THE UNIVERSITY OF
SYDNEY

Hierarchical models of goal-directed and automatic actions

Amir Dezfouli
March 2015

University of Sydney

Faculty of Medicine

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

Statement of Authentication

This thesis is submitted to the University of Sydney in fulfilment of the requirements for the Degree of Doctor of Philosophy.

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that I have not submitted this material, either in full or in part, for a degree at this or any other institution.

Signature  Date 29 March, 2015

To my wife and parents.

Acknowledgements

I would like to express my deepest gratitude to my advisor Professor Bernard Balleine for his guidance and encouragement, and providing me with an excellent atmosphere for doing research. I was extremely fortunate to have Bernard as my supervisor, who gave me the freedom to explore the world of neuroscience and psychology on my own, and at the same time the guidance to correct my steps when faltered. I am also grateful to the members of Balleine's Lab, with whom I have interacted and collaborated during the past years: Dr. Richard Morris, Dr. Nura Lingawi, Dr. Laura Bradfield, Dr. Vincent Laurent, Dr. Kristi Griffiths, Mr. Allan Arraf, Dr. Jesus Bertran-Gonzalez, Dr. Teri Furlong, Dr. Shauna Parkes, Dr. Genevra Hart, Dr. Beatrice Leung, Dr. Chin Chieng, and Dr. Serena Becci.

I would like to acknowledge Dr. Majid Nili Ahmadabadi; my motivation to study psychology and neuroscience originated from the course "istributed Artificial Intelligence", which he presented at the University of Tehran. Later on, I had the privilege of working with Dr. Caro Lucas (R.I.P), Dr. Laleh Ghadakpour, and Dr. Hamed Ekhtiari, who had a profound impact on my research, and also my motivation to pursue a PhD. I would also like to thank my former colleagues Dr. Mahdi Keramati, and Payam Piray, for sharing ideas and insightful discussions. I would also like to extend my special thanks to Dr. Yael Niv, and Dr. Nathaniel Daw, for their help and encouragement, which had a great impact on my research and my scientific career.

I would like to thank my parents Giti and Hamid, and my brothers, Behnam, Majid, and Mehrdad for their love and encouragement during the past years. I would also like to thank my in-laws, Dr. Reza Berangi and Mrs Seddigheh Ahmadian who supported me in every way possible throughout this time, and shared their experiences of a similar journey with me.

Above all, I would like to thank my wife Tahereh for her love and constant support, which

made the current thesis possible.

Sydney, 30 March, 2015 A. D.

Abstract

Decision-making processes behind instrumental actions can be divided into two categories: goal-directed processes, and automatic (or habitual) processes. Goal-directed processes evaluate actions based on the value of their consequences and produce a course of action that will lead to desired goals. This form of decision-making, however, is computationally demanding and prone to suffer from the "curse of dimensionality", i.e., it does not scale to complex environments. To overcome this limitation, individuals employ automatic actions, which are less flexible than goal-directed actions, but are also less expensive in terms of the required computational resources. The structure of such automatic actions, their interaction with goal-directed actions, and their behavioral and computational properties are the topics of the current thesis. We conceptualize the structure of automatic actions as sequences of actions that form a single response unit and are integrated within goal-directed processes in a hierarchical manner. We then represent this hypothesis using the computational framework of reinforcement learning and develop a new normative computational model for the acquisition of action sequences, and their hierarchical interaction with goal-directed processes. We develop a neurally plausible hypothesis for the role of neuromodulator dopamine as a teaching signal for the acquisition of action sequences. We further explore the predictions of the proposed model in a two-stage decision-making task in humans and we show that the proposed model has higher explanatory power than its alternatives. Finally, we translate the two-stage decision-making task to an experimental protocol in rats and show that, similar to humans, rats also use action sequences and engage in hierarchical decision-making. The results provide a new theoretical and experimental paradigm for conceptualizing and measuring the operation of goal-directed and automatic actions and their interactions.

Acknowledgements

Key words: Decision-making, Goal-directed actions, Habitual actions, Hierarchical decision-making, Reinforcement learning

Care and Use of Animals

The care and use of animals in the research presented in this thesis complies with the Rules Governing the Use of Animals in Research and Teaching at the University of Sydney with the Australian Code of Practice for the Care and Use of Animals for Scientific Purposes Act (1985 and its subsequent amendments). These experiments were approved by the Animal Ethics Committee at the University of Sydney.

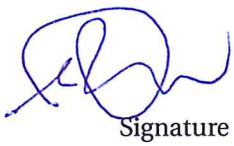
Collaborator and co-author declaration

We, the undersigned, acknowledge the following statement:

This thesis principally represents the work of Amir Dezfouli. Prof. Bernard W. Balleine provided considerable support in all the experiments and the preparation of the thesis. Prof. Balleine's support was most prominent in experimental design, as well as in the editing of written work.

Bernard W. Balleine

Date 25.3.15



Signature

Publications

- Dezfouli, A., & Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. European Journal of Neuroscience (Vol. 35, pp. 1036–1051).
- Dezfouli, A., & Balleine, B. W. (2013). Actions, Action Sequences and Habits: Evidence that Goal-Directed and Habitual Action Control are Hierarchically Organized. PLoS Computational Biology, 9(12).
- Dezfouli, A., Lingawi, N. W., & Balleine, B. W. (2014). Habits as action sequences: hierarchical action control and changes in outcome value. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1655), 20130482.

Contents

List of Figures

List of Tables

1 Introduction

Behavioral evidence suggests that decision-making systems behind the choices made by humans and other animals can be divided to two categories, known as goal-directed, and automatic processes (also referred to as habitual processes). Goal-directed action is a form of decision-making guided by encoding the relationship between actions and their consequences and the value of those consequences. In this form of decision-making an agent deliberates over the consequences of its actions, and chooses the course of action that leads to desired outcome. *Outcome devaluation* studies provide direct evidence that both humans (Valentin, Dickinson, & O'Doherty, 2007; Tricomi, Balleine, & O'Doherty, 2009) and other animals (C. D. Adams, 1982; Dickinson, 1994; Dickinson, Squire, Varga, & Smith, 1998; Tolman, 1948) engage in this form of action control. For example, in a typical experiment an agent is first trained to perform two different actions that earn different food outcomes. After this training, an outcome devaluation treatment is conducted off baseline or *offline*; i.e., in a situation where the outcome is presented without the action being performed, a treatment that generally involves sating the animals on one of the two outcomes to decrease its value. Subsequently, back *online*, a test is conducted in which choice between the two actions is assessed in the absence of the outcome. Typically, when given this choice, humans and other animals decrease their performance of the action that previously delivered the now devalued outcome, demonstrating that such actions reflect both the relationship to their consequences and the

value of those consequences.

Extended training makes goal-directed actions habitual or automatic (we will use these two terms interchangeably hereafter). This automaticity has two manifestations: (i) inflexibility of actions to the offline changes in the value of their outcomes (Coutureau & Killcross, 2003; Dickinson et al., 1998; Quinn, Pittenger, Lee, Pierson, & Taylor, 2013; Wassum, Cely, Maidment, & Balleine, 2009; Yin, Knowlton, & Balleine, 2004, 2005)(Figure 1.1), and (ii) the concatenation of actions executed together to form action sequences that are then treated as a single response unit (Lashley, 1951; Book, 1908)(Figure 1.2). These two aspects of automaticity share a similar neural structure, however, computationally, they have been attributed to two different models: insensitivity to changes in outcome value has often been interpreted as evidence for a model-free reinforcement learning (RL) account of instrumental conditioning (Daw, Niv, & Dayan, 2005), whereas the development of action sequences has been linked to hierarchical RL (Botvinick, 2008; Ito & Doya, 2011).

The current thesis develops the hypothesis that a goal-directed hierarchical RL system can explain various aspects of decision-making processes, including the behavioral phenomena which are attributed to model-free RL. We develop a new computational model, and test its predictions in both humans and rats. The structure of the thesis is as follows: in chapter 2 we review multiple forms of decision-making systems in the brain. In chapter 3 we provide a new neurally plausible computational model for action sequence learning, and investigate its behavioral power in explaining automatic actions. In chapter 4 we provide data from a decision-making task in humans, and we directly compare the proposed hierarchical RL account and the model-free RL account, in explaining choices. In chapter 5 we develop a task similar to that we developed in humans, and we provide evidence for the operation of the hierarchical RL in rats. Finally, in chapter 6 we provide the summary and conclusion of the thesis.

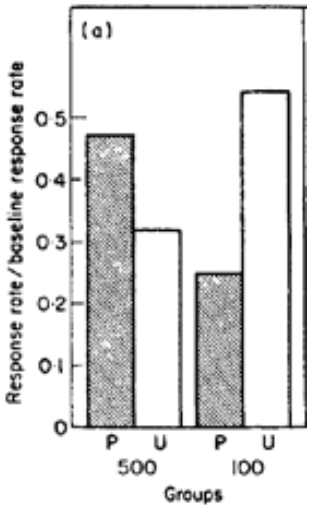


Figure 1.1 – The effect of training on the sensitivity of actions to the value of their outcomes. Animals were trained to press a lever in order to earn sucrose pellets. One group of animals received 100 pairings of lever presses with sucrose pellets (Group 100), while the other group received 500 pairings (Group 500). Before the extinction test, within each group, for some of the animals the outcome was devalued (Group 100-P and Group 500-P), while for the rest the outcome remained valued (Group 100-U and Group 100-U). As the figure shows, the number of responses in Group 500-P is not effected by the devaluation of the outcome, while in Group 100-P, the responses for earning the devalued outcome have decreased. From (C. D. Adams, 1982)

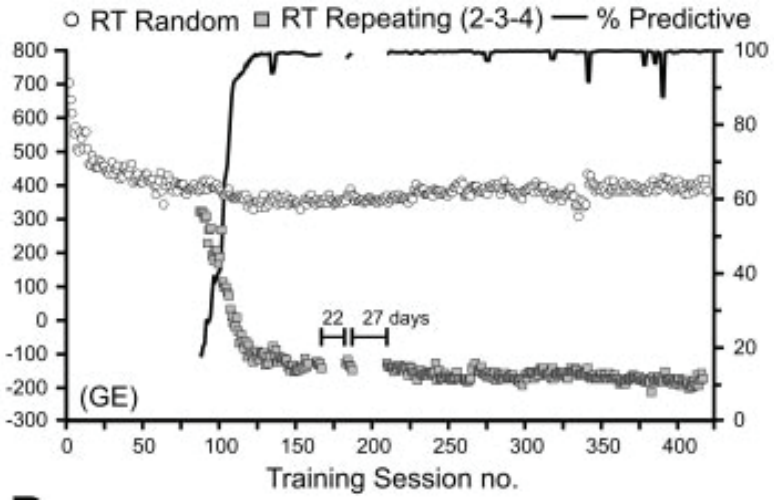


Figure 1.2 – Evidence for the formation of action sequences in the sequential reaction time task (SRTT). When the subjects were asked to perform a fixed sequence of actions (RT repeating), reaction times (RT) decreased, which is a sign of the formation of action sequences. The solid line represents 'predictive responses', i.e., the responses that elicit even before the corresponding stimulus signals that the response should be made. As the figure shows, in the late stages of training almost all the responses are predictive, which is another sign of the formation of action sequences. In contrast, when the sequence of actions to be performed was variable (RT random), reaction times remained slow relative to repeating trials even after extended training. From (Matsuzaka et al., 2007)

.

