Okinawa Institute of Science and Technology

Internal Report

A Critical View on Model-free / Model-based Decision Learning

Chris Reinke

May 2014

Supervisor: Prof. Dr. Kenji Doya

Abstract

Decision making is a cognitive process to select an option (action) from a set of alternatives. Learning becomes important where the outcomes of decisions are unknown and need to be learned. Decision learning describes this process of learning to make decisions.

Reinforcement learning, a field in machine learning, provides a framework to describe the cognitive processes of decision learning. Two theories about decision learning have been developed in the past using the reinforcement learning framework. The model-free/model-based (MF/MB) theory proposes that the general procedure of decision learning is composed of two distinct processes. One using model-free mechanisms and the other using model-based mechanisms. The second theory, the actor-critic theory, proposes the existence of two components to explain model-free mechanisms. The critic learns the values for certain states and the actor learns which actions to select to reach states with high values. Current MF/MB models are critic-only models, i.e. actions are only selected based on values and not by an separate actor component. This is contrast to the actor-critic theory about model-free processes.

I developed an experimental design for a decision task with the goal to identify decision learning behavior that can not be modeled with critic-only methods, but which needs an actor-critic model to be explain. I collected preliminary data from human participants and showed that current MF/MB model can not explain some of their behavior. The results suggest that current MF/MB models need to be adapted, possibly by introducing an actor-critic procedure for their model-free component. Improved models can help model human decision learning better and to facilitate our understanding of its cognitive processes in the brain.

Contents

1 Introduction

Decision making is a cognitive process to select an option, e.g. an action, from a set of alternatives. Learning becomes important in environments where decisions have to be repeatedly made and the outcomes of decisions are unknown and need to be learned. Examples can be found in daily life such as: Which road should we use to drive to the supermarket? There could be two possibilities: 1) using the highway or 2) using local roads. The highway is usually faster but during certain time periods, e.g. rush hour, the local roads are better. If we recently moved to a new city we need to learn the different roads and their travel times. Learning is often performed in a trial and error way by using different roads at different time points. Decision learning describes this process of learning and decision making.

Reinforcement learning (Sutton & Barto, 1998), a field in machine learning, provides a value-based decision making framework to describe the cognitive processes of decision learning. The framework assigns the expected reward (for the road example the expected traveltime) as a value to actions. Actions are then chosen based on their values with a preference for higher valued actions. Two kinds of techniques use this approach. Model-free algorithms learn the values for each possible action with help of the temporal difference error, i.e. the difference between the observed reward and the expected reward. In contrast, model-based algorithms learn models of the environment, i.e. what is the next state after an action is performed. The model is then used to simulate possible outcomes of actions to calculate their value.

Two theories about the cognitive processes behind decision learning have been developed using the reinforcement learning framework. The model-free/model-based (MF/MB) theory looks at the general structure of decision learning and proposes that it is composed of two distinct processes (Daw et al., 2005). One using model-free mechanisms and the other using model-based mechanisms. The actor-critic theory proposes the existence of two components to explain model-free mechanisms (Barto, 1995). The critic learns the values of states and the actor learns which actions to select to reach states with high values. The model-based mechanisms are less clear (Daw, 2012) but it is proposed that they could also be explained by an actor-critic framework (Bornstein & Daw, 2011).

Current MF/MB models are critic only models, i.e. actions are only selected based on values. This is contrast to the actor-critic theory about model-free processes in the brain. I developed a new experimental design with the goal to identify decision learning behavior that can not be modeled with critic only methods, but which needs actor-critic models to be accounted for. The next section introduces the background in reinforcement learning and the cognitive models that are based on it. Afterward, the experiment and its preliminary results are described.

2 Literature Review

The literature review introduces the basic concepts and techniques in reinforcement learning which form the basis for current theories in the field of decision learning. This is followed by a discussion of the MF/MB and the actor-critic theory with a review of experimental studies that support their claims. Moreover, cognitive models based on the theories are reviewed.

2.1 Reinforcement Learning

Reinforcement learning is a field in machine learning studying how an agent should learn to choose actions, i.e. to make decisions, in tasks with multiple decision steps to acquire a maximum sum of reward. Reinforcement learning provides the mathematical framework of Markov decision processes (MDP's) to describe such learning situations and introduces several algorithms as solutions. Value-based and non value-based algorithms exist where value-based are further distinguished in critic-only and actor-critic algorithms (Table 1). All algorithms can be further distinguished in model-free and model-based methods. This review focuses on model-free and model-based critic-only and model-free actor-critic algorithms because they form the basis of the theories to describe cognitive mechanisms of decision learning. First the definition of an MDP is given followed by an introduction in value-based model-free and model-based algorithms.

	value-based		**non value-based**
	critic-only	**actor-critic**	**actor-only**
model-free	TD-Algorithms such as Q-Learning, SARSA (Sutton & Barto, 1998)	(Barto, 1995)	REINFORCE (Williams, 1992) Natural Gradient (Kakade, 2001)
model-based	Dynamic Programming (Sutton & Barto, 1998)	Value Gradient Based Policy Continous RL (Doya, 2000)	PILCO (Deisenroth & Rasmussen, 2011)

Table 1: Classification of reinforcement learning methods with example algorithms.

Markov Decision Process

Reinforcement learning uses the framework of MDP's to describe decision problems (Bellman, 1957; Sutton & Barto, 1998). A MDP consists of four elements:

$$\text{MDP} = (S, A, Pr(s_{t+1}|s_t, a_t), R(s_t, a_t), \gamma), \tag{1}$$

a finite number of states S and actions A. Performing an action a_t while in state s_t results in a transition to the next state s_{t+1}. Each transition has a probability defined by the transition probability $Pr(s_{t+1}|s_t, a_t)$. The reward function $r_{t+1} = R(s_t, a_t) \in \mathbb{R}$ defines the rewards that can be gained by a transition. The goal is to find an action selection strategy, called a policy $\pi(s, a) = Pr(a|s)$ which is usually represented as a probability distribution over actions for each state, that maximizes the discounted reward sum given a discount factor $\gamma \in [0\ 1]$:

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \tag{2}$$

Value-Based Algorithms

Reinforcement learning algorithms solve MDP's. Value-based algorithms use a value function which represents, for different states or actions, how much reward can be expected in the future by using them if afterwards a certain policy is followed. These algorithms form the basis for

many cognitive theories about decision learning. Non value-based algorithms such as policy search methods (Deisenroth et al., 2011) try to optimize directly the policy to obtain a high reward. The review focuses on value-based methods. The next part will introduce the value function which forms the basis for these methods.

The concept of a value function was introduced to solve MDP problems with a dynamic programming approach by dividing the problem in smaller sub problems and then by solving them individually to solve the overall problem (Bellman, 1957; Sutton & Barto, 1998). The function can be formulated as a state-value function $V(s)$ or an action-value function $Q(s, a)$ (also called Q-function). The value of a state is the expected discounted reward sum by starting at state s and using a certain policy π in its successor states:

$$V^{\pi}(s) \;=\; E_{\pi}\{R_t|s_t = s\} \;=\; E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t = s\right\} \tag{3}$$

The value of a state-action pair (Q-value) is the expected discounted reward sum by starting from s using action a and then following a certain policy:

$$Q^{\pi}(s, a) \;=\; E_{\pi}\{R_t|s_t = s, a_t = a\} \;=\; E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t = s, a_t = a\right\} \tag{4}$$

The Bellman equation reformulates these into recursive formulations were the value of a state $V(s_t)$ or an action $Q(s_t, a_t)$ depends on its successor's value $V(s_{t+1})$, $Q(s_{t+1}, a_{t+1})$:

$$V^{\pi}(s_t) \;=\; \sum_{a_t \in A}\left[Pr(a_t|s_t)R(s_t, a_t) + \gamma \sum_{s_{t+1}\in S} Pr(s_{t+1}|s_t, a_t)V(s_{t+1})\right] \tag{5}$$

$$Q^{\pi}(s_t, a_t) \;=\; R(s_t, a_t) + \gamma \sum_{s_{t+1}\in S}\left[Pr(s_{t+1}|s_t, a_t) \sum_{a_{t+1}\in A}[Pr(a_{t+1}|s_{t+1})Q(s_{t+1}, a_{t+1})]\right] \tag{6}$$

Bellman proofed that by finding for each recursive sub problem, i.e. for each state, the action that results in the optimal value (V^*, Q^*) the overall optimal value function for the problem is found:

$$V^*(s_t) = \max_{a_t}\left[R(s_t, a_t) + \gamma \sum_{s_{t+1}\in S} Pr(s_{t+1}|s_t, a_t)V^*(s_{t+1})\right] \tag{7}$$

$$Q^*(s_t, a_t) = R(s_t, a_t) + \gamma \sum_{s_{t+1}\in S} Pr(s_{t+1}|s_t, a_t) \max_{a_{t+1}}[Q^*(s_{t+1}, a_{t+1})] \tag{8}$$

Importantly is that by learning the optimal value function the optimal policy is also acquired. In the case of the state-value function by using in each state the action that leads to the successor state with the highest value. In the case of the Q-function by using the action with the highest Q-value.

