Goldman-Rakic, P. S. (1998). Widespread origin of the primate mesofrontal dopamine system. *Cerebral Cortex*, 8(4):321–345.

Williams, Z. M. and Eskandar, E. N. (2006). Selective enhancement of associative learning by microstimulation of the anterior caudate. *Nat Neurosci*, 9(4):562–568.

Wise, S. P. (2008). Forward frontal fields: phylogeny and fundamental function. *Trends in Neurosciences*, 31(12):599 – 608.

Wunderlich, K., Dayan, P., and Dolan, R. J. (2012a). Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience*, 15(5):786–791.

Wunderlich, K., Smittenaar, P., and Dolan, R. J. (2012b). Dopamine enhances model-based over model-free choice behavior. *Neuron*, 75(3):418 – 424.

Yeterian, E. and Pandya, D. (1998). Corticostriatal connections of the superior temporal region in rhesus monkeys. *The Journal of Comparative Neurology*, 399(3):384–402.

Yeterian, E. H. and Hoesen, G. W. V. (1978). Cortico-striate projections in the rhesus monkey: The organization of certain cortico-caudate connections. *Brain Research*, 139(1):43 – 63.

Yeterian, E. H. and Pandya, D. N. (1991). Prefrontostriatal connections in relation to cortical architectonic organization in rhesus monkeys. *The Journal of Comparative Neurology*, 312(1):43–67.

Yin, H. H., Knowlton, B. J., and Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *European Journal of Neuroscience*, 19(1):181–189.

Yin, H. H., Knowlton, B. J., and Balleine, B. W. (2005a). Blockade of nmda receptors in the dorsomedial striatum prevents action–outcome learning in instrumental conditioning. *European Journal of Neuroscience*, 22(2):505–512.

Yin, H. H., Knowlton, B. J., and Balleine, B. W. (2006). Inactivation of dorsolateral striatum enhances sensitivity to changes in the action–outcome contingency in instrumental conditioning. *Behavioural Brain Research*, 166(2):189 – 196.

Yin, H. H., Ostlund, S. B., Knowlton, B. J., and Balleine, B. W. (2005b). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, 22(2):513–523.

Ó Scalaidhe, S. P., Wilson, F. A., and Goldman-Rakic, P. S. (1999). Face-selective neurons during passive viewing and working memory performance of rhesus monkeys: Evidence for intrinsic specialization of neuronal coding. *Cerebral Cortex*, 9(5):459–475.

Öngür, D., An, X., and Price, J. (1998). Prefrontal cortical projections to the hypothalamus in macaque monkeys. *The Journal of Comparative Neurology*, 401(4):480–505.

Öngür, D. and Price, J. (2000). The organization of networks within the orbital and medial prefrontal cortex of rats, monkeys and humans. *Cerebral Cortex*, 10(3):206–219.

Appendix A

Behavioural analysis and recordings in Subject J

The disparity in the number of recorded neurons between the two subjects is because of a health-related issue in Subject J. Subject J exhibited an isolated epileptic episode just prior to the start of electrophysiological recordings. After several days of observation and behavioural testing without another epileptic incident, electrophysiological recordings commenced. We collected 16 sessions (of the total 27 sessions) worth of behavioural and electrophysiological data without incident, when epileptic episodes returned. Under the advice of the Named Veterinary Surgeon and the Home Office Inspectorate, Subject J was placed on a long-lasting course of phenobarbital and over the next four months of observation and adjusting drug levels, we settled on a dose of 30mg phenobarbital given in the morning and evening of each day (60mg total/day).

Subject J maintained task motivation on this drug dosage and could perform the MB task normally. We collected another 11 sessions of behavioural and neurophysiological data (left hemisphere mostly and only one of the sessions from right hemisphere) with the subject on this drug dosage without any epileptic incidents before we ceased data collection with this subject. An analysis of the behavioural data, when considering the session effects present in both subjects (see Fig. A.1 and A.2; Table A.1), failed to demonstrate that the drug had a main effect on behaviour. The overall percentage of task-selective neurons (see Fig. A.3 and A.4) as well as the profiles of the population coding (see Fig. A.5 and A.6) were also relatively similar in the drug off/on sessions, as such, we pooled all sessions together in both the behavioural and neuronal analyses.