2 Multiple forms of decision-making

Humans and other animals are able to adjust their decisions by learning about their surrounding environment. Evidence shows that such decisions, are driven by multiple decision-making processes, which have partly overlapping neural, computational, and behavioral characteristics. The aim of this chapter is to provide an overview of each system, and also how these systems interact in order to produce final actions. We start by introducing some concepts from *reinforcement learning (RL)* literature (Sutton & Barto, 1998) - a widely used framework for studying decision-making processes in the brain - and then, building on this theoretical framework, we review the psychological and neural properties of each system. The chapter is divided into two-sections: the prediction problem (section 2.1), and the control problem (section 2.2), which as their titles indicate, correspondingly address how predictions about future events are made, and how the predictions are used in conjunction with actions in order to control the environment. These two sorts of problems roughly map onto what are known as *Pavlovian conditioning*, and *Instrumental conditioning* in the animal learning literature, which we will present in turn.

2.1 Pavlovian conditioning: the prediction problem

In a classic experiment, Pavlov (1849-1936) trained a dog in a condition in which the presentation of a stimulus (e.g., a bell) was followed by food. He observed that the dog began to salivate in response to the presentation of the stimulus, even before the food was being delivered (Pavlov, 1927), which suggested that the dog had associated the food with the bell. This phenomenon is called *Pavlovian conditioning*, which in general is composed of three elements: the predicting stimulus (the bell in this example), which is called *conditioned stimulus (CS)*, the predicted stimulus (food in this example), which is called *unconditioned stimulus (US)*, and the response that is triggered in anticipation of the US (salivation in this example), which is called *conditioned responses (CR)*.

The association between a CS and a US can take place at different levels. Indeed, both the CS and US have a variety of sensory and motivational features, which can enter into association. In general, features of a US can be divided into three categories: *sensory features*, *motivational features*, and *a reinforcer component*. The sensory features of a US define the identity of the US, which for example can refer to the specific taste of a food or its texture that uniquely distinguishes it from other stimuli. The motivational features are the representation of the US across the dimensions of primary incentives such as nutrients, salts, fluids, etc. Finally, the reinforcer component represents the hedonic component of a US. For example, for a hungry animal, receiving two units of food pellets has a higher reinforcer component than receiving one unit of food pellet. During the conditioning process, an animal associates all, or a subset of the US features with the CS. Later on, whenever the CS is present, the animal elicits CRs in anticipation that the US will shortly occur (Figure 2.1). In the computational models of conditioning processes (next sections), the reinforcer component is usually called a *reward component* and hereafter we use the term reward component instead of the reinforcer component.

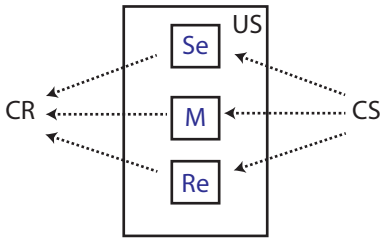


Figure 2.1 – An unconditioned stimulus is generally composed of sensory features (Se), motivational features (M), and a reward (reinforcer) component. A CS can enter into an association with either of these components, which will then trigger the associated conditioned response whenever the CS is present.

2.1.1 Markov Reward Process

With some simplifications, the structure of Pavlovian conditioning can be formalized using *Markov Reward Processes*. A Markov Reward Process is a tuple $(\mathscr{S}, \{P(.|s)\}, R)$, where:

- $\mathscr{S}$ is a set of states,
- $\{P(.|s)\}$ is the transition model. $P(s'|s)$ represents the probability of transition to state s' if the current state is s.
- R is a reward function. $R(s)$ is the reward of state s.

The term *state* in the above definition refers to distinguishable situations or events in the environment that an individual perceives through its sensory inputs. In Pavlovian conditioning, for example, presentation of a CS and US correspond to two different states of the environment, which we can denote by s^{CS}, and s^{US} correspondingly. In the previous example, the state s^{CS} is identified by the presentation of the bell signaled by its sensory features such as the sound that it makes, and state s^{US} is identified by the sensory features of the US signaled by the sensory features of the food such as its taste and texture. As such, among the three mentioned features (sensory, motivational, reward), the sensory features of the US and CS are represented by the states of the environment. Similarly, $R(s)$ represents the reward component of a state, as for example, $R(s^{\text{US}})$ represents the reward component of the US. Finally, the transition model corresponds to the relation between the states (for example, if the CS (the bell) is always

followed by the US, then we will have $P(s^{\text{US}}|s^{\text{CS}}) = 1$).

Thus, in summary, according to a Markov Reward Process, (i) an environment is composed of several states, (ii) each state has an amount of reward associated with it, and (iii) there is a transition model that defines how states are related to each other. As such, there is no inherent difference between CS and US; both are states in the environment with certain sensory and reward features. Based on this, we can refer to the states in the environment by their temporal order instead of stating whether they are US or CS: the state of the environment at time t is referred to as s_t. For example, within the Pavlovian conditioning, if we assume the origin of time is when the CS is presented, then we have $s_0 = s^{\text{CS}}$, and $s_1 = s^{\text{US}}$. Similarly, the reward received at time t is referred to as r_t (for example, $r_0 = R(s^{\text{CS}})$ and $r_1 = R(s^{\text{US}})$).

The definition of the Markov Reward Process provides a computational description of the elements of the Pavlovain conditioning process. However, remaining question is how an animal learns and represents a Markov Reward Process. Here, evidence shows that animals learn about a Markov Reward Process in two different ways. In the first way, they associate the CS with the sensory feature of the US, i.e., by presenting the CS, sensory features of the US can be predicted. In the second way, by presenting the CS, only the reward component of the US can be predicted. These two forms of Pavlovian conditioning, at the computational level, correspond to what are known as *model-based* and *model-free* learning. We will discuss these in turn in the next two sections.

2.1.2 Model-based Pavlovian conditioning

In model-based Pavlovian conditioning, an individual learns the transition model of the environment (and this is why this form of learning is called model-based), which allows it to predict the next state, given the current state of the environment. An example is predicting the sensory features of the upcoming US during the CS presentation. Such learning can take place for example using the Rescorla-Wagner rule (Rescorla & Wagner, 1972), according to which the association between the CS and the US gradually increases with each pairing of

the US and CS. Within the context of the Markov Reward Process, the Rescorla-Wagner rule corresponds to the learning of the transition model of the environment[1], i.e., the strength of the association between a CS and a US in the Rescorla-Wagner model is equivalent to the probability of the CS given the US. Inspired by the Rescorla-Wagner rule, for each pairing of CS and US, the transition model can be updated as follows:

$$\Delta P(s^{\mathrm{US}}|s^{\mathrm{CS}}) = \alpha(1 - P(s^{\mathrm{US}}|s^{\mathrm{CS}})) \tag{2.1}$$

$$\Delta P(s^{\overline{\mathrm{US}}}|s^{\mathrm{CS}}) = -\alpha P(s^{\overline{\mathrm{US}}}|s^{\mathrm{CS}}) \tag{2.2}$$

where α is the learning rate, and for example it can depend on the salience of the CS ($0 < \alpha < 1$)[2]. $s^{\overline{\mathrm{US}}}$ is any state other than s^{US}. The above equation intuitively implies that with each paring of CS and US, the predicted probability of the US after the CS increases, and the probability of no US after the CS decreases, and the rate of these increases and decreases depends on α.

The above conception of model-based Pavlovian conditioning is simplistic, and indeed, recent theoretical work indicates that the structure of the Pavlovian learning can be more complicated than learning the transition model, and can involve learning the *latent causes* in the environment (Courville, Daw, & Touretzky, 2006; Gershman, Blei, & Niv, 2010; Gershman & Niv, 2012). Within these accounts, instead of predicting the probability of the CS given the US ($P(s^{\mathrm{US}}|s^{\mathrm{CS}})$), the learning involves predicting both the CS and US given a latent cause (LC) in the environment ($P(s^{\mathrm{US}}, s^{\mathrm{CS}}|LC)$). Pavlovian conditioning in this conception involves inferring the likelihood that a latent cause is present and, based on that, predicting the probability that the US will happen.

One line of behavioral evidence supporting model-based Pavlovian conditioning comes from *post-conditioning stimulus revaluation* experiments. In such experiments, after an initial Pavlovian conditioning phase and the establishment of CRs, the ability of the US to induce

[1] It should be noted that the Rescorla-Wagner model is relevant only to the conditions in which US and CS occur at the same time (see (Niv & Schoenbaum, 2008) for a discussion). However, here this learning rule is applied in conditions that US and CS do not occur at the same time, in order to draw a parallel between this learning rule and model-based Pavlovian conditioning.

[2] Note that α is usually multiplied by another factor denoted by β, which is the associative value of the US ($0 < \beta < 1$)). Here for simplicity this factor is not included.

Table 2.1 – Post-conditioning stimulus revaluation. The experiment consists of three phases. The first is the establishment of CRs. In the second phase, for one group of the subjects, the ability of the US to induce CRs is manipulated (revalued). In the third phase (test), the ability of the CS to trigger CRs is measured.

Group	Phase 1	Phase 2	Test
Revalued	CS → US	US revaluation	CS?
Control	CS → US	-	CS?

unconditioned responses is manipulated (decreased or increased). Then, in the test phase, the ability of the CS to induce CRs is measured in extinction (Table 2.1). Evidence from rats (Rescorla, 1973; P. Holland & Rescorla, 1975; Cleland & Davey, 1982), humans (White & Davey, 1989; Davey & McKenna, 1983), and non-human primates (Gallagher, McMahan, & Schoenbaum, 1999; Izquierdo, Suda, & Murray, 2004) generally indicates that the stimulus revaluation procedure changes the ability of the CS to induce CRs, a finding that indicates animals have access to the sensory features of the US during the presentation of the CS, and thus they have learned the transition model of the environment during the conditioning phase.

In one example, Rescorla trained rats in a Pavlovian conditioning experiment in which a CS was paired with a loud noise (US) (Rescorla, 1973). After this training, animals showed suppression of lever presses during the CS. In the next phase of the experiment, a group of animals were habituated to the loud noise by the repeated presentation of the US, a treatment which decreases the effectiveness of the loud noise to induce unconditioned responses. In the third phase of the experiment, animals were tested in the presentation of the CS. Results indicated that the animals that had been habituated to the US, showed less suppression of lever presses in comparison to the control group, which were not habituated to the loud noise.

In another experiment (Gallagher et al., 1999), animals were first trained to associate the presentation of a CS with the delivery of food in a food cup. During the training animals started to approach the food cup during the CS, which is a type of CR in anticipation of the delivery of the US. In the devaluation phase the US was paired with illness by injection of LiCl after the consumption of the US in home cages. In the test phase the animals for which the

Table 2.2 – Design of the blocking and identity unblocking experiments. Please see the text for a description.

Group	Phase 1	Phase 2	Test
blocking	CS1 → US	CS1 + CS2 → US	CS2 → CR-
identity unblocking	CS1 → US1	CS1 + CS2 → US2	CS2 → CR+

US was devalued showed less approaches to the food cup during the CS, than the animals for which the US was not devalued.

Another line of evidence in support of the model-based Pavlovian conditioning comes from *identity unblocking* experiments (Table 2.2). The blocking paradigm has three phases: in the first phase, a CS is associated with a US (CS1 → US). In the second phase, a compound of the previous CS with a new CS (CS2) is associated with the US. Finally, in the test phase, the ability of CS2 to induce a CR is measured. The results generally indicate that CS1 blocks CS2 from being associated with the US, and during the test, CS2 is less able to induce a CR (Kamin, 1969). The general explanation for why blocking happens is, after the first phase, CS1 is a good predictor of the US, and thus CS2 does not become associated with the US.