The value function is learned by interacting with the environment and by trying different strategies, i.e. using different actions to explore their outcomes. To allow such an exploration not always the action with the current highest value should be chosen during learning. Therefore, a probability distribution $Pr(a|s)$ over actions is used as policy.

Value-based algorithms are further distinguished in critic-only and actor-critic algorithms (Table 1). Both use the value function but differ in their way to define the policy $\pi = Pr(a|s)$. Critic-only methods directly use the learned value function to decide which action to perform in a state. Most critic-only algorithms use a Q-function and the soft max action selection to define the probability distribution:

$$Pr(a|s) \;=\; \frac{\exp(\beta \cdot Q(s,a))}{\sum_{a'} \exp(\beta \cdot Q(s,a'))} \tag{9}$$

The soft max action selection prefers actions with higher Q-values where the strength of the preference can be controlled by the inverse temperature parameter β.

In contrast, actor-critic algorithms do not directly use the value function for the action selection. Instead they have an extra parameterized policy which is learned with help of the value function. Before actor-critic algorithms are further introduced critic-only algorithms and the distinction between model-based and model-free algorithms are discussed in more detail.

Model-Based and Model-Free Critic-Only Methods

Reinforcement learning algorithms are also distinguished in model-based and model-free algorithms (Table 1). This part describes their critic-only variants which are used to describe cognitive processes of decision learning. The algorithms differ in their way to learn the value function.

Model-based methods use an environment model, i.e. the transition probabilities $Pr(s_{t+1}|s_t, a_t)$ and the reward function $R(s_t, a_t)$. Dynamic programming algorithms such as Value Iteration were the first model based algorithms (Bellman, 1957). Knowledge of the environment model allows them to directly solve the Bellman equation by iterating several times over all states or state-action pairs to improve an initial value function over time. Current cognitive models often use a Forward algorithm (Gläscher et al., 2010; Daw et al., 2011; Wunderlich et al., 2012; Lee et al., 2014). Forward algorithms simulate all possible actions and outcomes from a start state to a final state in a forward tree search manner based on the environment model. The value for each state or state-action pair is the best value over all possibilities to reach the final state.

For most problems the transition probabilities $Pr(s_{t+1}|s_t, a_t)$ are not known and need to be learned. The probabilities are learned from observations, i.e. the agent interacts with the environment and observes the transition from s_t to s_{t+1} after an action is executed. The following rules can be for example used to update the probabilities:

$$\begin{aligned} Pr(s'|s,a) &\leftarrow Pr(s'|s,a) + \eta \cdot (1 - Pr(s'|s,a)) \\ Pr(s''|s,a) &\leftarrow Pr(s''|s,a) \cdot (1 - \eta) \end{aligned} \tag{10}$$

were s' is the observed state transition, s'' are all other possible transitions and η is a learning rate parameter. In summary, learning in model-based algorithms is associated with the learning of its environment model by observations. Based on the learned model the value function is computed for example with a dynamic programming or a tree search approach.

Model-free algorithms on the other side such as Q-learning or SARSA use no environment model and have therefore to learn the value function directly from observations. Algorithms are usually based on the temporal difference (TD) error and use a Q-function. Such TD algorithms start

with an initial function Q_0. The agent interacts with the world by performing an action a_t in state s_t. The resulting state s_{t+1} and the reward r_{t+1} are observed. Based on this information the TD error δ_t is computed:

$$\delta_t \;=\; r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \tag{11}$$

The error is the difference between the current reward prediction $Q(s_t, a_t)$ and the reward observation r_{t+1} plus the discounted Q-value of the next state $Q(s_{t+1}, a_{t+1})$. The action a_{t+1} depends on the TD algorithm, e.g. Q-learning is using the action with the highest value. Following Equation 6 if the true Q-function would have been learned then $Q(s_t, a_t) = E[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1})]$. Thus δ_t is the error between the current expected reward prediction and the true Q-value. It is used to update the prediction with:

$$Q(s_t, a_t) \;\leftarrow\; Q(s_t, a_t) + \alpha\delta_t \tag{12}$$

The learning rate α defines the update strength. With an appropriate setting of α the Q-function will converge over time to the true function. In combination with an action selection algorithm such as soft max that prefers actions with higher values the function will converge to the optimal function Q^*.

Actor-Critic Methods

Another class of algorithms used for cognitive models are model-free actor-critic methods (Barto, 1995). Their advantage to TD algorithms is instead of a Q-function the use of a state-value function $V(s)$ which is easier to represent and to learn. Actor-critic algorithms consist of a critic and an actor component. The critic is the value function and the actor a parameterized policy learned with help of the value function. Therefore, in contrast to critic-only methods the action selection depends not directly on the value function.

The basic actor-critic approach which underlies different cognitive models is illustrated on the algorithm given by Sutton & Barto (1998). The critic uses a the TD approach to learn the state-value function:

$$\delta_t \;=\; r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{13}$$

$$V(s_t) \;\leftarrow\; V(s_t) + \alpha\delta_t \tag{14}$$

The actor is parameterized by $\theta_{s,a}$ holding a preference value for each state-action pair. The actors policy is defined in similar way to the soft max method:

$$Pr_\theta(a|s) = \frac{\exp(\theta_{s,a})}{\sum_{a'} \exp(\theta_{s,a'})} \tag{15}$$

The higher a preference value of an action the higher is its probability to get selected. The preferences $\theta_{s,a}$ are learned by the TD error δ_t which is computed by the critic after an action was performed. If the action resulted in a positive TD error, i.e. the action resulted in a higher reward prediction than expected, the probability of performing the action should be increased. If instead the last action produced a negative TD error, i.e. the action resulted in a lower reward prediction than expected, the probability of performing the action again should be decreased. For the given actor-critic algorithm this update mechanism is implemented by:

$$\theta_{s,a} \;\leftarrow\; \theta_{s,a} + \kappa\delta_t \tag{16}$$

were κ is the learning rate of the actor. The difference to critic-only methods is that the policy is defined by the parameters θ and not by the value function.

Conclusion

The discussed reinforcement learning approaches differ in regard to their computational properties. Model-free critic-only methods use a Q-function because they need to store for each possible action a value to compare them and to make a decision between them. Model-based algorithms have the advantage to use only a state function which is easier to represent and to learn. Their environment model allows them to anticipate the outcome of an action, i.e. with which probability it transitions in another state. Therefore, actions can be compared based on the value of the states were they lead to. Actor-critic methods allow to combine a model free critic with a state-value function. For all important is the concept of the value function and for the model free methods the concept of the TD error.

The introduced algorithms form the computational theory behind most cognitive theories about decision learning. The next two sections introduce two important theories and their evidence. The MF/MB theory is first discussed followed by the actor-critic theory.

2.2 Model-Free / Model-Based Mechanisms in the Brain

The MF/MB theory states that decision learning can be explained by model-free and model-based computations (Doya, 1999). Daw et al. (2005) proposed that both processes are distinct from each other and explain the difference between goal-directed and habitual behavior. Goal-oriented behavior is performed to achieve a certain goal and is therefore aware of its outcomes. Habitual behavior represents a state-response mapping performed in certain situations without awareness of its outcome. Before the theory of separated model-free and model-based process in the brain is further discussed general evidence for the existence of both processes is reviewed. The last part of the section lists cognitive models based on the theory.

Evidence for Model-Free Mechanisms in the Brain

First, studies are reviewed supporting the hypothesis that model-free value-based reinforcement learning mechanisms explain the brain processes of learning behavior. The first line of evidence shows that the activity of dopaminergic neurons in the basal ganglia, a system of subcortical nuclei in the brain important for learning, is correlated to TD error computations. The second line of evidence shows that striatal neurons in the basal ganglia seem to encode a value function. TD error computations and the representation of a value function are the important properties of model-free TD mechanisms and provide strong evidence that the brain is using such mechanisms for decision learning.

The evidence for TD error computations in the brain comes from findings that activity of dopamine neurons in the substantia nigra (SN) and the ventral tegmental area (VTA) follow TD error characteristics (Schultz et al., 1997). The neurons increase activity in response to unexpected rewards and show a reduction in activity at time points when rewards should be expected but it is not given. Furtermore, fMRI studies show that the activity of the projection areas of these neurons in the striatum and the frontal cortex is also correlated with reward prediction error computations (Pagnoni et al., 2002; Pessiglione et al., 2006). The neurons release the neurotransmitter dopamine according to their activation which is therefore regarded as a TD error learning signal reinforceing actions or states previous to its release. Experiments which stimulate dopamine neurons with intracranial methods such as with optogenetics support

this hypothesis. Rats prefer a lever which stimulates dopamine neurons over one that does not (Adamantidis et al., 2011) and they prefer locations were dopamine responses are elicited (Tsai et al., 2009). Moreover, human fMRI studies show that reward prediction error related activity can be measured in participants that learn a task in contrast to non learning participants (Schönberg et al., 2007). That dopamine is important for learning is also supported by findings that dopamine modulates plasticity in the striatum (Reynolds & Wickens, 2002). In summary, the findings suggest that dopamine neurons in the SN and VTA encode the TD error which gets projected into the striatum and the prefrontal cortex via the neurotransmitter dopamine.