To note that for the analysis we adopted a comparative analysis process between first part (i.e., for subject J it corresponds to before the medication, 16 sessions out of 27; for

subject C it corresponds to 18 sessions out of 30 as this is the approximate same proportion of sessions out of the total, around 60%, as in subject J) and second part (remaining 11 sessions in subject J; remaining 12 sessions for subject C), and using two-sample t-tests or χ^2 tests. Statistical significance was evaluated at $\alpha = 0.05$ and $\alpha = 0.01$.

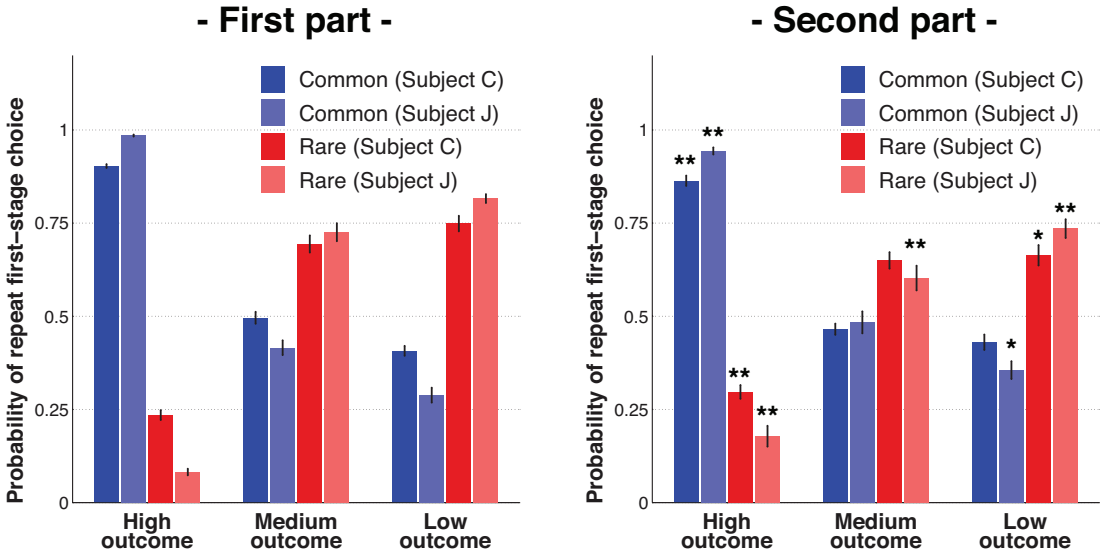


Fig. A.1 **Comparison between first and second parts of recordings for the impact of both reward and transition information on observed first-stage choice behaviour** The probability of repeating the same first-stage choice, averaged across sessions for each subject, as a function of outcome level and transition type (common transition in blue; rare transition in red) of the previous trial. First part corresponds to the first 18(C)/16(J) sessions out of a total of 30(C)/27(J). Error bars depict standard errors of the mean. ** for $\alpha = 0.01$ and * for $\alpha = 0.05$ in two-sample t-test with null-hypothesis that the probabilities of repeating across sessions between the corresponding first and second parts come from independent random samples from normal distributions with equal means and equal but unknown variances.