The blocking effect can be *unblocked* if, in the second phase of a blocking experiment, the *identity* of the US changes from that in phase 1. In identity unblocking, in the second phase, the US is replaced by another US with different sensory properties, but the same reward component. For example the US in the first phase can be food pellets with a certain flavor, and in the second phase it can be the same food pellets with a different flavor. In this condition, CS1 is still able to predict the reward component of the US, but not the sensory features of it. Results of such experiments generally indicate that the CS2 triggers CRs during the test (McDannald, Lucantonio, Burke, Niv, & Schoenbaum, 2011; P. C. Holland, 1984), which means that the target of the learning process in the second phase is the prediction of the sensory features of the US, consistent with model-based Pavlovian conditioning.

The identity unblocking effect is not always observed in experiments (e.g., (Ganesan & Pearce, 1988)). In fact, under some experimental conditions, blocking occurs even when the identity of the outcome in the second phase is different from the one in the first phase. Such

experimental conditions are not yet systematically investigated, however, the fact that under some conditions the blocking effect is not observed, suggests that the Pavlovian conditioning process is derived by multiple learning processes, including model-based learning. One of such processes is model-free Pavlovian conditioning, which will be discussed in the next section.

2.1.3 Model-free Pavlovian conditioning

In model-free Pavlovian conditioning, an individual learns the reward component of the upcoming state, without necessarily learning the sensory features of that state. For example, an animal can learn that it will receive one unit of reward shortly after the CS, while not being able to predict the identity of the upcoming state (in contrast to the model-based Pavlovian conditioning). Within the Markov Reward Process context, instead of learning $P(s^{\mathrm{US}}|s^{\mathrm{CS}})$ (predicting the sensory features of the US during the CS), animals predict the amount of reward that they will receive from CS onwards. Take for example a condition in which a CS is followed by a US1 which constitutes one unit of reward, and after that there will be another US (US2) which constitutes two units of reward. In this condition, the total amount of reward that will be earned since the CS presentation is the sum of the rewards earned during the CS, US1, and US2, which is called the *value* of the CS (Figure 2.2). The value of the CS, therefore, equals 3 in this example and, in general, the total amount of reward that can be earned in a state, and its future states is called the *value* of that state, denoted by $V(s)$:

$$V(s) = E\left[\sum_{\tau=0}^{T-1} R(s_\tau)|s_0 = s\right] \tag{2.3}$$

where E is expectation over future states, and T is the duration of the experiment. The goal of model-free Pavlovian conditioning is then to directly learn the amount of reward that can be earned in the future states. Note that such values can also be calculated by predicting the future states using model-based Pavlovian conditioning. However, the goal of model-free Pavlovian conditioning is to predict such values without requiring the prediction of future states. The way these values are learned is described in the rest of this section; before that we

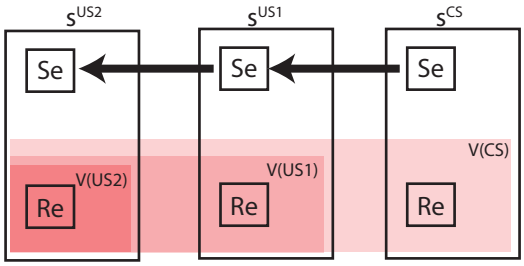


Figure 2.2 – The structure of the environment is such that a CS is followed by US1, and US1 is followed by US2. Each stimuli is composed of sensory features (Se), and a reward component (Re). The model-based Pavlovian conditioning is learning to predict the sensory information of the next state (arrows shown in the figure). On the other hand, model-free Pavlovian conditioning is to learn the value of each state, which is the total amount of reward that will be earned during or after that state. In this illustration, the value of the CS, denoted by $V(CS)$, includes the reward earned during the CS, plus the rewards earned during US1 and US2. The value of US1 is the reward earned during US1 and US2. Finally, the value of US2 is the reward earned during US2.

will review some evidence in support of such form of Pavlovain conditioning.

One line of evidence supporting model-free Pavlovian conditioning comes from *quantity unblocking experiments* (Rescorla, 1999; Blaisdell, Denniston, & Miller, 1997; McDannald et al., 2011). In these experiments (Table 2.3), in the first phase, a CS is being associated with a US. In the second phase, the reward component of the US is manipulated by changing the quantity of the US. For example, instead of earning one food pellet, the subjects earn two food pellets at each pairing of the CS and US (see also (P. C. Holland, 1988) for the effect of downshift in the magnitude of the US). This manipulation does not change the identity of the US, but it changes the reward component of the US. In this condition, since CS1 is able to predict the identity of the US in the second phase, model-based Pavlovian conditioning predicts that CS1 should be able to block CS2 from being associated with the US[3]. In contrast, evidence shows that changing the magnitude of the US unblocks CS2, which indicates that predicting the reward component of the US is also a target of the Pavlovian conditioning (see also aversive superconditioning (Dickinson, 1977) for a similar line of evidence).

[3]Note that in model-based Pavlovian conditioning, CS1 is associated with the identity of US, not amount of reward received during US. The amount of reward received during US will be associated with the sensory features of US (not CS).

Table 2.3 – Design of the blocking and quantity unblocking experiments. Please see the text for a description.

Group	Phase 1	Phase 2	Test
blocking	CS1 → US	CS1 + CS2 → US	CS2 → CR-
quantity unblocking	CS1 → US	CS1 + CS2 → 2 × US	CS2 → CR+

Equation 2.3 defines the value of a state as the total amount of reward that can be earned in the future, starting from that state. However, a remaining question that needs to be answered is how an agent learns this value. Here, neural and behavioral evidence shows that the value of a state is learned using the *temporal-difference learning* method (Sutton, 1988). According to this method, an agent starts learning by an initial guess about the value of each state. Then, during the learning process, as the agent receives rewards, it updates the values of states according to how much the value of each state was mis-predicted. Formally, let's assume that the agent is in state s, and receives reward r, and then transfers to state s'. Also, let's denote the current estimate of the value of state s with $\hat{V}(s)$, which means that the agent *expected* to earn $\hat{V}(s)$ amount of reward in the future, but, what it has actually earned in the current trial is r amount of reward, plus whatever it will earn in the next state ($\hat{V}(s')$), which will sum to $r + \hat{V}(s')$. This difference between what the agent expected to earn, and what it actually earned is the *reward prediction error*, denoted by δ:

$$\delta = r + \hat{V}(s') - \hat{V}(s) \tag{2.4}$$

The calculation of the above error signal is one form of the *bootstrap* method, since the agent is not sure about the value of the next state ($\hat{V}(s')$), but uses it anyway to calculate the prediction error of the current state, with the hope that after several learning trials, all the values gradually converge to their true values. The calculated reward prediction error is then used to adjust the value of the states:

$$\Delta\hat{V}(s) = \alpha\delta \tag{2.5}$$

where α is a learning rate. This update rule changes $\Delta\hat{V}(s)$ in the direction that decreases

the error signal, and eventually when the error signal becomes zero, the true value of $\Delta\hat{V}(s)$ is learned. One attractive feature of the above learning rule is the correspondence between the error signal (δ), and the activity of mid-brain dopamine neurons (Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997; Houk, Adams, & Barto, 1994). Indeed, there is extensive evidence in rodents and non-human primates that the activity of the mid-brain dopamine neurons encodes the prediction error (see (Clark, Hollon, & Phillips, 2012) for a review), and therefore, the model-free conception of the Pavlovian conditioning provides a neurally plausible way of learning the value of different stimuli. We will present more evidence in section 2.4.2, when explaining model-free instrumental learning.

2.2 Instrumental conditioning: the control problem

In the Pavlovian conditioning an individual learns to make predictions about the future states of the environment, however, it cannot control those future states. In contrast, in instrumental conditioning, an individual can influence future states of the environment by taking *actions*. The goal of taking actions is to earn the maximum amount of reward, and the minimum amount of punishment during the course of a task. This problem is known as reinforcement learning (RL) in the machine learning literature, and it refers to the problem of learning to choose a course of action so as to maximize a long-term objective. In a typical RL setting, an agent (or a controller) interacts with an environment (or a system) by executing actions in the environment and transitioning to a new state in the environment (Figure 2.3). Similar to Pavlovian conditioning, each state is composed of a sensory component and a reward component that the agent experiences by transitioning to that state.

2.2.1 Markov Decision Process

In section 2.1, we presented the Markov Reward Process as a way to formally represent Pavlovian conditioning. In the case of instrumental conditioning, we augment a Markov Reward Process with a set of actions, turning the Markov Reward Process into a Markov Decision

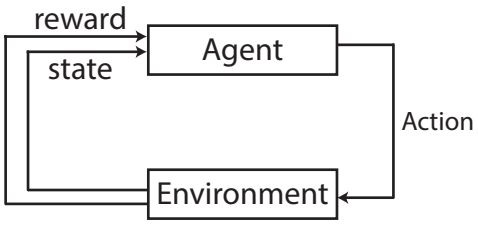


Figure 2.3 – A typical RL setting. An RL agent makes actions in the environment, receives a reward and transitions to the new state of the environment after executing each action.

Process (MDP). An MDP can formally be represented by a five tuple $(\mathscr{S}, \mathscr{A}, \{P(.|s,a)\}, R, T)$, where:

- A set of states $\mathscr{S}$, which is called the *state space*.
- A set of actions $\mathscr{A}$,
- The transition model $P(.|s,a)$ ($s \in \mathscr{S}$, $a \in \mathscr{A}$). $P(s'|s,a)$ represents the probability of transition to s' by taking action a in state s,
- The reward function R, which represents the reward associated with each state. $R(s)$ ($s \in \mathscr{S}$) is the reward received by transitioning to state s.
- T is the number of decisions that are to be made.

Take for example the situation in a Skinner box, in which there are two levers available (left lever and right lever), and pressing the left lever leads to the delivery of food pellets in the food magazine, and pressing the right lever leads to the delivery of a sucrose solution into the magazine. This environment consists of three states. The first state is the initial condition of the box in which the outcome has not yet been delivered (s^{i}). By pressing the left lever, the environment enters the second state in which a food pellet is delivered (s^{p}), and similarly by pressing the right lever, the environment enters the third state in which a sucrose pellet is delivered (s^{s}). An agent perceives entering into these states through the information that it receives via its sensors. If for example a rat is interacting with this environment, then entering into state s^{p} is signaled to the animal by the specific sensory properties of the food pellets that

the animal experiences during consumption and, similarly, transition to state s^{s} is signaled by the specific sensory properties of sucrose. The state that an action leads to is sometimes referred to as the *outcome* of that action. For example, in the above example, food pellets are the outcome of the left lever, and the sucrose solution is the outcome of the right lever. Within this example, there are two actions available, pressing the left lever (A_{l}) and pressing the right lever (A_{r}). Since by taking action A_{l} the environment transitions to state s^{p} we will have $P(s^{\mathrm{p}}|s^{\mathrm{i}}, A_{\mathrm{l}}) = 1$, and likewise, $P(s^{\mathrm{s}}|s^{\mathrm{i}}, A_{\mathrm{r}}) = 1$. As another example, assume a condition in which on average every four left lever presses lead to the delivery of a food pellet (random-ratio 4 schedule), which means that $P(s^{\mathrm{p}}|s^{\mathrm{i}}, A_{\mathrm{l}}) = \frac{1}{4}$ (these probabilities are sometimes called *action-outcome contingencies (A-O)* in psychological terms).