The second aspect of model-free value-based algorithms is their use of a value function. Correlations between the firing pattern of striatal neurons and values of TD models have been identified during electrophysiological experiments in monkeys and rats (Samejima et al., 2005; Ito & Doya, 2009). The animals performed a decision learning task and their choice behavior was modeled with TD algorithms to compute the value function which the animals seem to use. Electrical recordings of neurons showed a correlation between their activity and the values computed by the TD algorithms providing evidence that the animals may encode a value function in the striatum. This is further supported by human fMRI experiments which also found a correlation of activity in the striatum and the prefrontal cortex with values based on TD computations (Kable & Glimcher, 2007; Gläscher et al., 2010; Daw et al., 2011; Wunderlich et al., 2012; Lee et al., 2014). In conclusion, the evidence for a neural representation of values and the TD error supports the theory that the brain is using a model-free value-based mechanisms.

Evidence for Model-Based Mechanisms in the Brain

Research by Tolman (1948) showed that behavior can not exclusively explained by model-free mechanisms but that model-based mechanisms are also important. Tolman studied how rats learn to find the optimal path in mazes to reach a goal box with food. Before the experiment some rats were allowed to explore the maze. During this exploration phase the rats did not receive rewards and therefore could not learn values for actions based on TD computations. Nonetheless, rats that were allowed to explore the maze before the experiment learned more quickly to reach the goal box. Tolman concluded that the rats must have learned a cognitive map of the maze which helped them afterwards during the learning phase. Such learning can not be explained by model-free mechanisms but by model-based mechanisms which use the cognitive map as an environment model.

Evidence for Distinct Model-Free / Model-Based Mechanisms

The reviewed evidence suggests the existence of model-free and model-based mechanisms in the brain, but leaves open how they relate to each other. The MF/MB theory proposes that both are separate processes (Daw et al., 2005). The theory is based on the distinction between goal-directed and habitual behavior (Dickinson, 1985) which are shown to be distinct processes in the brain (Yin et al., 2004, 2005). The learning of goal-directed behavior is associated with model-based processes and habitual behavior with model-free processes. Therefore, model-based and model-free processes should also be separate processes. Before the MF/MB theory is discussed in more detail the distinction between goal-directed and habitual behavior is introduced.

Goal-directed and habitual behavior can be identified with help of devaluation experiments (Dickinson, 1985) which consist of three phases: I) A rat learns that pressing a lever results in a

food reward. II) The value of the reward gets devaluated before the next phase for example by feeding the rat. III) The rat is given again the chance to press the lever but without receiving any rewards. In the last phase it is measured if a difference in the rate of lever pressing exists between devaluated and non-devaluated sessions. Depending on the number of training trials in the first phase the animal learns either goal-directed or habitual behavior.

In the case of goal-directed behavior the rat anticipates the outcome of the action that pressing the lever will give food. Therefore, the rat will not press the lever in the third phase if the food is devaluated because it is not hungry and does not want food. In the case of habitual behavior the rat forms a habit to press the lever and is not anticipating its outcome. Thus, the rat presses the lever regardless of whether or not it is hungry. Goal-directed behavior is seen in moderately trained rats, whereas habitual behavior is seen in over trained rats.

Lesion studies have shown that both learning behaviors can separately exist from each other. Animals with lesions in the dorsomedial striatum show disrupted goal-directed behavior but they are able to learn habitual behavior (Yin et al., 2005). Vice versa, lesions in the dorsolateral striatum disrupt learning of habitual behavior but not goal-directed behavior (Yin et al., 2004).

Daw et al. (2005) suggest the MF/MB theory to explain the different behaviors in devaluation experiments. The theory states that a separate model-free and a model-based component exist in the brain explaining habitual and goal-oriented behavior respectively. A cognitive model was developed in which each component computes an independent Q-value. The model-based component learns the transition probabilities of the environment, i.e. what are the outcomes of actions. Based on the model Q-values for actions are computed with a Forward tree search method. In the first phase of devaluation experiments the model-based component learns the environment model, i.e. pressing the lever results in food. In the third phase the outcome of a lever press can be anticipated with help of the model. Therefore it can anticipate that food would result and not press the lever if food was devaluated resulting in goal-oriented behavior.

In contrast, the model-free component explains habitual behavior. Q-values of actions are learned on the history of perceived rewards with help of a TD mechanism. In the first phase of the devaluation experiment a positive Q-value for pressing the lever is learned. In the third phase the lever gets pressed because the action has a high Q-value. That pressing the lever would result in devaluated food is not anticipated and habitual behavior is obtained.

Which of the two Q-values (model-free or model-based) is used to select an action depends on the certainty of the components, i.e. how accurate their reward prediction is. After moderate training the certainty of the model-based component is higher and the resulting behavior is goal-oriented. With more training the certainty of the model-free component becomes stronger resulting in habitual behavior.

Cognitive Models of the Model-Free / Model-Based Mechanisms

Further cognitive models based on the MF/MB theory were developed to explore its properties and to find the neural substrates responsible for model-free and model-based processes (Gläscher et al., 2010; Daw et al., 2011; Wunderlich et al., 2012; Lee et al., 2014). This section introduces the models and their findings.

All studies used fMRI experiments and fitted the parameters of their MF/MB model such as learning rates to reproduce the behavior of participants. The Q-values of the model-based

and model-free component were then correlated to the brain activity of participants during the experiment. Distinct areas correlating with model-based values (prefrontal cortex areas, caudate) and model-free values (ventral striatum, putamen) have been identified supporting the theory of two distinct learning systems. Concerning the model-based processes the studies by Gläscher et al. (2010) and Lee et al. (2014) used the concept of a state prediction error (SPE) to learn the environment model in the model-based component. The SPE is the prediction error of the environment model in predicting the outcome of an action. The error is used to update the model in a similar way to the TD error. Correlated brain activity to the SPE was found in the lateral prefrontal cortex (latPFC) and the intraparietal sulcus (pIPS) supporting its existence.

A major difference between the cognitive models is their way to arbitrate between the Q-functions of the model-based and the model-free component. All models use critic-only mechanisms with the soft max action selection (Equation 9). Therefore an arbitration between the two Q-functions in necessary. The model of Daw et al. (2005) uses a winner takes all method by selecting the Q-value of the system with the higher certainty. Another study found that decision behavior seems not to be purely controlled by either the model-based or model-free component but it is more integrative and a combination of both (Daw et al., 2011). This evidence was accounted for by introducing a weighted linear sum of the model-based and model-free Q-value to compute the resulting value:

$$Q_t(s,a) \;=\; w_t Q_{MB}(s,a) \;+\; (1-w_t) Q_{MF}(s,a) \quad \mid w_t \in [0\ 1] \tag{17}$$

The first approach (Gläscher et al., 2010) uses an exponentially decaying weight: $w_t = Ie^{-kt}$. As a result the Q-values of the model-based component dominate in the beginning of learning while the influence of the model-free component becomes stronger over time until it takes over. This follows the findings that goal-oriented behavior is more dominant after moderate training followed by habitual behavior after over training (Dickinson, 1985).

The model by Lee et al. (2014) improves the idea of a weighted linear sum by introducing the concept of reliability similar to the concept of uncertainty (Daw et al., 2005). A system is reliable when its predictions about the future are correct. Q-values of the more reliable system get a higher weight with a bias towards to model-free component. Therefore model-free behavior will be dominant if over training occurs and the predictions of both systems have a high reliability. FMRI results of Lee et al. (2014) suggest that the anterior cingulate cortex compares reliability signals of the model-based and model-free systems. The inferior lateral prefrontal cortex and the frontopolar cortex arbitrate between the systems by controlling the inhibition of the input from the model-free system to the ventro medial prefrontal cortex and the orbitofrontal cortex were the values seem to get integrated.

In conclusion, the MF/MB theory provides a description of the general structure of the cognitive processes behind decision learning. Its success is based on its ability to explain different aspects of behavior, most importantly the distinction of goal-directed and habitual behavior. Its assumption of separated model-based and model-free mechanisms is supported by lesion studies and fMRI experiments.

2.3 Actor-Critic Mechanisms in the Brain

A second key theory influenced by reinforcement learning is that model-free learning processes are implemented by an actor-critic framework within the basal ganglia (Barto, 1995). This

section introduces the neurobiological evidence supporting this theory and gives an overview of cognitive models that use the actor-critic framework.

Evidence for Actor-Critic Mechanisms in the Brain

The actor-critic theory suggests that an actor-critic system with separate areas for the critic and the actor in the basal ganglia is responsible for model-free learning processes. The general existence of such model-free TD mechanisms has been discussed in Section 2.2. This section concentrates on the evidence for the hypothesis that these computations are implemented by a separate critic and actor component. The ventral striatum (VS) is associated with the role of the critic and the dorsal striatum (DS) with the role of the actor. First an experiment is reviewed that dissociates both areas based on their activity providing evidence that both areas have separate functions. This is followed by a review of specific neurobiological properties of both areas which support their associated computational role in an actor-critic framework.