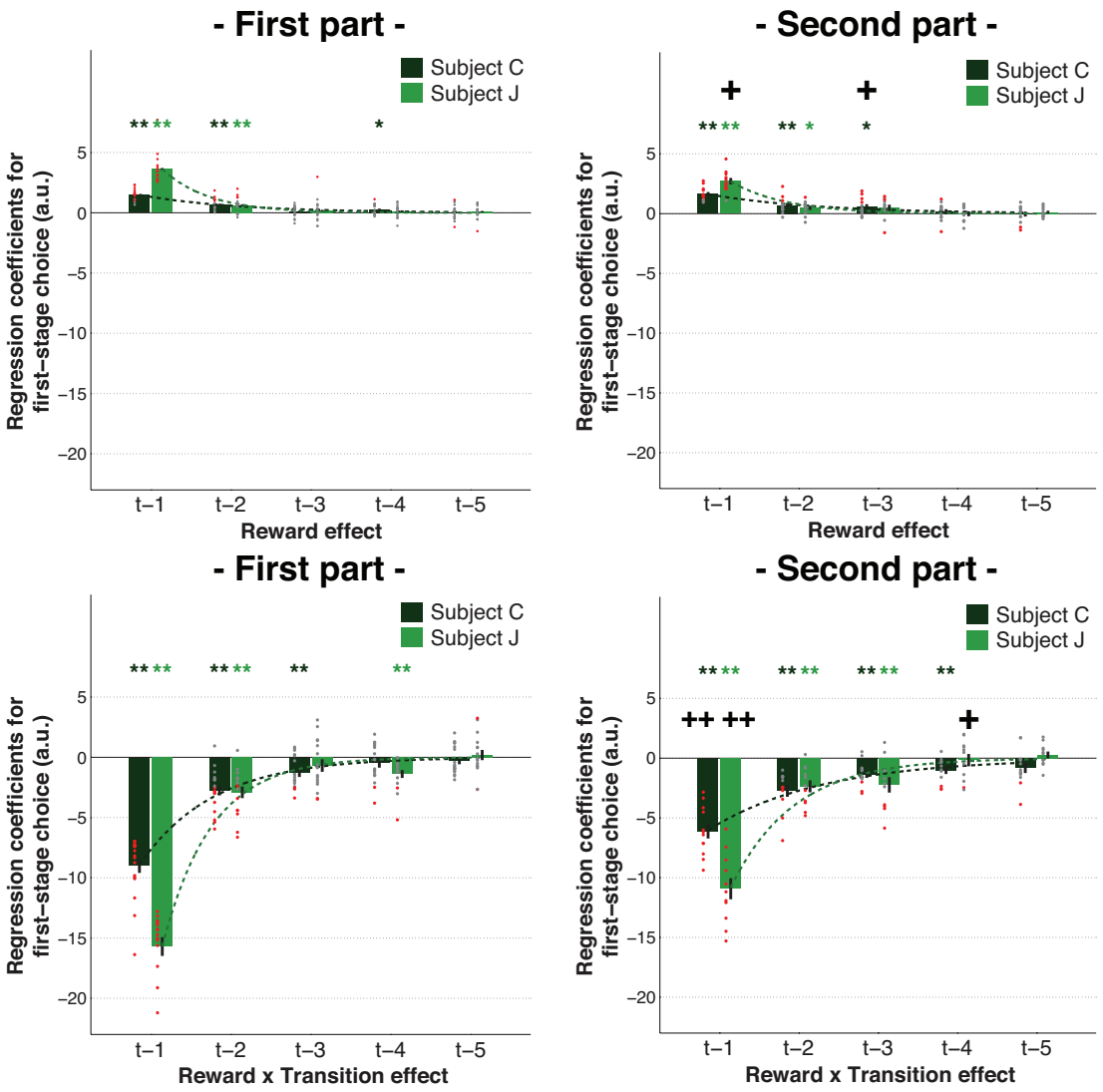


Fig. A.2 **Comparison between first and second parts of recordings for the logistic regression on observed first-stage chosen picture using predictor variables from the five previous trials.** For the given trial t, the variables used as predictors of the dependent variable first-stage choice (1=car picture, 0=watering can picture) were: C is first-stage choice (1=car picture, 0=watering can picture); R is outcome level (assumed as continuous and with low=1, medium=2, high=3); and T is transition (rare=1, common=0). Const is the constant term. Predictors were mean centred and continuous variables were also scaled by dividing them by two standard deviations (adjustments made before the computation of the interaction terms). First part corresponds to the first 18(C)/16(J) sessions out of a total of 30(C)/27(J). Each dot represents the fixed-effects regression estimate for a given session (coloured red when $p < 0.05$ and grey otherwise). Bar and error bar values correspond, respectively, to the mean and SE of the fixed-effects coefficients. $^{++}$ for $\alpha = 0.01$ and $^{+}$ for $\alpha = 0.05$ in two-sample t-test with null-hypothesis that the regression coefficients across sessions between the corresponding first and second parts come from independent random samples from normal distributions with equal means and equal but unknown variances