Finally, similar to the Markov Reward Processes, there is a reward associated with each state. In the above example, when the animal consumes food pellets, it assigns an incentive or a reward to the current state of the environment, s^{p} (this process is called *incentive learning* (Dickinson & Balleine, 2002)). This reward is denoted by $R(s^{\mathrm{p}})$ for the pellet, and $R(s^{\mathrm{s}})$ for the sucrose. For example, if a rat values food pellets twice as much as the sucrose solution, then we can assume $R(s^{\mathrm{p}}) = 1$, and $R(s^{\mathrm{s}}) = 0.5$. Based on these factors, the MDP describing the environment will be:

- $\mathscr{S} = \{s^{\mathrm{i}}, s^{\mathrm{p}}, s^{\mathrm{s}}\}$,
- $\mathrm{A} = \{A_{\mathrm{l}}, A_{\mathrm{r}}\}$,
- $P(.|s, a) = \{P(s^{\mathrm{p}}|s^{\mathrm{i}}, A_{\mathrm{l}}) = 1, P(s^{\mathrm{s}}|s^{\mathrm{i}}, A_{\mathrm{r}}) = 1\}$
- $R(s^{\mathrm{p}}) = 1$, $R(s^{\mathrm{s}}) = 0.25$
- $T = 10$

where we assumed that there are 10 trials in which the animal is allowed to make a decision ($T = 10$).

The action selection *policy* of an agent indicates which action will be taken in each state. The policy is denoted by $\pi(.|s), s \in \mathcal{S}$, and it is the probability of taking action a in state s. In the above example, if the agent always chooses the left lever, then we will have $\pi(A_l|s^i) = 1$, which means that the probability of taking action A_l in state s^i is one. The value of a policy, V^π, can be defined as the total reward that the agent earns if it follows that policy:

$$V^\pi(s) = E\left[\sum_{t=0}^{T-1} r(s_t)\right] \tag{2.6}$$

For instance, in the condition described above, if the agent follows a policy that indicates taking action A_l in all the trials, ($\pi(A_l|s^i) = 1$), then $V^\pi(s) = T$ since it will earn one amount of reward in each trial. If the agent takes an action randomly ($\pi(A_l|s^i) = 0.5$, $\pi(A_l|s^i) = 0.5$), then half of the time it receives $R(s^p)$, and the rest of the times it receives $R(s^s)$, and thus, the value of the policy will be $V^\pi(s^i) = 0.75T$. Among all the action selection policies that an agent can take, there is one policy which yields the highest amount of future rewards, which is called the optimal policy, and is denoted by π^*.

As mentioned earlier, the goal of a RL agent is to take the course of action that maximizes the reward earned, or in other words, the goal is to discover the optimal policy (or close to optimal). Such a policy is not known to the agent at the beginning, and it needs to be learned through the interaction of the agent with the environment, a process which is called instrumental conditioning. During instrumental conditioning, the agent builds a representation of the environment, and then uses that representation to decide which action should be taken at each state. Such a representation, however, is not unique, and there are multiple ways that the agent can represent the environment, and yet maintain a good action selection policy. In fact, evidence shows that depending on the internal and environmental conditions, individuals can switch between different representations. In particular, in some conditions instrumental actions exhibit the properties of *goal-directed actions*, and in other conditions they seem to be *automatic actions*. These types of actions (instrumental actions) are explained in detail in the following sections, and their behavioral and theoretical aspects are the main focus of the current thesis.

2.3 Instrumental conditioning: goal-directed actions

2.3.1 Behavioral properties

Goal-directed action selection can be defined as an action selection policy which is sensitive to the changes in the underlying MDP of the environment. In simple words, in goal-directed decision-making, an agent that previously took an action in order to attain a certain goal, will no longer take that action if the value of the goal decreases for it, or it no longer believes that the action will lead to the goal that it wanted to attain. Guided by this definition, a commonly used experimental method for determining whether action selection is goal-directed, is to manipulate the reward of the states of the environment, and then test whether it affects action selection. A typical experiment for assessing whether action selection is goal-directed is called *outcome revaluation* (Table 2.4). In such experiments, an agent is first trained to perform two different actions that earn different food outcomes (i.e, two different actions that lead to two different states). After this training, an *outcome revaluation* treatment is conducted, in which the value of one of the outcomes (i.e., one of the states) is manipulated, a treatment that generally involves sating the animals on one of the two outcomes to decrease its value. Subsequently, back *online*, a test is conducted in which the choice between the two actions is assessed in the absence of the outcome. Typically, when given this choice, humans and other animals decrease their performance of the action that previously delivered the now devalued outcome, which can be regarded as an operational definition that action selection has been goal-directed (Colwill & Rescorla, 1985).

Changes in the value of outcomes is usually conducted in an *offline* manner, in which the value of the outcome is manipulated without taking actions. In contrast, in online outcome devaluation, the value of an outcome is manipulated during the performance of instrumental actions. For example, assume that an animal has learned that food pellets can be earned by pressing the left lever in the operant chamber. In this condition, the offline outcome devaluation of food pellets can be achieved by feeding the animal with food pellets (without requiring it to take any instrumental action) until it gets sated. In contrast, online outcome

Table 2.4 – Design of revaluation experiments. During the training phase, subjects learn that the outcome of one of the actions (A1) is O1, and the outcome of the other action is O2. Then, the value of one of the outcomes is increased (inflation), or decreased (outcome devaluation). Finally, during the test phase, animals are given a choice between A1 and A2, in order to probe whether changes in the value of the outcomes affect their action selection.

Experiment	Training	Devaluation/Inflation	Test
Devaluation	A1→O1 A2→O2	O1↓	A1 vs A2
Inflation	A1→O1 A2→O2	O1↑	A1 vs A2

devaluation of food pellets can be achieved by allowing animals to earn food pellets in the training chamber by taking the instrumental action (pressing the left lever in this example). The criteria for action selection to be goal-directed includes both online and offline outcome devaluation. In practice, however, only an offline outcome devaluation test is conducted, which is presumably a more stringent test than online outcome devaluation. It should also be noted that, the test of outcome devaluation experiments is usually conducted in extinction conditions in which the agent does not receive outcomes as a result of taking instrumental actions. The benefit of testing in extinction conditions is that action selection only reflects what the agent has learned prior to the test, and is not confounded by online learning during the test.

The two common ways of manipulating outcome values are *specific satiety*, and *conditioned taste aversion*. In specific satiety, animals are pre-fed with the outcome which we want to devalue. For example, if the outcome of one of the actions is a food pellet, and the outcome of the other action is a sucrose pellet, then changing the value of the sucrose pellet is achieved by pre-feeding animals with sucrose pellets for one hour before the test. An alternative way to change the value of an outcome is conditioning a taste aversion to the outcome. Animals readily associate gastric malaise with specific foods and tastes (Garcia, Kimeldorf, & Koelling, 1955). Lithium chloride (LiCl) induces a gastric malaise in rats when injected intraperitoneally and, by pairing the consumption of the outcome with injections of LiCl, animals attribute the illness to the outcome they have just consumed, and in this way, the value of the outcome decreases.

2.3.2 Computational models: model-based RL

The observation that an offline change in the value of a state influences action selection, reveals two aspects of goal-directed learning. Firstly, it shows that the learning algorithm builds a representation of the transition model of the environment, or in other words, it has learned action-outcome contingencies; otherwise, it is not possible for an agent to know which action's value will be affected by the change in the value of a state (as it has not learned the links between actions and states). Secondly, it shows that in addition to the transition model of the environment, actions are also guided by the reward component of states, otherwise changing the reward component of the states by devaluation would not affect actions. Action outcome contingencies (or the transition model of the environment in RL terms), and the reward component of different outcomes are referred to as a *model of the environment*, and a RL algorithm that builds a representation of the model of the environment is called *model-based RL*.Within this framework, the explanation of a typical outcome devaluation experiment is as follows: during the initial training, the agent learns the value of different states, as well as the transition model of the environment. After the initial learning, during the outcome revaluation phase, the value of one of the states updates, which leads to a change in the reward function. Finally, during the test, the values of actions are calculated using the new reward function, which will cause a decrease in the performance of the action that has a devalued outcome. As such, model-based RL can produce behavior that is similar to the results of the outcome revaluation experiments, and it is suggested, therefore, to be a model of goal-directed learning (Daw et al., 2005; Keramati, Dezfouli, & Piray, 2011).

The next issue is how an agent learns the model of the environment. As mentioned earlier, the model of the environment has two components: the transition model of the environment, and the reward of each state. Let's assume an agent is in state s, and performs action a and enters a new state, s'. Given these experiences, the estimated probability of reaching s' given s and a increases, for example using the following rule:

$$\Delta P(s'|s,a) = \alpha(1 - P(s'|s,a)) \tag{2.7}$$

where α is a learning rate. Similarly, the estimated probability of reaching all the states other than s' (denoted by $\bar{s}$) decreases:

$$\Delta P(\bar{s}|s,a) = -\alpha P(\bar{s}|s,a) \tag{2.8}$$

where α is a learning rate. In this way, after each experience of a transition to a new state the agent updates its estimates of the transition model of the environment. At the same time, by entering a new state, the agent also receives a reward by which the reward function will be updated. For example, if the agent receives reward r by entering into state s, the reward of state s updates as follows:

$$\Delta \hat{R}(s) = \alpha(\hat{R}(s) - r) \tag{2.9}$$

where α is a learning rate. By having an estimate of the model of the environment, an agent can calculate the value of each action. To see how this can be done, let's denote the value of action a in state s as $Q(s,a)$. The value of such a state-action pair is the reward received in state s ($R(s)$), plus the value of each new state that the action might lead to ($V(s')$) weighted by the probability of reaching each new state ($P(s'|s,a)$):

$$Q(s,a) = R(s) + \sum_{s'} P(s'|s,a)V(s') \tag{2.10}$$

the above equation defines the value of each state-action pair using the values of states ($V(s')$). Under the optimal policy, we can assume that in each state the agent will take the action that has the highest value, and thus the value of a state will be equal to the value of the highest state-action in that state:

$$V(s) = \max_a Q(s,a) \tag{2.11}$$

where $V(.)$ and $Q(.,.)$ refer to the values of states, and state-action pairs respectively.

Equations 2.10 and 2.11 are called *Bellman optimality equations*, which can be solved using

different methods (Puterman, 1994). One such method is the *tree search* method in which, starting from a state, the agent unfolds each node of the tree using equation 2.10, and continues this process until it reaches a *terminal state*, i.e., a state which does not lead to any other state (or an *absorbing state*). There are methods other than the tree search to solve Bellman equations such as *value iteration* and *policy iteration* (e.g., (Puterman, 1994)). However, such methods often suffer from the *curse of dimensionality* (Bellman, 1961): the complexity of the state-space of the problem (and therefore finding the course of action that maximize the total reward) often increases exponentially with the number of features or dimensions that define the problem. For example, in the case of the tree search, for each step further that the agent desires to predict the consequence of its actions, the number of nodes of the tree that need to be unfolded increases exponentially. In the next section, we look more closely at this issue.