The VS and DS show distinct activities in human fMRI experiments (O'Doherty et al., 2004). The experiments consist of two task settings. The first is a classical conditioning task in which the association of a stimulus and a reward that it predicts is learned. Thus, this would involve only the critic component to learn the value function, i.e. the state of seeing the cue to be associated with the reward value. The second setting consists of an instrumental conditioning task were the values of actions need to be learned to make the correct decision between them. This would involve to learn the value function for the actions and to learn the policy for the action execution. The experimental results show that in both tasks reward prediction related activity was correlated with the VS pointing to its role as a critic. Whereas only in the instrumental conditioning setting where the learning of a policy takes place also activation in the DS could be correlated pointing to its role as an actor.

Other evidence shows that the specific neurobiological properties of the the VS and DS are in accordance with their role as a critic and an actor. First the VS and its role as a critic is discussed. If the VS has the role of a critic it needs to represent and to learn the value function. One aspect for representing a value function is that information about the current environmental state is needed. The VS fulfills this criteria because it receives input from different areas such as the basolateral amygdala, the hippocampus and prefrontal areas (Voorn et al., 2004) which include state and value information. Another property concerns the connections a critic must have in relation to the TD error. To learn the value function it needs the TD error information (Equation 14). Moreover, to compute the TD error the value function is itself needed (Equation 13). The connections of the VS with the two dopaminergic nuclei which represent the TD error are in line with these properties. The VS receives input from the VTA, one of the dopaminergic nuclei, and has therefore the information to learn the value function. Important is that it projects itself back to both dopaminergic nuclei, the SN and the VTA (Joel & Weiner, 2000), which can therefore compute the TD error and project it to the VS and other areas. Electrophysiological experiments provide more evidence for the role of the VS as a critic by showing that its neural activity is correlated to reward expectations which is the defining property of a value function (Schultz et al., 2000; Wan & Peoples, 2006). The VS has the important properties that a critic has to fulfill in an actor-critic framework. It receives as input state information and the TD error, it represents the value function and it projects value information back to the areas which compute the TD error.

The DS is associated with the role of an actor. The actor learns and defines the policy which

controls actions and movements (Equation 15). This is in line with the findings that the DS is part of the motor cortical loop in the basal ganglia which is involved in movement control (Voorn et al., 2004). To learn the policy the TD error is needed (Equation 16). This is fulfilled for the DS because it receives connections from the SN, one of the dopaminergic nuclei (Joel & Weiner, 2000). Moreover, electrophysiological studies in monkeys and rats show that DS neurons encode action initiation that is learned over time (Schultz & Romo, 1988; Jog et al., 1999). This is in line with the properties of an actor to represent a policy that is learned over time with help of the TD error.

In conclusion, different lines of research link the VS to the role of a critic component. Whereas, the DS fulfills the properties related to the function of an actor. These findings support the theory that model-free learning processes could be organized in an actor-critic framework in the striatum. Please note that the review concentrated on the role of the striatum in an actor-critic frame work. Nonetheless, other brain areas are also associated with the role of a critic such as the orbitofrontal cortex (Schultz et al., 2000) or the amygdala (Belova et al., 2008).

Cognitive Models of Actor-Critic Mechanisms

Several cognitive models with the actor-critic framework have been proposed. Some are based on detailed neurobiological findings about involved brain areas to show their function in an actor-critic framework. Others use the framework to explain general learning behavior.

Barto (1995) suggested the use of the actor-critic framework to explain learning behavior and identified the basal ganglia as responsible for such computations. Several models followed this suggestion and linked specific neural substructures in the basal ganglia to certain computations in an actor-critic framework (Houk et al., 1995; Suri & Schultz, 1998, 1999; Brown et al., 1999; Contreras-Vidal & Schultz, 1999; Suri et al., 2001). They assume the critic is represented by striosome neurons and the actor by matrisome neurons which can be found in the whole striatum. This view was challenged by the findings that ventral and dorsal areas in the striatum have different connections to the dopaminergic nuclei (Joel & Weiner, 2000) leading to the current theory were the critic is in the VS and the actor is in the DS (Joel et al., 2002). A further model about the actor-critic structure in the basal ganglia by Joseph et al. (2010) concentrates on the role of different basal ganglia nuclei and their involvement in exploratory behavior.

Besides the basal ganglia, the actor-critic framework is also proposed to account for mechanisms in the anterior cingulate cortex (ACC). Two models have been developed based on neurobiological findings of the ACC and related learning behavior. Both concentrate on the role of a critic in the ACC which learns either to predict outcomes of actions (Alexander & Brown, 2011) or values of actions (Silvetti et al., 2011). The actor is hypothesized to be in the most rostral and caudal areas of the ACC whereas the critic is associated with the mid-third areas (Silvetti et al., 2013).

Other cognitive models use the actor-critic framework to explain general learning phenomena without a strong focus on its neurobiological basis. Nakahara et al. (1998) modeled sequence learning behavior. Drug addiction mechanisms are explained with help of the actor-critic framework by Takahashi et al. (2008). Maia (2010) explains different behavioral aspects of avoidance conditioning and Barnes et al. (2014) use the actor-critic architecture to model learning and attention effects during categorization learning.

In summary, the actor-critic theory suggests that model-free decision learning processes are

implemented by a critic and an actor component. The critic is associated with the VS and other areas such as the orbitofrontal cortex and the amygdala. They learn and represent a value function. Whereas the actor which is mainly associated with the DS learns and represents a policy. The theory is supported by neurobiological evidence and by its ability to model different reward related learning behavior.

3 Experiment

The experiment is designed to test the critic-only design of current MF/MB models. Preliminary data was collected and one of the recent MF/MB cognitive models was analyzed based on its ability to reproduce the behavioral data. The next section describes the experimental design followed by the behavioral results and the model analysis.

Recent MF/MB models use critic-only mechanisms where Q-values are learned and the soft max procedure is used for action selection (Gläscher et al., 2010; Daw et al., 2011; Wunderlich et al., 2012; Lee et al., 2014). Actions will therefore exactly follow the learned Q-values. The goal of the experiment was to test if there could exist conditions under which the action selection does not follow learned values. In such a scenario a critic-only approach would not be able to explain the divergence between Q-values and behavior. Instead this could provide evidence for an actor-critic framework responsible for the decision learning behavior. In actor-critic frameworks the policy is separately learned from the values and could therefore diverge from the values.

The experiment was a computer task and consisted of three phases (Figure 1). In Phase 1 the participants were overtrained in learning the optimal decisions for four different states (180 trials for each state). States were distinguishable by the background color of the screen. In each state participants had to decide between two options (left or right). Each option had a certain reward probability to win a point which was given by a pseudorandom process. The probabilities for the states were (0.75; 0.5), (0.75; 0.5), (0.5; 0.25) and (0.25; 0).

In Phase 2 the reward probability of the previous optimal option o_1 for state s_1 was reduced from 0.75 to 0.25 becoming the non-optimal option. The probabilities for the other states remained the same. Participants had only a short training of 10 trials for each state. It was predicted that for state s_1 the participants would continue to choose the previous optimal option due to short relearning period. If the prediction holds then in the case of a critic-only framework the Q-value for the changed option o_1 should adapt only slowly and should still be higher than for the other option o_2 with a reward probability of 0.5. This would explain the non-changing behavior of the participants. In contrast, if the Q-value for the changed option adapts faster and goes below the Q-value of option o_2 then the non-changing behavior could not be explained by a critic-only approach.

Phase 3 was intended to test both cases by measuring the Q-value of the changed option o_1. The Q-value was inferred by measuring the preference between the four states. In each trial participants were given the choice between two of the states with 20 trials for each combination. The decision was presented as a choice between two options (left or right) which where visualized in the color of their associated state. After participants selected a color they went in the associated state and could win a point by choosing between its two options as done in Phase 1 and 2. It was assumed that the preference between the four states should follow the reward rate observed in them and therefore follow the Q-values of the options that are mostly used in them. Thus, if the optimal choices are learned state s_2 (0.75; 0.5) should for example be

preferred over state s_3 (0.5; 0.25) because its optimal option has a higher reward probability. Additionally, the value of state s_3 (0.5; 0.25) should be similar to the Q-value for the unchanged option o_2 in s_1 because they have the same reward probability. Based on these assumptions and the prediction that participants learn the optimal options for states s_2, s_3, s_4 and if they keep their preference for the changed option o_1 in s_1 the Q-value of o_1 should be able to be inferred from the state preferences. If state s_1 is preferred over s_3 and s_4 ($s_1 > s_3/s_4$) the Q-value of option o_1 should be higher than the optimal value from s_3 which has a 0.5 reward probability. If instead s_3 and s_2 are preferred over s_1 ($s_1 < s_3/s_2$) the value for o_1 should be lower than for a reward probability of 0.5. In the first case a critic-only framework could explain the behavior of keeping o_1 as the preferred option in state s_1 because it would have a higher Q-value than the second option o_2. In the second case a critic-only framework can not explain why option o_1 would still be preferred.