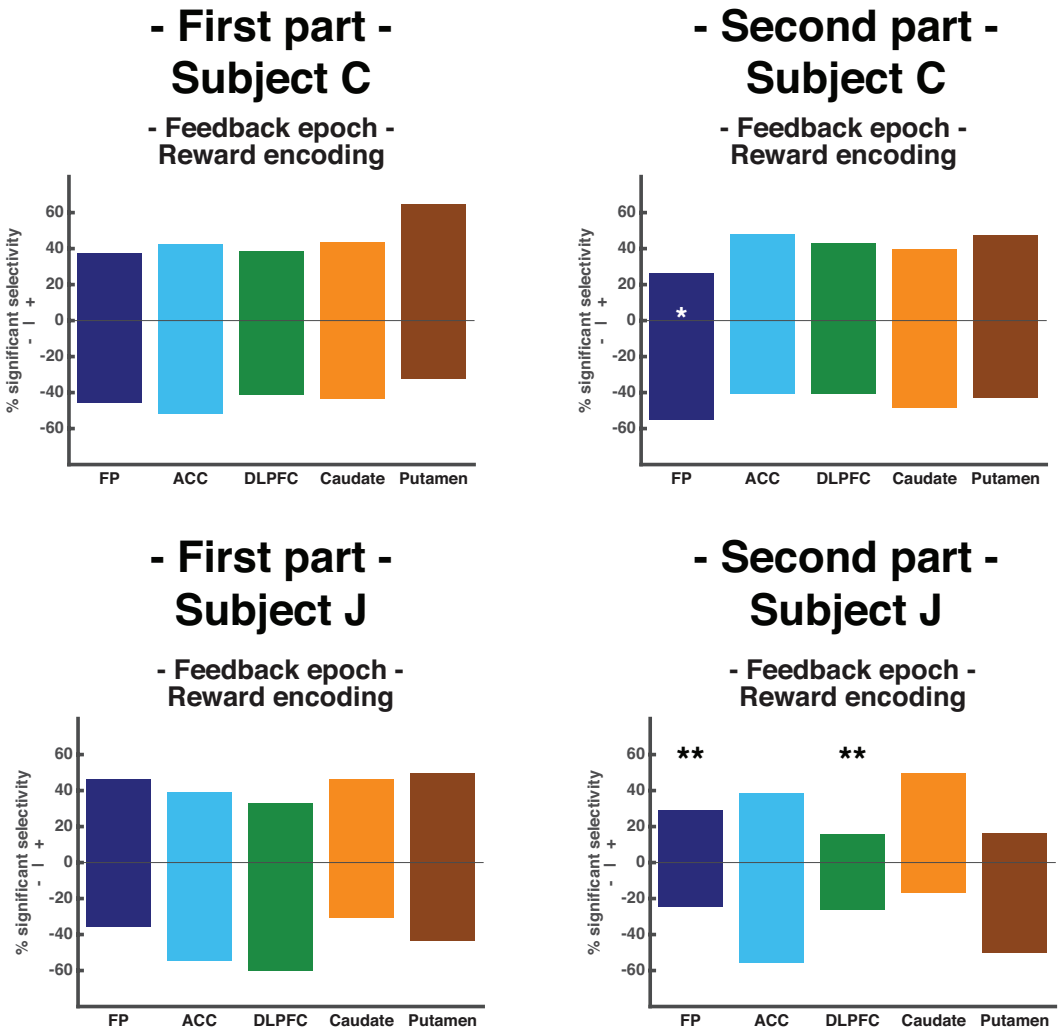


Fig. A.3 **Comparison between first and second parts of recordings for the encoding of reward at the single-neuron level.** Bar plot with the prevalence of neurons significantly encoding reward, based on the sign of the regression coefficient (+ or -). Double black asterisks for $\alpha = 0.01$ and single black asterisk for $\alpha = 0.05$ in chi-squared tests for differences between respective first and second-parts in the number of selective cells overall; double white asterisks, $p < 0.05$ for the proportion of neurons with positive or negative regression coefficients different from the chance 50%-50% split (binomial test).