2.3.3 Computational models: complexity measures

The computational cost of finding the (near) optimal solution of an MDP is often called *computational complexity*, and as we mentioned before the model-based approaches suffer from high computational complexity (see (Littman, 1996) for a survey). For example, in a standard model-based RL algorithm, known as $R_{\texttt{max}}$, the computational complexity for planning the next action is $\Omega(|S|^2|A|)$ where $|S|$ is the size of the state space, and $|A|$ is the size of the action space. As we will discuss in the next section, there are other algorithms which are more efficient than model-based RL in terms of computational complexity.

Other than computational complexity, there are two more aspects that should be considered when evaluating decision-making algorithms: *sample complexity* (Kakade, 2003), and *space complexity*. Sample complexity, roughly means how much data an agent needs to gather in order to make optimal decisions. As such, the sample complexity of an algorithm is basically a reflection of how the algorithm trades-off between exploration and exploitation strategies. The space complexity refers to the amount of space or memory an algorithm needs in the worst case scenario. For example in a standard model-based algorithm, called $R_{\texttt{max}}$ (Brafman & Tennenholtz, 2003), the space complexity is $\Theta(|S|^2|A|)$, which is intuitive since the probability

of transition from each state to other states by each action should be stored by the algorithm.

There are algorithms other than model-based RL for solving an MDP, which have different complexity properties. For example, a class of algorithms called *model-free RL* have lower computational and space complexity than model-based RL, but can have higher sample complexity than model-based RL. As such, it is not surprising that the brain employs different algorithms depending on its available memory, computational resources, and environmental conditions. In fact, in the next section, we provide behavioral evidence that the decision-making process in the brain deviates from simple model-based RL. In particular, we review behavioral evidence that support the brain using model-free RL (section 2.4.1), and *hierarchical RL* (section 2.4.4). Decision-making in such processes can be seen as a more *automated* form of action selection, since such decision-making processes usually have lower computational or space costs compared to model-based RL. As such we call such processes *automatic processes*, which is the topic of the next sections.

Before getting on to the properties of automatic actions, it should be noted that automatic actions are not the only method that the brain utilizes to overcome the curse of dimensionality. Some theories suggest that goal-directed decision-making collaborates with the Pavlovian process for action control. Imagine that an agent desires to perform a tree search to make a goal-directed action, but the depth of the tree is beyond the available computational resources of the agent. In such a situation the agent can use some heuristics to approximate action values. One psychologically plausible way to perform such heuristics is to use model-free Pavlovian values of the states instead of calculating the value of each action in that state (see (Huys et al., 2012) for a related approach). For example, a person might deliberate over the available paths (actions) that can be taken to reach home, but will not calculate the values of the actions that might be taken (expanding the tree) after reaching home, and will instead use a model-free Pavlovian value that has previously been assigned to the state of being at home.

Another way of restricting the size of the forward search is searching the tree bottom-up instead of top-down. This means that an agent first determines which outcome is desired, and

Table 2.5 – Design a typical specific Pavlovian-instrumental transfer experiment. Please see the text for description.

Pavlovian phase	instrumental phase	Test
S1→O1	A1→O1	S1: A1 vs A2
S2→O2	A2→O2	S2: A1 vs A2

then calculates which course of action leads to that outcome. While in the top-down approach, the agent first considers which actions are available, and calculates all the potential outcomes of the available actions[4]. A manifestation of such a process (bottom-up) can probably be seen in a phenomenon known as *specific Pavlovian-instrumental transfer (sPIT)*. A typical sPIT experiment has three phases (Table 2.5). In the first phase (Pavlovian phase), two neutral stimuli (S1 and S2) are being paired with two distinguishable outcomes (S1-O1, S2-O2). In the second phase (instrumental phase), the outcomes are paired with two different actions (A1-O1 and A2-O2). Finally, in a test phase, animals are given a choice between A1 and A2 in the presence of S1 or S2. Results generally show that in the presence of S1, animals perform the action with similar outcome to that paired with S1 (action A1), and similarly, in the presence of S2, animals perform action A2 (e.g., (Trapold & Overmier, 1972; Baxter & Zamble, 1982; Kruse, Overmier, Konz, & Rokke, 1983; Colwill & Motzkin, 1994; Corbit, Muir, & Balleine, 2001)). One interpretation of such results is that, the presence of say S1, primes the representation of O1, and then the subject searches for the action that leads to O1, i.e., A1, which biases the action selection toward the choice of action that has a consistent outcome with the presented stimuli (see (Balleine & O'Doherty, 2010)). Such kinds of interactions between instrumental and Pavlovian systems are beyond the scope of the current thesis, and we will not discuss them further.

2.4 Instrumental conditioning: automatic actions

In certain situations, actions made by an individual are not consistent with the goal-directed schema explained in the previous section. Such deviations can generally be categorized into

[4] Note that the relative efficiency of the bottom-up search to the top-down search depends on the structure of the decision-making tree, e.g., number of available actions and available outcomes, and it is not always necessarily the case that the bottom-up search is better than the top-down search.

two forms of behavior: (i) insensitivity to outcome devaluation and contingency degradation, and (ii) performance of action sequences. In the following sections, we will explain each form in turn, and then we will explore the theoretical aspect of each form.

2.4.1 Behavioral properties: outcome devaluation

As mentioned in the previous section, goal-directed actions are sensitive to changes in the value of their outcomes. However, evidence shows that *overtraining* changes this property of actions and makes them: (i) insensitive to changes in the value of the outcome (C. D. Adams, 1982; Dickinson, Balleine, Watt, Gonzalez, & Boakes, 1995; P. C. Holland, 2004; Killcross & Coutureau, 2003; Yin et al., 2004; Lingawi & Balleine, 2012; Gremel & Costa, 2013); and (ii) insensitive to changes in the causal relationship between the action and outcome delivery (Dickinson et al., 1998). Here we describe, for illustration, data from a recent study in which we were able to observe both effects in the same animals, in the same experiment, comparing moderately trained and overtrained actions for their sensitivity to devaluation, induced by outcome-specific satiety, and contingency degradation, induced by the imposition of an omission schedule in rats. The data is presented in Figure 2.4. Rats were trained to press a lever for sucrose and, after the satiety treatment, given a devaluation test (Figure 2.4A). After retraining, they were given the contingency degradation test (Figure 2.4B). In the first test, moderately trained rats showed a reinforcer devaluation effect; those sated on the sucrose outcome reduced performance on the lever compared to those sated on another food. In contrast, groups of overtrained rats did not differ in the test. In the second test, rats exposed to the omission contingency were able to suppress the previously trained lever press response to get sucrose, but only if moderately trained. The performance of overtrained rats did not differ from the yoked, non-contingent controls.

There is an important difference between the experiment described above, and the experiments described in the previous section for the test of goal-directed actions. In the above experiment, there is only one instrumental action available, however, in the experiments explained in the previous section, there are two actions, and two outcomes available in the

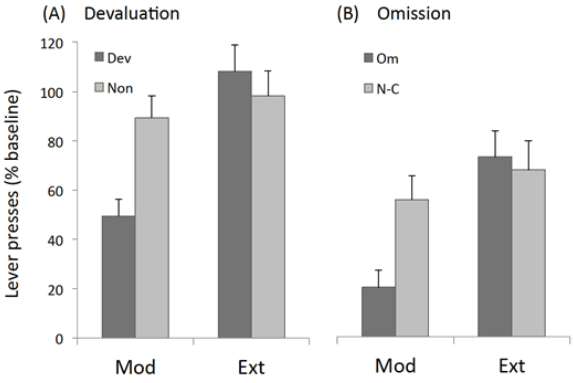


Figure 2.4 – Four groups of rats (n=8) were trained to lever press for a 20% sucrose solution on random-interval schedules (RI1, RI15, RI30, RI60) with moderately trained rats allowed to earn 120 sucrose deliveries and overtrained rats 360 sugar deliveries (the latter involving an additional 8 sessions of RI60 training with 30 sucrose deliveries per session). (A) For the devaluation assessment, half of each group was then sated either on the sucrose or on their maintenance chow before a 5-min extinction test was conducted on the levers. As shown in the figure (panel A), moderately trained rats showed an reinforcer devaluation effect; those sated on the sucrose outcome reduced performance on the lever relative to those sated on the chow. In contrast, groups of overtrained rats did not differ in the test. Statistically we found a training x devaluation interaction, $F(1,28)=7.13$, $p<0.05$, and a significant devaluation effect in the moderately trained, $F(1,28)=9.1$, $p<0.05$ but not in the overtrained condition ($F<1$). (B) For the contingency assessment, after devaluation all rats received a single session of retraining for 30 sucrose deliveries before the moderately trained and overtrained rats were randomly assigned to either an omission group or a yoked, non-contingent control group. During the contingency test the sucrose outcome was no longer delivered contingent on lever pressing and was instead delivered on a fixed time 10 sec schedule. For rats in the omission groups, responses on the lever delayed the sucrose delivery by 10 sec. Rats in the yoked groups received the sucrose at the same time as the omission group except there was no response contingency in place. As is clear from the figure (panel B), rats exposed to the omission contingency in the moderately trained group suppressed lever press performance relative to the non-contingent control whereas those in the overtrained groups did not. Statistically, there was a training x degradation interaction, $F(1,28)=5.1$, $p<0.05$, and a significant degradation effect in the moderately trained, $F(1,28)=7.8$, $p<0.05$, but not in the overtrained condition ($F<1$).

environment. In fact, evidence indicates that when animals have a choice between two instrumental actions, their choices remain sensitive to outcome devaluation, even after extended training (Kosaki & Dickinson, 2010), and therefore, the training conditions for the study of automatic actions usually consist of only one instrumental action.

In fact, even when only one instrumental action is available, other factors promote the formation of automatic actions. In particular, random interval schedules of reinforcement (RI), enhance the formation of automatic actions, in comparison to ratio schedules of reinforcement or fixed-interval schedules (Dickinson, Nicholas, & Adams, 1983; Gremel & Costa, 2013; Derusso et al., 2010) (see also (Derusso et al., 2010)).

2.4.2 Computational models: model-free RL

As mentioned in the previous sections, the learning and expression of goal-directed actions is guided by what has been learnt about the model of the environment. The observation that over-training renders such goal-directed actions insensitive to outcome revaluation, reveals that the selection of automatic actions cannot be guided by the model of the environment. Associative learning theories attribute such actions to *stimulus-response (S-R)* learning, which means than an agent learns to elicit a response (R) when faced with a certain stimulus (S). In essence, S-R theories of instrumental conditioning argue that whenever the performance of a response is followed by positive *reinforcement*, then the strength of the association between the response and the antecedent stimulus increases. Conversely, negative reinforcement after a response decreases the strength of the S-R association (Thorndike, 1911).

Within S-R theories, reinforcement can be regarded as the reward component of the state that the response (or action) leads to. As such, the strength of an S-R association will be proportional to the reward that an action entails in the future: actions that have led to higher rewards in the past are more likely to be taken in similar situations (or similar states in RL terms) in the future. In simple words, this form of learning implies that whenever an action leads to reward, the agent marks that action as a *good* action, and then in the future, the agent

takes the actions that have been marked as good actions, without necessarily knowing what the exact consequence of those actions will be.

As such, similar to goal-directed instrumental actions, the agent will take the actions with the highest values. However, the difference is that the values of actions are updated at the time of experiencing the reward, while in the goal-directed case, the values of actions are calculated at the choice points. In other words, during the formation of S-R associations, by experiencing a reward after taking an action, the value of that action updates and becomes *cached* for future use, while in the goal-directed case, values are calculated at the choice points. Since values of actions are cached, they remain unchanged after off-line devaluation of outcomes, and hence can be regarded as computational basis for the insensitivity of overtrained actions to outcome devaluation (Daw et al., 2005).