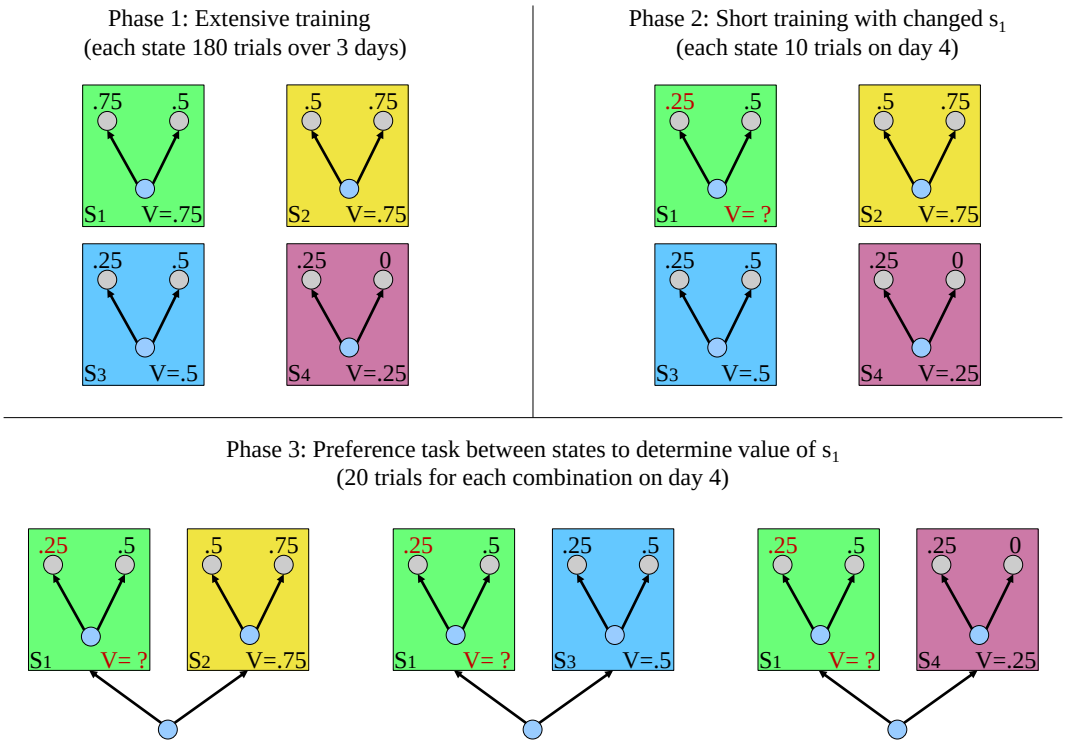

Figure 1: Experimental design with 3 phases and four different states. Each state has two choice options (blue circles with arrows). Each option has a certain probability of wining a point (gray circles with probability above). The value of each state should follow the maximum reward probability if the participant learns the optimal behavior. In phase 2 the reward probability for the optimal option in state 1 changes from .75 to .25. In phase 3 the new value of state 1 is measured by giving the participant the ability to choose at first between two states. All six possible state combinations are tested but only three of them are shown.

4 Behavioral Results

Preliminary data from seven participants was collected (Figure 2). Phase 1 was performed over three days with 60 trials for each state per day. Phase 2 and 3 were performed on the fourth

day. The color associated with a state and the association of a reward probability with the left and right option for each state was randomized over participants.

To infer the Q-value of the non-optimal option o_1 in s_1 the participants needed to learn the optimal options during Phase 1. A criteria of at least 80% of optimal choices for each state during the last day was set. Unfortunately, no participant fulfilled the criteria for all states. One fulfilled the criteria for three states, one for two states, one for one state and four for no states. An effort was made to find a design that makes it easier for the participants to learn the optimal options by changing the reward probabilities for some participants to (0.9; 0.5), (0.9; 0.5), (0.5; 0.1) and (0.1; 0). In addition the sequence of states was changed from a random sequence to a sequence were the same state repeats over several trials. The sample size for each change was too small to give conclusive results, but none of these changes seemed to improve the performance of the participants.

The data from the participant who fulfilled the learning condition in three states and the data from the other participants are separately discussed (Figure 2, top and bottom). The selected participant reached only 62% of optimal choices for state s_4 on the third day. Nonetheless, the data could be used to infer the Q-value of option o_1 because only the preference between state s_1 and s_3 is necessary. Therefore, at the current point only the data of this participant was used for the cognitive modeling. Data of one participant is not enough to make general conclusions but it provided in this preliminary state of the research the possibility to test if different cognitive models could reproduce its behavior.

The first prediction was that the behavior in state s_1 from Phase 1 should be kept during Phase 2 and 3 because of the short relearning time. The selected participant confirmed this prediction by choosing the changed option o_1 in 78.6 percent of the trials during Phase 2 and 3 ($n = 28$). The other participants had a lower mean preference of 51.5% which resulted most likely from their already lower preference during Phase 1 with 59.2% on the third day.

The preference between states measured in Phase 3 (Figure 2 right) followed for all participants the general assumption that states with a higher reward probability are preferred over states with a lower probability ($s_2 > s_3 > s_4$). In the case of the participant who partly fulfills the performance criteria the preference of state s_1 followed ($s_1 < s_3/s_2$) and ($s_1 > s_4$). For the other participants it followed ($s_1 < s_2$) and ($s_1 > s_3/s_4$). Based on the discussion in the previous section this suggests the observed behavior in the selected participant should not be describable by a cognitive model with a critic-only framework. For the other participants a critic-only framework should be able to describe their behavior.

5 Cognitive Modeling

Different cognitive models were analyzed based on the collected experimental data from the one participant who partially fulfilled the learning performance criteria (Figure 2, top). The first model is an adapted version of the Decay MF/MB model by Gläscher et al. (Gläscher et al., 2010). The model-free component uses the SARSA TD algorithm (Equation 11 and 12) and the model-based component uses a FORWARD algorithm that uses the model in a tree-search manner (Section 2.1). The model-free and model-based component compute Q-values which are integrated with a weighted sum (Equation 17). The integration weight decays over time with $w_t = I \cdot e^{-kt}$. Actions are selected based on the integrated Q-values with the soft max method (Equation 9). The original model has five parameters. The learning rate α controls the

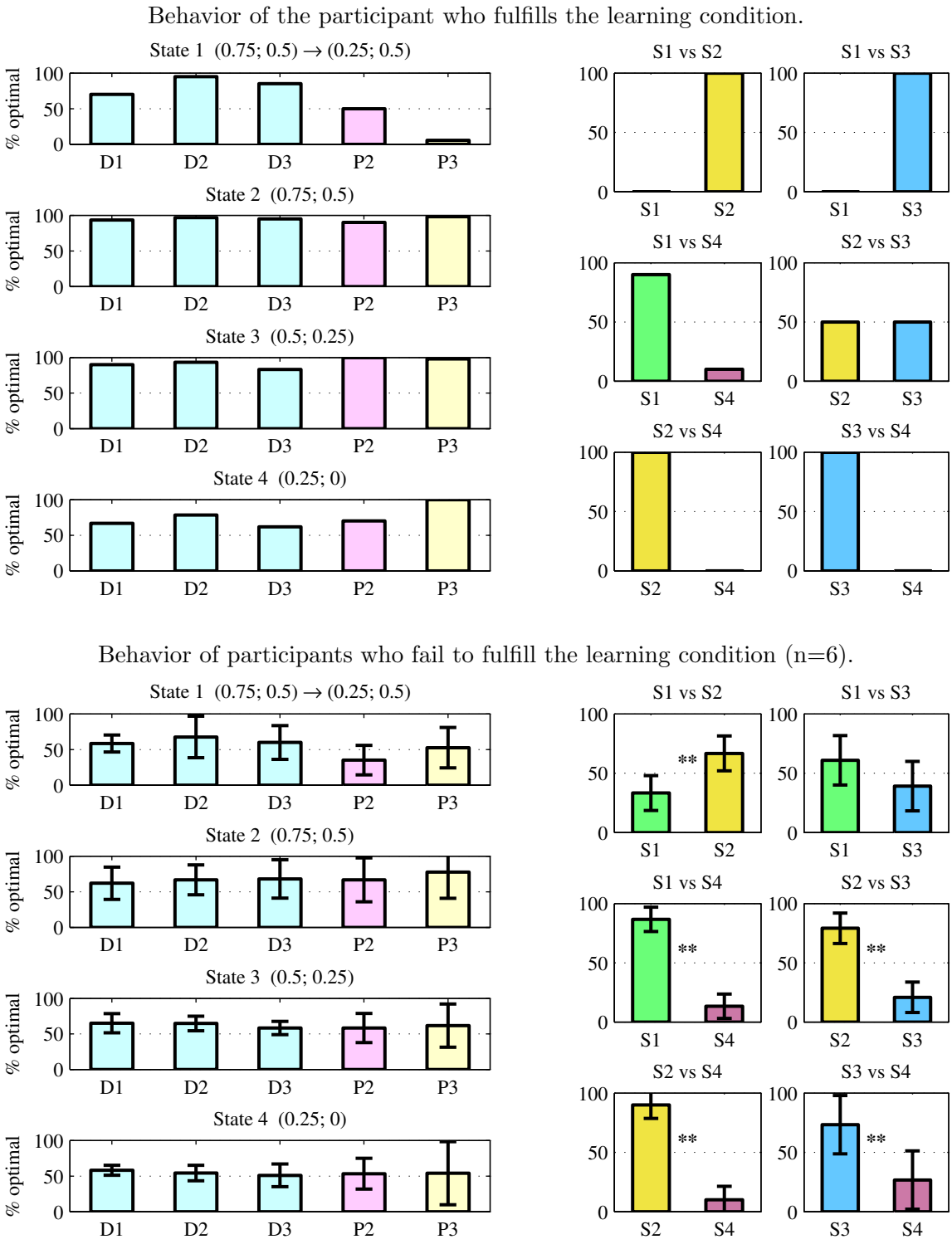

Figure 2: (left) Preference for the option with the higher reward probability for each state of each day in Phase 1 (D1, D2, D3), Phase 2 (P2) and Phase 3 (P3). **(right)** Preference between states during Phase 3. Errorbars depict the standard deviation. Significant differences between state preferences are depicted with ** ($p < 0.01$).