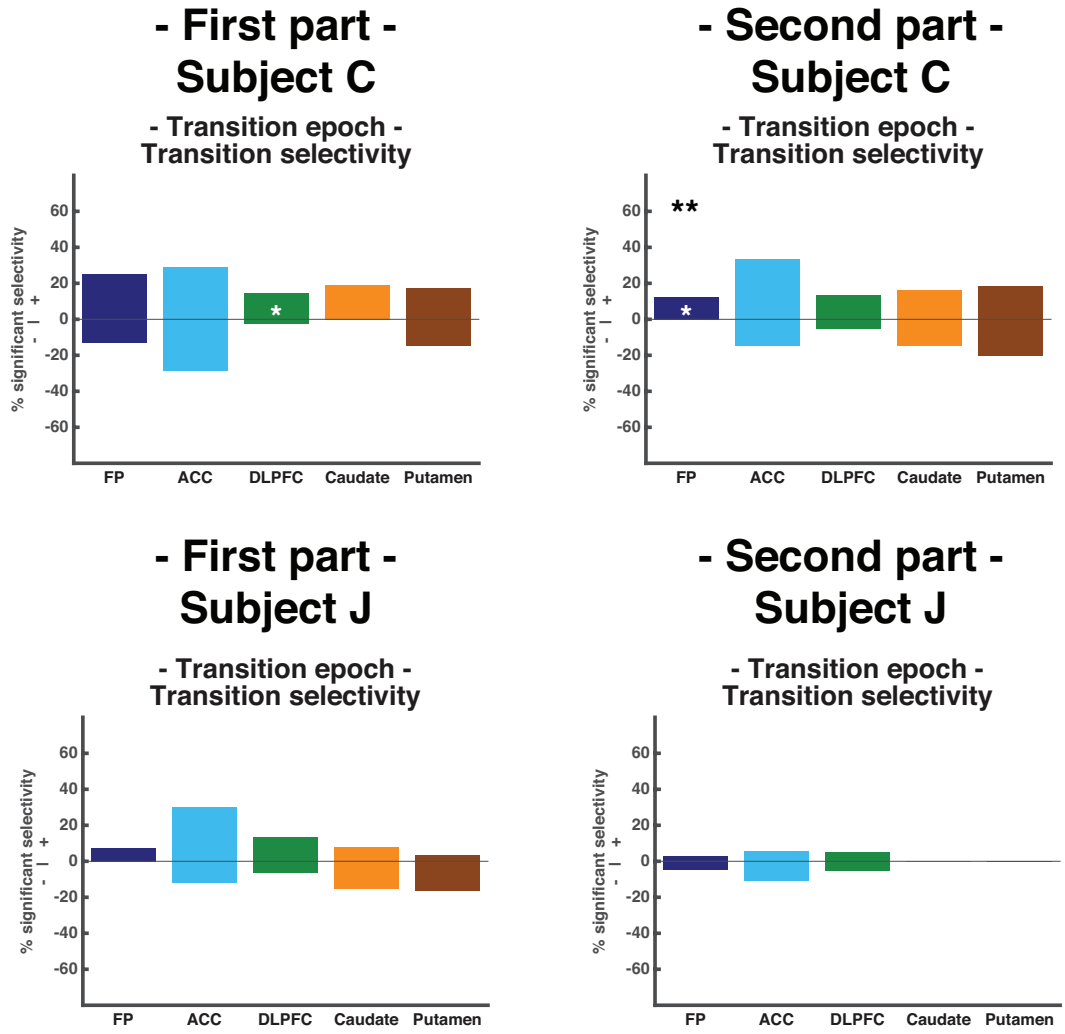


Fig. A.4 **Comparison between first and second parts of recordings for the encoding of state-transition information at the single-neuron level.** Bar plot with the prevalence of neurons significantly encoding transition type, based on the sign of the regression coefficient (+/− for increased firing rate if rare/common transition, respectively). Double black asterisks for $\alpha = 0.01$ and single black asterisk for $\alpha = 0.05$ in chi-squared tests for differences between respective first and second-parts in the number of selective cells overall; double white asterisks, $p < 0.05$ for the proportion of neurons with positive or negative regression coefficients different from the chance 50%-50% split (binomial test).