One way of learning the cached values of actions is through *model-free RL*. The goal of model-free RL is to learn the value of each action in each state, without learning the model of the environment. Following (Watkins, 1989), we denote the value of action a in state s with $Q(s,a)$. Now, assume that an agent is in stage s, and performs action a and enters a new state s', and receives reward r. The amount of mis-prediction in the value of the action generates a reward prediction error:

$$\delta = r + V(s') - Q(s,a) \tag{2.12}$$

where $V(s')$ is the value of state s', which is equal to the value of the best action in state s'. Then, using this error signal, the value of the action adjusts as follows:

$$\Delta Q(s,a) \leftarrow \alpha\delta \tag{2.13}$$

The above form of learning is called $Q-$learning (Watkins, 1989). There is another variant of model-free RL, called SARSA (Sutton & Barto, 1998), in which the error signal is calculated based on the value of the actual action that will be taken in the state s', i.e., the error signal

will be as follows:

$$\delta = r + Q(s', a') - Q(s, a) \tag{2.14}$$

where a' is the action that will be taken in state s'. There is evidence in support of both forms of learning in the brain (Morris, Nevet, Arkadir, Vaadia, & Bergman, 2006; Roesch, Calu, & Schoenbaum, 2007).

As it is clear from the above equations, the learning process is similar to the model-free Pavlovian conditioning, but the difference is that here, the values of *actions* are being learned, while in the Pavlovian case the agent learns the value of *states*. Indeed, it is suggested that the prediction error which is used in equation 2.13 is generated by the Pavlovian system. That is, the prediction error that is generated by the Pavlovian model-free system is also sent out to the instrumental model-free system for the purpose of learning the state-action values (Barto, 1995; Joel, Niv, & Ruppin, 2002). This form of learning is called *actor-critic*, in which an actor selects and learns the value of actions, and a critic provides the error signal that adjusts action values.

The information that reward prediction conveys is only regarding the reward component of the outcome of actions and, thus, the identity of the consequence of actions will not be represented during the learning process. The advantage of such a schema is that at the choice points values are already calculated, and there is no need to deliberate over the consequence of actions (as in the goal-directed process). However, choices will be insensitive to offline changes in outcome values, since values are cached and updated only after the online experience of actions. Due to this property of model-free RL, previous works have suggested that it is a computational substrate for the insensitivity of overtrained actions to outcome devaluation (Daw et al., 2005).

Evidence for model-free RL in the brain- As we will show in chapter 3, the model-free account is

not the only way to explain insensitivity to outcome devaluation, therefore, further behavioral or neural evidence is required to confirm the operation of the model-free RL behind choices. One line of behavioral evidence can be provided based on the equation 2.13, which implies that the effect of past rewards on current choices decays exponentially (with the rate of α). This prediction has been confirmed in previous studies in humans, non-human primates, and rodents (Lau & Glimcher, 2005; Sugrue, Corrado, & Newsome, 2004; Ito & Doya, 2009). However, alternatively, it can be argued that a model-based learning algorithm can also predict the same pattern; indeed, in equations 2.7 and 2.8, the effect of past experiences on the current contingencies decreases exponentially, and since values are calculated using action-outcome contingencies, the effect of past experiences on the current action values, and choices decays exponentially, similar to the pattern predicted by model-free RL. Therefore, this behavioral evidence alone is not sufficient to conclude that performance reflects the operation of model-free instrumental conditioning.

Another line of evidence in support of model-free instrumental learning comes from the connection between dopamine and the reward prediction error (equation 2.12). In fact, similar to the Pavlovian conditioning, it is suggested that the phasic activity of dopamine neurons code the prediction error, and serve as the teaching signal for the value of *actions* (where in the Pavlovian case the signal was involved in learning the value of *states*). Causal evidence in support of this hypothesis comes from studies showing that Parkinson patients, who suffer from the degeneracy of dopamine neurons, are impaired in making optimal choices, which can be recovered by the administration of L-DOPA (which is a dopamine agonist) (Frank, Seeberger, & O'reilly, 2004; Bódi et al., 2009; Cools, Altamirano, & D'Esposito, 2006; Rutledge et al., 2009). One notable feature of these experiments is that, subjects were presented with a set of stimuli on the screen, and they were instructed to choose one of them, for example, by pressing the corresponding button. Although this process can be interpreted as if subjects are learning the value of actions, it is also consistent with the subjects learning the Pavlovian value of states (stimuli). That is, subjects learn the Pavlovian value of the stimuli, and then they evaluate actions in a model-based manner, and take the actions which correspond to

the states with high values. As such, these experiments do not provide direct evidence for the involvement of dopamine in learning the value of actions.

There is similar evidence from optogenetic studies in mice showing that, animals return to the location in which they received stimulation of dopamine neurons in VTA (Tsai et al., 2009), avoid locations in which the activity of dopamine neurons in ventral tegmental area (VTA) was inhibited (Tan et al., 2012), and learn to nose-poke when nose-poking is followed by the stimulation of dopamine neurons (K. M. Kim et al., 2012). In such experiments, the responses that the animal presents can be seen as forms of the Pavlovian condition response (CR) (e.g., approaching the locations with high value), and thus the learning process can be interpreted in terms of Pavlovian conditioning. There is another study in which animals needed to perform an instrumental action in order to receive the stimulation (Adamantidis et al., 2011). This study showed that animals preferred to take the action which delivered both food and stimulation, over the action that only delivered food. However, similar to the experiments in Parkinson's patients, it can be argued that here the stimulation increases the value of the 'food state' through a Pavlovian learning mechanism, which implies a high value for the goal-directed action that led to the food (see (Domingos et al., 2011) for a similar experiment using water).

Thus, in summary, although there is behavioral and neural evidence for the involvement of model-free instrumental conditioning, such evidence can mainly be explained using a combination of model-free Pavlovian conditioning, and model-based instrumental conditioning process. The above argument, however, only implies that the mentioned effects *can* be explained using model-based instrumental conditioning, and since such experiments are not accompanied by tests such as outcome devaluation it cannot be claimed that the underlying process is necessarily model-based RL.

2.4.3 Computational models: arbitration rules

In section 2.3.2 we argued that model-based RL provides a plausible computational substrate for goal-directed actions. Also, as argued in section 2.4.2, previous authors suggest that

model-free RL exhibits properties that can explain insensitivity to outcome devaluation. Now, a remaining question to be answered is why decision-making is model-based early in training, whereas later in training it becomes model-free (since actions will be insensitive to outcome devaluation after overtraining). Here, it is suggested that a third system, the *arbitrator*, coordinates the contribution of the systems to choice. That is, at each choice point, the arbitrator selects one of the systems, or a combination of both systems, to control actions. Several models have been previously suggested for how the arbitrator works (Daw et al., 2005; Keramati et al., 2011; Pezzulo, Rigoli, & Chersi, 2013; Lee, Shimojo, & O'Doherty, 2014), that we illustrate them in this section.

Within a normative perspective, the logic of the arbitrator, i.e, shifting from model-based to model-free, should stem from the algorithmic properties of each process. In section 2.3.3 we mentioned that model-based RL is generally expensive in terms of computational and space complexity, while model-free RL has a lower space complexity, which is $O(|S||A|)$ compared to $O(|S|^2|A|)$ in model-based RL. This means roughly that the operation of the model-free system requires less memory than model-based RL. An implication of this observation is that, when the agent has memory constraints, it would be more beneficial to rely on the model-free system (Otto, Gershman, Markman, & Daw, 2013; Otto, Raio, Chiang, Phelps, & Daw, 2013). Similarly, the computational complexity of model-free learning is $O(\ln|A|)$ (if Q-values are presented by a heap data structure), compared to model-based RL which has a higher computational complexity of $\Omega(|S|^2|A|)$. As such, the model-free approach will be more appropriate when computational resources are limited, or decisions need to be made under temporal constraints.

But neither space nor computational complexity are directly related to the shift to model-free RL after extended training, since memory and temporal constraints are not manipulated during the course of training. However, what might be underlying the shift to model-free RL is sample complexity: model-based RL generally requires less amount of data to make optimal choices in comparison to model-free RL (see (Li, 2009, p.82) for a discussion). As such, early in training when data is limited, the agent uses model-based RL, while later in training when

enough data exists, the agent uses model-free RL which is more efficient in terms of space and computational complexity. Inspired by this argument, previous works have suggested several arbitration rules for how and why the brain switches between model-free and model-based decision-making modalities. Such arbitration rules can be divided into two classes, which we will discuss in the following section.

Class I arbitration rules- In this class of arbitration rules (Keramati et al., 2011) (see also (Pezzulo et al., 2013)), it is assumed that the choices made by the goal-directed process are always better than the choices of the model-free system, i.e., the model-based process has the *perfect information* about the environment, which is consistent with the assumption that the model-based system has lower sample complexity. On the other hand, it is assumed that the model-based system takes longer to calculate which action should be taken, which is again consistent with the fact that model-based RL has a higher computational complexity. Based on this, the arbitration rule becomes as follows: if the *value of perfect information (VPI)* provided by the model-based system exceeds the opportunity cost of waiting for the time consuming model-based approach, then the model-based system will control actions, otherwise, the model-free system will control actions. The value of perfect information is equivalent to the extra amount of reward that can be gained if the model-free system knows the exact values of actions (by querying the model-based process). The opportunity cost of the slowness in the model-based system can be quantified using the concept of *average reward rate ($\bar{R}$)* (Niv, 2007). The average reward rate represents the amount of reward that the agent can gain in a unit of time, and thus the opportunity time of waiting for model-based calculations will be $\bar{R}\tau$, where τ is the time taken by the model-based process.

Within the above schema, early in training when the estimations of the model-free system are inaccurate, the value of perfect information is high, and thus control will be goal-directed. Later in training when the estimations of the model-free system become rather accurate, the value of perfect information drops, and the model-free system dominates behavior. An important property of this class is that, the information that the arbitrator relies on for choosing a system to control actions, is solely based on the estimation of the model-free system, and the

model-based system is called only when it is required. This property allows an agent to exhibit fast reaction times when relying on the model-free system.

Class II arbitration rules- This class of arbitration rules (Daw et al., 2005; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Gläscher, Daw, Dayan, & O'Doherty, 2010; Lee et al., 2014) assumes that there are conditions under which the values estimated by the goal-directed process, i.e., model-based system, *can be* less accurate than the values estimated by model-free RL. For example, if the available memory is limited the goal-directed process might sacrifice the accuracy of the value computation for fitting the required computation in the available memory. The model-free process, however, suffers less from such memory limitations, and after sufficient training it eventually surpasses the model-based accuracy. This is in contrast to the first class of arbitration rules, which assumed that the estimated values of the goal-directed process are always more accurate than those of model-free RL. As such, in this class of arbitration rules, the arbitrator receives inputs from both model-free and model-based RL, and these inputs convey information about the accuracy of the estimates of each system. Then, the arbitration allows each system to exert control over the actions proportional to the certainty that they have about their decisions.

The major difference between this class of arbitration rules and the first class is that, in this class, the goal-directed process is always engaged in decision-making, while in the first class, the goal-directed process will only be engaged when required. In other words, in this class of arbitration rules, even when the overt behavior is model-free, the model-based system is still working behind the scene to feed its inputs into the arbitrator. (Keramati et al., 2011) showed that the way in which class I arbitration rules conceptualize the arbitration is more consistent with behavioral results.