model-free component and η the model-based component. The parameters I and k define the starting value of the weight for the Q-value integration and its decay rate. The action selection is controlled by the inverse temperature parameter β.

The original model was modified to account for the assumption that uncertainty is the main factor to control the integration weight of the components (Daw et al., 2005; Lee et al., 2014). Therefore the weight should change in Phase 2 and 3 for state s_1 because its reward probabilities change resulting in increasing uncertainty. In the original model the weight w can not adapt to such changes. The parameter I_r is added to allow such an adaptation. For state s_1 in Phase 2 and 3 the weight is calculated by $w_t = I_r \cdot e^{-kt}$ with t restarting from zero. The calculation of the weight for all other states does not change. The second and third model are the SARSA learner (MF only) and the FORWARD learner (MB only) alone. They have their respective learning parameter α or η and the inverse temperature β as parameters.

Model & Parameters	**-LL**	**AIC**	**BIC**
Decay MF/MB with parameters from Gläscher et al. (2010) ($\alpha : 0.20$, $\eta : 0.21$, $\beta : 4.91$, $I : 0.63$, $k : 0.09$, $I_r : 0.63$)	395.6	803.2	821.6
Decay MF/MB (slow MB) ($\alpha : 0.10$, $\eta : 0.007$, $\beta : 7.87$, $I : 0.25$, $k : 0$, $I_r : 1$)	364.5	741.0	759.4
Decay MF/MB (slow MF) ($\alpha : 0.06$, $\eta : 0.43$, $\beta : 6.23$, $I : 0.21$, $k : 0$, $I_r : 0$)	367.2	746.4	764.8
SARSA (MF only) ($\alpha : 0.07$, $\beta : 6.01$)	372.8	749.6	755.7
FORWARD (MB only) ($\eta : 0.11$, $\beta : 7.41$)	418.5	841.0	847.1

Table 2: Negativ log likelihood (-LL), BIC and AIC scores for different models based on the behavior of the participant who partially fulfilled the performance criteria (Figure 2, top).

All parameters for each model were identified to produce the best fit to reproduce the behavior of the selected participant who partially fulfills the learning condition (Table 2). The identification followed the general procedure described in Appendix A. A constraint nonlinear optimization algorithm was used to identify the minimum of the negative log likelihood (Equation 20) for each model. Besides the negative log likelihood the AIC and BIC scores were computed. For the Decay MF/MB model two local minima with a similar performance but with different parameter settings could be identified. One uses a slow learning model-based component (slow MB) with $\eta = 0.007$ and the other a slow learning model-free component (slow MF) with $\alpha = 0.06$. The model scores were also calculated for the Decay MF/MB model with the parameters given by Gläscher et al. (2010) to evaluate how well the parameters could be generalized to another experimental task. For the new parameter I_r was the setting of the existing parameter $I = 0.63$ used. Concerning the negative log likelihood the Decay MF/MB (slow MB) had the best performance which was slightly better than the slow MF. They were followed by the SARSA learner, the Decay MF/MB with the parameters from Gläscher et al. and the FORWARD model. The AIC scores followed the order of the negative log likelihood. For the BIC the SARSA model came first because the Decay MF/MB models were penalized due to their higher number of parameters.

The models were further analyzed to test if they are able to reproduce the behavior of the

selected participant. The analysis concentrated on the two highlighted aspects of the participant behavior. First, if the models were able to reproduce the effect that the participant kept selecting the previous optimal option o_1 of state s_1 during Phase 2 and 3. The second aspect was the strong preference of the participant of states s_2 and s_3 over s_1 in Phase 3. Each model was used to simulate the experimental task 100 times (Table 3). The full behavior of the slow MB Decay MF/MB model is shown in Figure 3. The behavior of the slow MF model is similar but no significant difference between the preference of state s_1 and state s_3 existed. The other models differed mainly in their ability to keep the learned behavior for s_1 from Phase 1 also in Phase 2 and 3. The analysis is discussed in the following part. It showed that the slow MB, slow MF and the SARSA model could reproduce the behavior of keeping the non-optimal option o_1 but failed to show a preference of state s_3 over s_1.

	Preference for o_1 of s_1 in Phase 2 and 3	**Preference for s_1 in Phase 3**		
		s_1 vs s_2	s_1 vs s_3	s_1 vs s_4
Participant	78.6%	0.0%	0.0%	90.0%
Decay MF/MB (Gläscher)	37.3% (10.0)	21.1% (16.5)	45.5% (19.8)	76.0% (14.4)
Decay MF/MB (slow MB)	72.8% (7.28)	25.1% (19.6)	54.9% (23.3)	75.2% (14.7)
Decay MF/MB (slow MF)	68.5% (12.2)	28.4% (15.6)	49.1% (16.8)	69.6% (15.5)
SARSA (MF only)	63.7% (14.6)	36.6% (23.7)	49.1% (21.7)	59.9% (16.4)
FORWARD (MB only)	43.3% (9.89)	13.5% (8.00)	45.8% (13.3)	84.4% (9.92)

Table 3: Behavioral results of the selected participant and the models. First, the preference for the choice option o_1 in state s_1 which changes the reward probability from 0.75 to 0.25 for Phase 2 and 3 is listed. Secondly, the preference of state s_1 in Phase 3 when it gets compared to s_2, s_3 and s_4 is given. Each model performed the task 100 times. The standard deviation of their behavior is given in brackets.

The first aspect of the analysis was the high rate of 78.6% to keep choosing the previous optimal option o_1 in state s_1 during Phase 2 and 3 of the selected participant. The slow MB, slow MF and the SARSA model seemed to be able to replicate the effect with a choice rate of 72.8%, 68.5% and 63.7%. For the Decay MF/MB models this was explained by the small learning rate of either the model-free or the model-based component. This slow component could not adapt fast enough to the new reward probability. Therefore, in Phase 2 and 3 this component dominated the Q-values used for state s_1 and the behavior of Phase 1 was still observed. For the slow MB model the I_r was 1. Therefore, the Q-values of the slow learning model-based component dominated the weighted Q-value. In contrast the I_r of the slow MF model was 0 resulting in using only Q-values of the model-free component. The SARSA model could replicate the effect because it had a small learning rate that is used for all states in each phase. The FORWARD and the Decay MF/MB model with the parameters from Gläscher et al. could not replicate the effect. They had a rate of 43.3% and 37.3% of choosing the previous optimal option. The learning rate for both models was too high making the adaptation to the new reward probability too fast.

The second aspect to analyze the models was the participant's strong preference of states s_2 and s_3 over s_1 in Phase 3. All models showed a preference of state s_2 over s_1 but not to the same extent as the participant (Table 3). Moreover, the models struggled to replicate a preference of state s_3 over s_1. The participant had a preference of 0% whereas all models had a mean preferences around 50%. In the case of the models which showed a stable preference for the non-optimal option in state s_1 during Phase 2 and 3 (slow MB, slow MF Decay MF/MB and