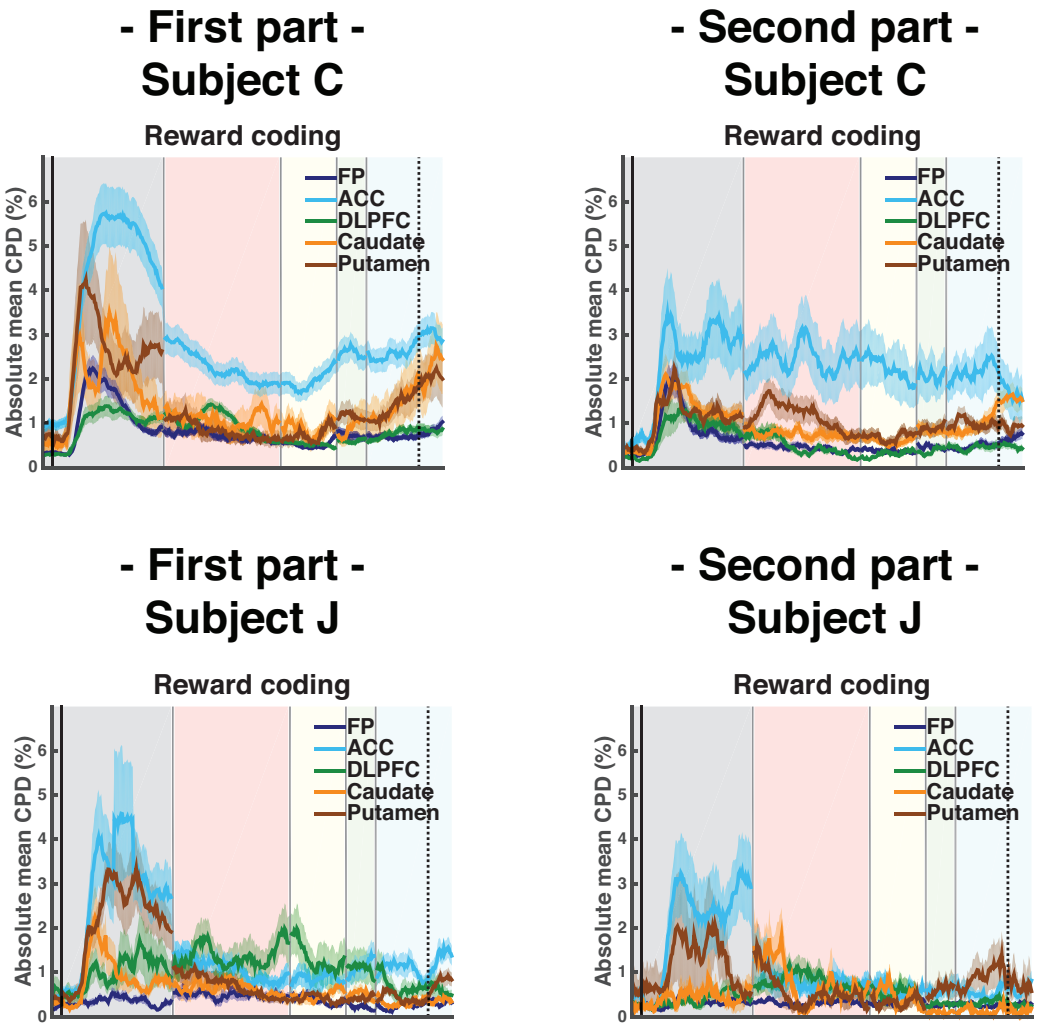


Fig. A.5 **Comparison between first and second parts of recordings for the population encoding of reward.** Time course of the reward coding at the population level, as determined by the absolute coefficient of partial determination (CPD) value, from feedback to choice1 epochs. First part (left-column) correspond to the first 18(C)/16(J) sessions out of a total of 30(C)/27(J). The 5% trimmed absolute mean (solid coloured line) and respective SEM (shading) across recorded neurons was used; solid vertical line corresponds to secondary reinforcer presentation; dotted vertical line represents the mean first-stage reaction time across subjects and sessions.