2.4.4 Behavioral properties: action sequences

In the previous section we presented a property of automatic actions, which was their insensitivity to outcome devaluation. In this section, we present another property of automatic

actions, known as *action chunking*. Evidence shows that with practice, actions that are frequently executed together concatenate to form *action sequences* (also known as *action chunks*, *macro actions*, *skills*, *tasks*, *sub-tasks* or *schema*) which are then treated as a single response unit (Book, 1908; Miller, Galanter, & Pribram, 1960; Pew, 1966) (see (Rhodes, Bullock, Verwey, Averbeck, & Page, 2004) for a review). Within the domain of motor control, such action sequences can be a sequence of motor commands, as for example typing a new word for the first time is a disjoint set of key presses (actions), while practice renders the whole set of key presses as an integrated sequence of muscle contractions. More generally, actions within an action sequence can be at a higher level than motor commands. For example, within the area of executive functions, with practice, making a cup of tea becomes a sequence of actions consisting of 'boiling the water', 'adding sugar', etc.

The problem of learning action sequences was called *action syntax* by Karl Lashley (1890-1958), who is usually credited with pointing out the centrality of serial order learning in skilled actions (Lashley, 1951). Lashley proposed a new theoretical point of view that rejected the associative account of action sequence learning, called *response chain theory (or associative chain theory)* (James, 1890). The response chain is basically an extension of the stimulus-response learning theory (Guthrie, 1952; Hull, 1952), which postulates that within an action sequence, the stimulation caused by an action triggers the next action in the sequence:

$$\cdots \rightarrow R_i \rightarrow S_i \rightarrow R_{i+1} \rightarrow S_{i+1} \rightarrow \ldots$$

where action R_i causes stimulus S_i, which then triggers R_{i+1}. Take for example typing the word "are". The response chain theory suggests that seeing the word "are", triggers the action of pressing key "a"; then, the feedback produced by pressing the key "a" (e.g., kinesthetic feedback, or perhaps the sight of letter "a" in the screen) triggers pressing the next key ("r"), and this process continues up to the end of the word.

In contrast to the response chain theory, Lashley postulated that all actions that are to be executed are determined at the onset of an action sequence within a *motor program* (Keele,

1968; Henry & Rogers, 1960). Lahsley provided three lines of evidence in support of his theory. Firstly, he argued that during the performance of an action sequence, reaction times are faster than can be attributed to the feedback received from previous actions. For example, (Rumelhart & Norman, 1982) argued that the mean interval between key presses in world champion typists is 60 milliseconds, which is close to the neural transmission time between the spinal cord and the effector, and thus there is not sufficient room for the feedback to guide actions. This argument, however, was based on the assumption that the minimum reaction time for responding to kinesthetic stimulation is 100 milliseconds (Glencross, 1977), whereas later evidence showed that reaction times can be faster than was initially thought (e.g., (J. A. Adams, 1976)). Moreover, not all reaction times are so fast that they cannot be explained based on the feedback received from previous stimuli (Bruce, 1994).

Secondly, Lashley provided evidence from a man with damaged nerves leading to his leg (which caused anesthesia for the movements of his knee joint, although he was still able to voluntarily move his leg) (Lashley, 1917), arguing that even in the absence of sensory feedback, individuals are able to make motor movements. Again here one can argue that in this situation the feedback can be provided using other senses (Bruce, 1994). The third line of argument is structure of the errors in the production of action sequences, which is not consistent with the response chain theory. For example, *transposition errors* during typing (e.g., typing "becuase" instead of "because") is one of the most common errors, which cannot be explained easily using response chain theory (Rumelhart & Norman, 1982).

Other evidence has also been found, that is more consistent with Lashley's account than the response chain account. Firstly, it is reported that the time required to initiate a sequence of movements increases in proportion to the length and complexity of the action sequence, which presumably indicates that the elements that are to be executed are prepared at the beginning of the sequence. In a seminal study, (Henry & Rogers, 1960) asked subjects to perform three different actions in response to a tone stimulus. In one portion of the study, subjects were instructed to simply withdraw their hands; in the second portion of the study, subjects were asked to make two rapid movements after they withdrew their hands. Finally, in

the third portion of the study, subjects were instructed to make four hand movements after hand withdrawal. The result of the study showed that, in the second portion of the study, the reaction time to start responding was 23 percent greater than the first portion (simple hand removal). Similarly, the reaction time to start responding, in the longest sequence was 31 percent greater than the simple hand withdrawal.

Other studies have reported results similar to the above study (e.g., (Klapp, 1995, 1977; Sternberg, Monsell, Knoll, & Wright, 1978; Verwey, 1994; Canic & Franks, 1989)). However, the differences in initiation times of short and long sequences diminish with practice (Klapp, 1995; Hulstijn & van Galen, 1983; Verwey, 1994, 1999; van Mier & Hulstijn, 1993), which is taken as a sign of the development of *motor chunks* (Brown & Carr, 1989; Klapp, 1995), i.e., with practice, the whole sequence of movements is encapsulated within a motor chunk that can be programmed as a single unit during the initiation of an action sequence.

The second line of evidence supporting Lashley's argument comes from neurological studies. Averbeck and colleagues (Averbeck, Chafee, Crowe, & Georgopoulos, 2002; Averbeck, Sohn, & Lee, 2006; Averbeck, Chafee, Crowe, & Georgopoulos, 2003) trained monkeys to perform a sequence of strokes to draw a predetermined geometric shape. For example, in a sample trial, a monkey was shown a triangle, and was then required to draw a triangle using a joystick. Within this example, the action sequence is composed of three actions, each corresponding to drawing one edge of the triangle (there were four different shapes in the experiment). Recordings from neurons in area 46 of the pre-frontal cortex of the animals, revealed that, at the onset of the sequence of strokes (before the animal started to draw the shape), each element of the action sequence to be executed was represented in the brain. In addition to this, the activity of the representation of each action, corresponded to the order of that action in the sequence, i.e., the action with the highest representation activity was the first action to be executed, and this representation was deleted after the action was executed. Such findings are more consistent with Lashley's conception of action sequences than with response chain theories. However, regarding the interpretation of this experiment, one can claim that the neural activity reflects the sub-goals to be met (e.g., drawing each edge of the shape), rather

than the sequence of actions (e.g., arm movements in this example), as indicated with some studies (Mushiake, Saito, Sakamoto, Itoyama, & Tanji, 2006; Saga, Iba, Tanji, & Hoshi, 2011; Shima, Isoda, Mushiake, & Tanji, 2007).

We will review more behavioral properties of action sequences in section 3.1 within the context of the development of a new model for learning action sequences.

2.4.5 Computational models: hierarchical RL

Theoretical perspectives on action sequences have a long history (e.g., (Estes, 1972; Rumelhart & Norman, 1982; Grossberg & Pearson, 2008; Cooper & Shallice, 2006; Botvinick, Niv, & Barto, 2009; Nakahara, Doya, & Hikosaka, 2001; Helie, Roeder, Vucovich, Rünger, & Ashby, n.d.; Solway & Botvinick, 2012; Ito & Doya, 2011). See for a review (Botvinick, 2008)), and one of the most recent forms is borrowed from the computational theory of *hierarchical RL* (Botvinick, 2008). This theory has the appealing property that it can be readily integrated with model-based RL (which corresponds to goal-directed actions), and besides that, it is simple and benefits from a sound theoretical foundation. Based on this, in the following, we present a review of hierarchical RL, and in the next chapters, we will be using it for model development and simulation.

As mentioned in section 2.3.3, model-based RL is not efficient in environments with high numbers of states and actions, and therefore, it is not scalable to complex environments. Hierarchical RL offers a solution to this problem by introducing the notion of *extended actions* or *temporally extended actions*. Temporally extended actions are actions that take more than one time step to complete. This is in contrast to the model-based and model-free reinforcement learning frameworks that we presented in 2.3.2 and 2.4.2, in which an action only took one time step to finish. The concept of temporally extended actions is close to the notion of action sequences in psychology, and this is why hierarchical RL provides a suitable framework for studying action sequences.

In order to incorporate temporally extended actions into RL, we need to extend the notion of

MDPs (introduced in section 2.2) to semi-MDPs (SMDP), in which transition from one state to another state can take more than one time step. As such, instead of the transition model being $P(s'|s,a)$ it will be $P(s',\tau|s,a)$, which represents the probability of reaching state s' after the time step of τ, if action a was taken in state s. Within an SMDP, some actions can take more than one time step to finish, while others can take only one time step to finish, similar to MDPs. Actions that take one time step to finish are called *primitive actions*, as opposed to temporally extended actions. Given an SMDP, the role of a hierarchical RL agent is to take the course of action (either primitive or temporary extended actions) that maximizes the long-run expected cumulative reward. To achieve this aim, several frameworks have been suggested previously, such as *options* formalism (Sutton, Precup, & Singh, 1999), the HAMQ approach (Parr & Russell, 1998), and the MAXQ value decomposition (Dietterich, 2000). We will be using the options framework, due to its simplicity and popularity, and it will be presented in the next section.

Options framework- A commonly used environment for the study of hierarchical RL is depicted in Figure 2.5. The environment consists of four rooms, which are connected through hallways. An agent is randomly initialized in one of the cells in the top-left room, and it can move to other cells by taking actions. There are four primitive actions that the agent can take (left, right, top, down), which move the agent to the corresponding cells in a stochastic manner. The agent will receive a reward whenever it reaches the goal ('G' in the figure). In addition to the primitive actions, at each cell, the agent can also select from the two *options* that take it to the hallways of the current room. For example, if the agent is in the top-left room, one of the options takes it to the right hallway, and the other option takes it to the bottom hallway (there are eight options in total). Executing each option takes more than one time step, and thus, options play the role of temporary extended actions. One way of thinking about options is that they are solutions to sub-problems. For example, in the four-room task to reach the goal, the agent first needs to reach a hallway (sub-problem), which can be achieved by executing the appropriate option.

More generally, an option o is defined with three properties $o = \langle I^o, \pi^o, \beta^o \rangle$. I^o is called the

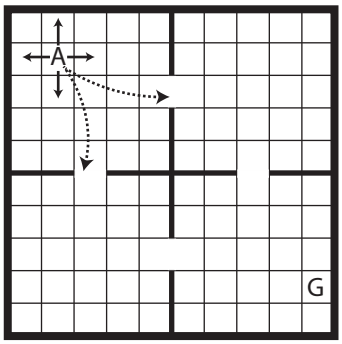


Figure 2.5 – The four-room environment, which is used for studying hierarchical RL. Solid arrows show primitive actions (left, right, up, down), and dashed arrows show available options.

initiation set, which is the set of states in which option o can be initiated. π^o is the internal action-selection policy of the option, and defines how the option will be implemented at the level of primitive actions. Finally, $\beta^o(s)$ is the termination condition of the option, which defines the probability of the termination of the option upon entering state s.