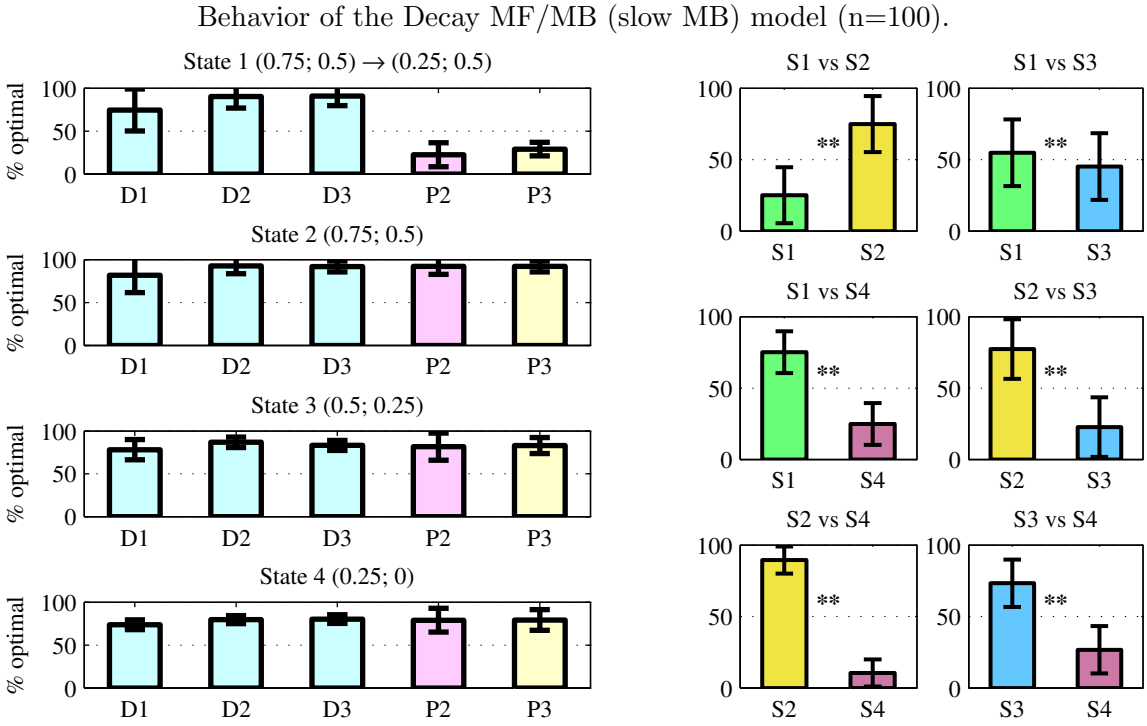

Figure 3: (left) Preference for the option with the higher reward probability for each state of each day in Phase 1 (D1, D2, D3), Phase 2 (P2) and Phase 3 (P3). **(right)** Preference between states during Phase 3. Errorbars depict the standard deviation. Significant differences between state preferences are depicted with ** ($p < 0.01$).

SARSA) the missing strong preference of s_3 over s_1 was caused by the model-free mechanisms. For both Decay MF/MB models has the model-free component in general a stronger influence on the Q-values than the model-based component because of $I = 0.25$ and $I = 0.21$. The effect that the model-free component shows no strong preference effect has the following cause. The decisions between the states s_1, s_2, s_3 and s_4 were modeled as new states and the model-free component started with 0 for the Q-values of each choice between states. It had only 20 trials to learn the Q-values and all three models used a relatively small learning rate ($\alpha = 0.1$, $\alpha = 0.06$, $\alpha = 0.07$). As a result the slow MB, the slow MF Decay MF/MB and the SARSA model had a problem reproducing the strong preference effect of s_3 over s_1 from the participant. The FORWARD and the Decay MF/MB model with the parameters from Gläscher et al. model had no strong preference between states s_1 and s_3 because they learned to choose the new optimal option in s_1 faster which had the same reward probability as the optimal option in s_3.

In summary, different cognitive models were fitted to the data of the one participant who fulfilled partially the learning criteria and showed the effect of a divergence between learned values and decision behavior. The slow MB, slow MF Decay MF/MB and the SARSA model were able to reproduce the participant's behavior of continuing to chose the previous optimal option o_1 in state s_1. Nonetheless, they were not able to show a strong preference of states s_2 and s_3 over s_1 as observed in the participant.

6 Discussion

The experiment tested the simplification of current MF/MB models that decisions are made with a critic-only approach. A behavioral effect where decision behavior diverges from Q-values was predicted which was predicted not to be reproducible by current models. Preliminary data from seven participants was collected. Six of the participant were excluded because their low learning performance interferes with the disposition of the effect. The participant who partially fulfilled the learning criteria showed the predicted effect.

The cognitive modeling showed that the MF/MB model by Gläscher et al. (2010) which uses a critic-only framework can not reproduce the effect of a divergence of values and decision behavior. This would lead to the conclusion that critic-only frameworks can not explain the observed behavior. Nonetheless, other MF/MB critic-only cognitive models such as by Daw et al. (2005) and Lee et al. (2014) need to be tested to validate that they can also not reproduce the observed behavior.

Models with actor-critic frameworks are predicted to be able to reproduce the effect by allowing a divergence between the values of their critic component and the learned decision behavior defined by their actor component. To test this prediction an actor-critic model needs to be fitted to the experimental data to see if it is able to reproduce the effect.

One major problem with the preliminary data is its small sample size which does not allow to draw consistent conclusions. More participants are needed to see if they also show the anticipated behavioral effect. A related problem is that most participants do not fulfill the learning criteria resulting in an exclusion of six out of seven participants. Therefore too many participants would be needed to gain a large enough pool of participants which fulfill the learning criteria. The learning task seems to difficult and its complexity should be reduced. One possibility is the reduction of the number of states from four to two (s_1, s_3). This will be the next step that will by tried besides the implementation and analysis of the missing MF/MB models and actor-critic alternatives.

Appendix A: The Cognitive Modeling Procedure

Cognitive modeling uses computer models to describe cognitive processes. The models have the advantage of being clear descriptions that produce testable predictions and can therefore be comparable. During the project different models will be implemented, analyzed and compared. The cognitive modeling consists of four steps: 1) identification of theories describing the cognitive processes, 2) formalization of the theories as computational models, 3) identification of model parameters and 4) the analysis and comparison of the different models. The applied modeling methods follow standard procedures as for example described by Daw (2011).

The first step is to identify different theories that describe the cognitive processes of interest. Existing theories will be identified in the literature and new theories will be developed. New theories could be based on existing ones by integrating or adapting them to cope with unexplained behavioral effects. The main focus will be on theories based on reinforcement learning.

The second step is to formalize the theories as computational models. Models have the form of agents in the MDP framework. They receive information about the current state s_t of the environment and have to perform an action a_t. Actions are drawn from a probabilistic policy $Pr(a_t|s_t)$. The resulting reward r_{t+1} and state s_{t+1} are provided to the agent. The cognitive processes are represented as computer algorithms (e.g. TD learning). Internal representations are the latent variables (e.g. Q-values) of the models. The behavior of the model is defined by parameters θ (e.g. the learning rate α). Reinforcement learning methods will mainly applied as algorithms. Nonetheless other methods will be considered such as Bayesian techniques, neural networks or dynamical systems if they are able to improve the models.

Step three is the identification of the parameters θ for each model that allow the model to optimally describe and predict behavior. The identification is based on data recorded during the human behavioral experiments that are part of the proposed PhD project. Publicly available data in research literature and from other research units may also be used to allow a better identification and analysis of models. The experiments follow the MDP framework and record for each participant a trace. Traces consist of the state s_t, the performed action a_t and the received reward r_{t+1} for each time step during the experiment:

$$\tau = \{(s_1, a_1, r_2),\ \dots\ , (s_T, a_T, r_{T+1})\} \tag{18}$$

How well a model describes the behavior with parameters θ is measured by the likelihood of the trace given the parameters $Pr(\tau|\theta)$. The likelihood is measured by simulating the experiment with the model as agent. Each time step the model receives the current state and reward as given by the trace. The model updates its internal variables and the probabilistic policy $Pr_\theta(a_t|s_t)$ based on the information, but instead of drawing a random action from the policy the action from the trace is used. The likelihood is defined as the probability of the perceived actions in the trace given the policy of the model:

$$Pr(\tau|\theta) = \prod_{t=1}^{T} Pr_\theta(a_t|s_t) \tag{19}$$

Usually the negative log likelihood is used instead because the product goes to zero making it hard to be represented numerically. The objective function becomes:

$$J(\theta) = -\sum_{t=1}^{T} \log Pr_\theta(a_t|s_t) \tag{20}$$

Standard nonlinear optimization methods are applied to find the minimum such as gradient decent techniques or evolutionary algorithms. For each participant and model individual parameters θ_τ will be identified.

Besides the individual parameters group level parameters will also be identified. They describe the behavior over many subjects, for example by combining all traces to a combined data set. Another option is the usage of mixture models such as the hierarchical Bayesian analysis. They hold a hyperprior distribution over the individual parameters $Pr(\theta|\phi)$ with hyperprior parameters ϕ. The hyperprior provides a description of the individual differences in participants and can be used to analyze such differences further.

The final step of cognitive modeling is to compare and analyze the models by different criteria. A key criteria is the goodness of fit which measures how well a model describes the data. Another is the model complexity. The goodness of fit is often measured by the likelihood (Equation 19) and the model complexity by the number of parameters θ of a model. Standard procedures such as AIC or BIC use both criteria to find the simplest model that can describe the data. The goodness of fit will also be tested by measuring how well models can generalize, i.e. how well they can predict behavior in novel experiments. Other criteria to analyze the models are their plausibility to exist as brain functions, their computational (cognitive) effort or their learning performance. The goal is to identify the properties of each model and to find the model which is able to describe best the cognitive processes behind decision learning.