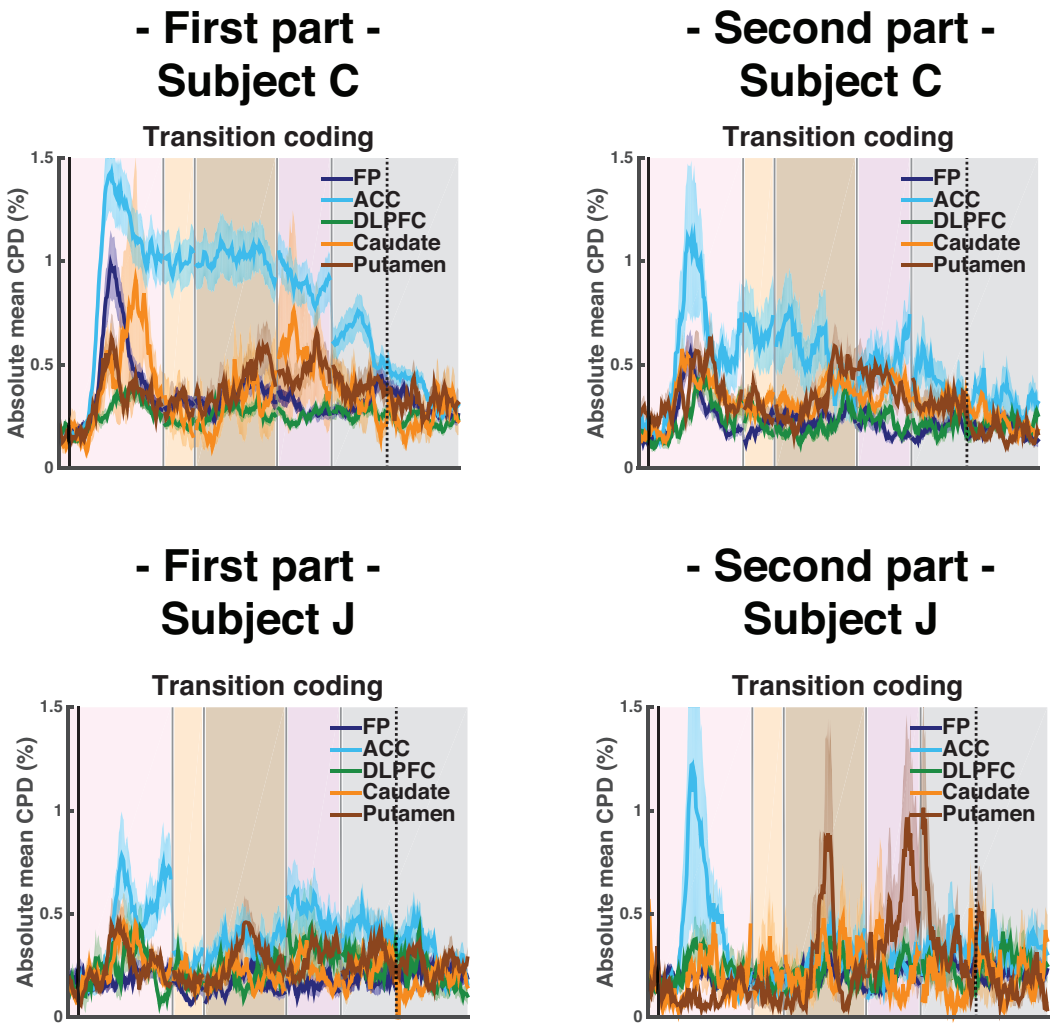


Fig. A.6 **Comparison between first and second parts of recordings for the population encoding of state-transition information.** Time course of the transition coding at the population level, as determined by the absolute coefficient of partial determination (CPD) value, from transition to feedback epochs. First part (left-column) correspond to the first 18(C)/16(J) sessions out of a total of 30(C)/27(J). The 5% trimmed absolute mean (solid coloured line) and respective SEM (shading) across recorded neurons was used; solid vertical line corresponds to the moment the background colour changes indicating the second-stage state, and hence the transition type; dotted vertical line represents the secondary reinforcer presentation.

Table A.1 **Comparison between first and second parts of recordings for the best fitting mixed-effects estimates from the best model.**

Model	Subject	α_1	α_2	β_1	β_2	κ_1	κ_2	λ	ω	L_1	L_2	L_3
Hybrid+												
Subject C												
	First part	0.77		4.77	2.57	0.06		-	0.86	0.26	-0.06	-0.09
	Second part	0.78		**4.27***	2.52	**0.05***		-	0.86	**0.25****	-0.06	-0.08
Subject J												
	First part	0.59		5.51	1.86	0.04	0.31	-	0.91	0.52	-0.11	-0.17
	Second part	0.56		**4.22****	1.84	0.04	0.31	-	0.91	**0.51****	-0.08	-0.15

Hybrid+ (was the best model) model included the $SARSA$ algorithm as model-free strategy and the $Forward_1$ model-based algorithm (see full text for details). First part corresponds to the first 18(C)/16(J) sessions out of a total of 30(C)/27(J). Values correspond to mean parameter estimates. ** for $\alpha = 0.01$ and * for $\alpha = 0.05$ in two-sample t-test with null-hypothesis that the parameters estimates across sessions between the corresponding first and second parts come from independent random samples from normal distributions with equal means and equal but unknown variances. Regarding the parameter nomenclature used (when placed in between parameters, the respective parameter estimate was shared between both first-stage and second-stage): learning rate for first-stage (α_1) and second-stage (α_2) choice; inverse temperature for first-stage (β_1) and second-stage (β_2) choice; perseveration for first-stage (κ_1) and second-stage (κ_2) choice; eligibility trace (λ); $L1_1$, $L2_2$ and L_3 are the reinforcement strength (or aversion) for high, medium and low outcome, respectively (see text for full details); ω is the model-based weight.