References

Adamantidis, A. R., Tsai, H.-C., Boutrel, B., Zhang, F., Stuber, G. D., Budygin, E. A., Touriño, C., Bonci, A., Deisseroth, K., & de Lecea, L. (2011). Optogenetic interrogation of dopaminergic modulation of the multiple phases of reward-seeking behavior. *The Journal of Neuroscience*, *31*(30), 10829–10835.

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action–outcome predictor. *Nature Neuroscience*, *14*(10), 1338–1344.

Atallah, H. E., Lopez-Paniagua, D., Rudy, J. W., & O'Reilly, R. C. (2006). Separate neural substrates for skill learning and performance in the ventral and dorsal striatum. *Nature Neuroscience*, *10*(1), 126–131.

Barnes, J. I., McColeman, C., Stepanova, E., Blair, M. R., & Walshe, R. C. (2014). Rlattn: An actor-critic model of eye movements during category learning. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In D. B. J.C. Houk, J.L. Davis (Ed.) *Models of Information Processing in the Basal Ganglia*, (pp. 215–232). Cambridge, MA: MIT Press.

Bellman, R. (1957). *Dynamic programming*. Princeton University Press.

Belova, M. A., Paton, J. J., & Salzman, C. D. (2008). Moment-to-moment tracking of state value in the amygdala. *The Journal of Neuroscience*, *28*(40), 10023–10030.

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

Brown, J., Bullock, D., & Grossberg, S. (1999). How the basal ganglia use parallel excitatory and inhibitory learning pathways to selectively respond to unexpected rewarding cues. *The Journal of Neuroscience*, *19*(23), 10502–10511.

Contreras-Vidal, J. L., & Schultz, W. (1999). A predictive reinforcement model of dopamine neurons for learning approach behavior. *Journal of Computational Neuroscience*, *6*(3), 191–214.

Corbit, L. H., Muir, J. L., & Balleine, B. W. (2001). The role of the nucleus accumbens in instrumental conditioning: evidence of a functional dissociation between accumbens core and shell. *The Journal of Neuroscience*, *21*(9), 3251–3260.

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. *Decision Making, Affect, and Learning: Attention and performance XXIII*, *23*, 3–38.

Daw, N. D. (2012). Model-based reinforcement learning as cognitive search: neurocomputational theories. *Cognitive Search: Evolution, Algorithms and the Brain*, (pp. 195–208).

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

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

Deisenroth, M., & Rasmussen, C. E. (2011). Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, (pp. 465–472).

Deisenroth, M. P., Neumann, G., & Peters, J. (2011). A survey on policy search for robotics. *Foundations and Trends in Robotics*, *2*(1-2), 1–142.

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.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*, *12*(7), 961–974.

Doya, K. (2000). Reinforcement learning in continuous time and space. *Neural Computation*, *12*(1), 219–245.

Doya, K. (2002). Metalearning and neuromodulation. *Neural Networks*, *15*(4-6), 495–506.

Doya, K., Samejima, K., Katagiri, K.-i., & Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural Computation*, *14*(6), 1347–1369.

Gläscher, J., Daw, N., Dayan, P., & 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.

Houk, J. C., Adams, J. L., & Barto, A. G. (1995). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In D. B. J.C. Houk, J.L. Davis (Ed.) *Models of Information Processing in the Basal Ganglia*, (pp. 249–270). Cambridge, MA: MIT Press.

Ito, M., & 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., & Doya, K. (2011). Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit. *Current Opinion in Neurobiology*, *21*(3), 368–373.

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, *15*(4), 535–547.

Joel, D., & 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.

Jog, M. S., Kubota, Y., Connolly, C. I., Hillegaart, V., & Graybiel, A. M. (1999). Building neural representations of habits. *Science*, *286*(5445), 1745–1749.

Joseph, D., Gangadhar, G., & Srinivasa Chakravarthy, V. (2010). Ace (actor–critic–explorer) paradigm for reinforcement learning in basal ganglia: Highlighting the role of subthalamic and pallidal nuclei. *Neurocomputing*, *74*(1), 205–218.

Kable, J. W., & Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, *10*(12), 1625–1633.

Kakade, S. (2001). A natural policy gradient. In *Advances in Neural Information Processing Systems*, vol. 14, (pp. 1531–1538).

Khamassi, M., Lachèze, L., Girard, B., Berthoz, A., & Guillot, A. (2005). Actor–critic models of reinforcement learning in the basal ganglia: from natural to artificial rats. *Adaptive Behavior*, *13*(2), 131–148.

Kurth-Nelson, Z., & Redish, A. D. (2009). Temporal-difference reinforcement learning with distributed representations. *PloS One*, *4*(10), e7362.

Kurth-Nelson, Z., & Redish, A. D. (2010). A reinforcement learning model of precommitment in decision making. *Frontiers in Behavioral Neuroscience*, *4*.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, *81*(3), 687–699.

Luman, M., Tripp, G., & Scheres, A. (2010). Identifying the neurobiology of altered reinforcement sensitivity in adhd: a review and research agenda. *Neuroscience & Biobehavioral Reviews*, *34*(5), 744–754.

Maia, T. V. (2010). Two-factor theory, the actor-critic model, and conditioned avoidance. *Learning & Behavior*, *38*(1), 50–67.

Nakahara, H., Doya, K., Hikosaka, O., & Nagano, S. (1998). Reinforcement learning with multiple representations in the basal ganglia loops for sequential motor control. In *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on*, vol. 2, (pp. 1553–1558). IEEE.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science*, *304*(5669), 452–454.

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, *5*(2), 97–98.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045.

Reynolds, J. N., & Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks*, *15*(4), 507–521.

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, *310*(5752), 1337–1340.

Schembri, M., Mirolli, M., & Baldassarre, G. (2007). Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot. In *Development and Learning, 2007. ICDL 2007. IEEE 6th International Conference on*, (pp. 282–287). IEEE.

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *The Journal of Neuroscience*, *27*(47), 12860–12867.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–1599.

Schultz, W., & Romo, R. (1988). Neuronal activity in the monkey striatum during the initiation of movements. *Experimental Brain Research*, *71*(2), 431–436.

Schultz, W., Tremblay, L., & Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral Cortex*, *10*(3), 272–283.

Schweighofer, N., Bertin, M., Shishida, K., Okamoto, Y., Tanaka, S. C., Yamawaki, S., & Doya, K. (2008). Low-serotonin levels increase delayed reward discounting in humans. *The Journal of Neuroscience*, *28*(17), 4528–4532.

Shiflett, M. W., & Balleine, B. W. (2010). At the limbic–motor interface: disconnection of basolateral amygdala from nucleus accumbens core and shell reveals dissociable components of incentive motivation. *European Journal of Neuroscience*, *32*(10), 1735–1743.

Silvetti, M., Alexander, W., Verguts, T., & Brown, J. W. (2013). From conflict management to reward-based decision making: actors and critics in primate medial frontal cortex. *Neuroscience & Biobehavioral Reviews*.

Silvetti, M., Seurinck, R., & Verguts, T. (2011). Value and prediction error in medial frontal cortex: integrating the single-unit and systems levels of analysis. *Frontiers in Human Neuroscience*, *5*.

Suri, R., Bargas, J., & Arbib, M. (2001). Modeling functions of striatal dopamine modulation in learning and planning. *Neuroscience*, *103*(1), 65–85.

Suri, R. E., & Schultz, W. (1998). Learning of sequential movements by neural network model with dopamine-like reinforcement signal. *Experimental Brain Research*, *121*(3), 350–354.

Suri, R. E., & 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., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge Univ Press.

Takahashi, Y., Schoenbaum, G., & Niv, Y. (2008). Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. *Frontiers in Neuroscience*, *2*(1), 86.

Tanaka, S. C., Schweighofer, N., Asahi, S., Shishida, K., Okamoto, Y., Yamawaki, S., & Doya, K. (2007). Serotonin differentially regulates short-and long-term prediction of rewards in the ventral and dorsal striatum. *PLoS One*, *2*(12), e1333.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189.

Tsai, H.-C., Zhang, F., Adamantidis, A., Stuber, G. D., Bonci, A., De Lecea, L., & Deisseroth, K. (2009). Phasic firing in dopaminergic neurons is sufficient for behavioral conditioning. *Science*, *324*(5930), 1080–1084.

Uchibe, E., & Doya, K. (2004). Competitive-cooperative-concurrent reinforcement learning with importance sampling. In *Proc. of International Conference on Simulation of Adaptive Behavior: From Animals and Animats*, (pp. 287–296).

Voorn, P., Vanderschuren, L. J., Groenewegen, H. J., Robbins, T. W., & Pennartz, C. (2004). Putting a spin on the dorsal–ventral divide of the striatum. *Trends in Neurosciences*, *27*(8), 468–474.

Wan, X., & Peoples, L. L. (2006). Firing patterns of accumbal neurons during a pavlovian-conditioned approach task. *Journal of Neurophysiology*, *96*(2), 652–660.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*(3-4), 229–256.

Wunderlich, K., Dayan, P., & Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience*, *15*(5), 786–791.

Yin, H. H., Knowlton, B. J., & 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., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *European Journal of Neuroscience*, *22*(2), 513–523.