Given a set of options and actions, the algorithm works as follows. In each state, the agent evaluates the options that can be taken in that state, and it also evaluates the available primitive actions in that state. If the agent chooses to run a primitive action, then it will execute that action, and the process restarts in the next step. If the agent chooses to run an option, then the option will start to select primitive actions according to its internal policy ($\pi(s)$) until it terminates. At each step, the option will be terminated with a certain probability. For example, in the above scenario, the options will be terminated whenever the agent reaches the intended hallway. In addition to such conditions (e.g., reaching the intended goal of executing the option), there is usually one more condition in which an ongoing option can be terminated. That situation occurs when, during the execution of an option, the agent realizes that there is a primitive action which has a higher value than continuing the option. For example, in the four-room scenraio, assume that an agent is in the top-left room, and starts an option in order to reach the eastern hallway. However, due to the unpredictability of the environment, after executing several actions, the agent finds itself near the southern hallway. In this condition, the agent terminates the ongoing option (to reach the eastern hallway), and exits the southern

hallway.

Options, from the respect that they are composed of several primitive actions, correspond to the concept of action sequences in psychology. The difference, however, is that the execution of options is based on the states of the environment, but the execution of action sequences can be independent of the state of the environment, as discussed by Lashley. We will discuss this issue in more detail in chapter 3. The second parallel is between the concept of option termination, and the *inhibition* of action sequences. As we will discuss in the next chapters, in certain conditions, the execution of an ongoing action sequence might be inappropriate, meaning that the action sequence should be terminated, which is conceptually similar to the termination of options.

model-based hierarchical RL- In the previous section we introduced the concept of options. Given a set of options and primitive actions, a remaining question is how an agent evaluates and selects actions. In principle, similar to primitive actions, options can be evaluated in both a model-based and model-free manner. In the next chapters of the current thesis, we will argue that a framework in which options and actions are evaluated in a model-based manner, provides a parsimonious explanation for the multiple forms of decision-making processes. Such a framework is called *hierarchical model-based RL*, which as its name implies, includes options (the hierarchical component), and evaluates actions in a model-based manner, i.e., an agent builds a representation of the model of the environment.

The representation of the model of the environment that the hierarchical model-based RL builds, consists of two parts. The first part is similar to the representation of the model-based RL introduced in section 2.3.2, i.e., an agent learns the transition function of the primitive actions ($p(s'|s,a)$), and also the reward of each state ($R(s)$). The second part involves learning about options, and has three components: (i) discovering options, i.e., learning how primitive actions can be organized in the form of useful options, (ii) learning the outcome of each option after discovering them, i.e., to which state each option leads to ($p(s'|o,a)$), and (iii) learning the total reward earned during the execution of the option ($R(o)$). $R(o)$ summarizes all the

events that happen during the execution of an option, in terms of the amount of rewards earned.

By building a representation of the environment, an agent can perform a tree search in order to select actions. In the case of primitive actions, this search is similar to the flat model-based RL (section 2.3.2). In the case of options, their value is composed of the value of the final state that the option leads to, and the amount of reward that can be earned during the option execution, which can be correspondingly calculated using $p(s'|s,a)$, and $R(o)$:

$$Q(s,o) = R(o) + \sum_{s'} P(s'|s,o)V(s') \tag{2.15}$$

where $Q(s,o)$ is the value of execution option o in state s. It is important to note that the term $R(o)$ represents the total reward earned during the execution of the option, without representing the actual states visited during the execution of that option[5]. In particular, since for the evaluation of the options, the states visited during the option execution are not considered individually, then off-line changes in the values of these middle states will not reflect on the values of the options that pass them. As we will discuss in the subsequent chapters, this property can render decisions insensitive to outcome devaluation, a characteristic of automatic actions.

There are two theoretical points worth mentioning regarding the utility of using options. The first one is regarding whether using options speeds up learning, i.e, how adding options to the learning algorithm affects the sample complexity of the algorithm. Experimental (Jong, Hester, & Stone, 2008) and theoretical (Brunskill & Li, 2014) studies indicate that adding options can *potentially* increase the speed of learning, depending on whether options are available in all, or a sub-set of states. In fact, it can be shown that in the four-room environment, if the agent is allowed to choose an option in all of the states, then the performance of the agent will be

[5] $R(o)$ in fact summarizes events that have occurred during the execution of option o. However, one can imagine other ways to summarize events that have occurred during the execution of option o. For example one can assume that a quantity called $S(o)$ tracks the probability of visiting each state during the execution of option o (i.e., cached states). Here, consistent with the current formulation of the options framework we assume that the agent represents $R(o)$.

even *worse* than an agent without options (Jong et al., 2008). For example, if the agent is near the goal, choosing to run an option will stop the agent from reaching the goal, because it will take the agent to the hallways which are far from the goal. However, in the rooms other than the goal room, using options can be beneficial. Similarly, theoretical results also suggest that using options can improve the sample complexity of the learning algorithm, if the options are available in a limited number of states, and in the rest of the states, only primitive actions are available.

The second theoretical point is how adding options changes the computational complexity of the learning algorithms. In discrete MDPs, there are some indications that using options can facilitate the tree-search value computation (He & Brunskill, 2011), however these results are restricted to the condition that only options can be selected by the agent (and not primitive actions) (see (Mann & Mannor, 2014) section 1 for a brief review, and a generalization of the results to the case of continuous MDPs). However, restricting the agent to select only options, can lead to sub-optimal policies, since for example achieving a goal might need utilizing fine-grain actions in addition to the options. As such, the utility gained by using options depends on the states in which options are available, and whether options augment the space of primitive actions or whether they replace primitive actions. In addition, how options are discovered, i.e., how primitive actions form an option, can also affect the performance of the learning algorithm. Obviously, if the options match the hierarchical structure of the environment, then they will benefit the agent in reaching its goals, otherwise, they might hinder performance. In chapter 3 we will introduce a new method for discovering action sequences.

2.4.6 Neural substrates

Previous studies have implicated several brain regions in the computational processes involved in decision making (model-based Pavlovian, model-free Pavlovian, model-based instrumental, model-free instrumental)(see for example (Balleine, Delgado, & Hikosaka, 2007; Dolan & Dayan, 2013) for reviews). Here, we will review the neural substrates that are more relevant to the current thesis. First, we briefly review the role of dopamine in automatic actions, and then,

we will review the role of striatal sub-regions in each form of automatic actions.

2.4.6.1 Role of dopamine

Insensitivity to outcome devaluation induced by over-training has been shown to depend on the ascending nigrostriatal dopamine pathway in rats (Faure, Haberland, Condé, & El Massioui, 2005), and the expression of NMDA receptors on dopamine neurons in mice (Wang et al., 2011). Dopamine has also been implicated in the operation of action sequences. Evidence suggests that the administration of a dopamine antagonist disrupts the chunking of movements into well-integrated sequences (in capuchin monkeys) (Levesque et al., 2007), which can be reversed by the co-administration of a dopamine agonist (Tremblay et al., 2009). In addition, motor chunking appears not to occur in Parkinson's patients (Benecke, Rothwell, Dick, Day, & Marsden, 1987) due to a loss of dopaminergic activity in the sensorimotor putamen, which can be restored in the patients on L-DOPA (Tremblay et al., 2010). The computational role that dopamine plays in the formation of action sequences is unclear, and in chapter 3 we develop a model for learning action sequences, which is based on the reward prediction error, and thus provides a basis for understanding the role of dopamine in learning action sequences.

2.4.6.2 Striatal sub-regions

While overtraining causes performance to become insensitive to outcome devaluation, lesions of the dorsolateral striatum (DLS; the sensorimotor striatum in rats) reverse this effect rendering performance once again sensitive to devaluation treatments (Yin et al., 2004). Likewise, muscimol inactivation of DLS has been found to render otherwise habitual performance, sensitive to changes in the action–outcome contingency (Yin, Knowlton, & Balleine, 2006). There are similar reports in humans showing the involvement of sensorimotor putamen in automatic actions (Tricomi et al., 2009; Wunderlich, Smittenaar, & Dolan, 2012).

Similarly, inactivation of the sensorimotor striatum disrupts the expression of previously learned motor sequences (Miyachi, Hikosaka, Miyashita, Kárádi, & Rand, 1997), and learning

new sequences (Yin, 2010). In humans, the blood oxygen level dependent activity in the sensorimotor putamen is correlated with the concatenation of action sequences (Wymbs, Bassett, Mucha, Porter, & Grafton, 2012), and the stage of training in action sequence learning (Lehéricy et al., 2005; Jueptner, Frith, Brooks, Frackowiak, & Passingham, 1997). Neural firing patterns recorded in rat's DLS have been reported to mark the start and end of action sequences in T-maze navigation (Thorn, Atallah, Howe, & Graybiel, 2010), and sequences of lever presses (Jin & Costa, 2010; Jin, Tecuapetla, & Costa, 2014). Furthermore, it is reported that most of the striatal neurons that were more active during the performance of a learned action sequence, were in the sensorimotor striatum, whereas neurons in the associative striatum responded more strongly to the performance of a new action sequence (Miyachi, Hikosaka, & Lu, 2002).

For example, (Jog, Kubota, Connolly, Hillegaart, & Graybiel, 1999) over trained rats in a T-maze and found that, as the component responses performed between the start and end points of the maze declined in latency, the neural activity in sensorimotor striatum specific to those points also gradually declined, to the extent that task-related activity was limited to the first and last responses in the maze, i.e. the points at which any response sequence or chunk should have been initiated and terminated. Similar effects have been observed by (Barnes, Kubota, Hu, Jin, & Graybiel, 2005) and in fact, using response reversals, they were also able to observe a collapse of the sequence, i.e. both the inter-maze responses and the neural activity initially associated with those responses, reemerged along with task-irrelevant movements. The inter-maze responses again declined with continued training post-reversal. Hence, changes in both striatal unit activity and incidental behavioral responses tracked the development of the sequence, as has been observed recently using a homogeneous sequence of lever press responses, described above (Jin & Costa, 2010), and in a head movement habit (Tang, Pawlak, Prokopenko, & West, 2007). Finally, (Kubota et al., 2009) reported observing electrophysiological effects associated with both the overall response sequence and the component response primitives as these were rapidly remapped onto new stimuli presented during the course of T-maze performance, suggesting that the mice (in this case) maintained

separate representations of the sequence and component movements. Interestingly, Redish and colleagues (Schmitzer-Torbert & Redish, 2004; A. Johnson, van der Meer, & Redish, 2007) reported neural signals in the striatum associated with sequence learning in a novel navigation task composed of a series of T-mazes reordered across days. Striatal activity differed across different sequences of turns in the maze, but was also highly correlated across repeated sequences suggesting, again, the co-occurrence of movement-specific and sequence-specific activity in the striatum.

Thus, in summary, there are considerable similarities between the neural structures mediating various forms of automatic actions, in terms of both striatal sub-regions and the role of dopamine. The role of the sensorimotor striatum in action sequences, however, is not without controversy (Turner & Desmurget, 2010). For example, inactivation of the globus pallidus internus - the principal output of the sensorimotor striatum - does not increase the reaction time in the 'sequential trials' of the sequential reaction times task (SRTT), which suggests that the performance of action sequences is not dependent on the sensorimotor striatum. However, the execution of sequence-based actions (as opposed to encoding) may not be dependent on the sensorimotor striatum. In fact, based on this assumption, a neurocomputational model has recently been developed (Helie et al., n.d.), in which the role of the basal ganglia is to train the cortical-cortical connections that mediate sequence production. The model has shown to be able to account for the data of (Desmurget & Turner, 2010) as well as some other aspects of sequence learning.

2.5 Summary

In this chapter we introduced different forms of decision-making processes in the brain: (i) Pavlovian conditioning and its two variants corresponding to model-based and model-free value learning, and (ii) instrumental conditioning, which can be in the form of goal-directed or automatic actions. We further presented three variants of RL models (model-based RL, model-free RL, and hierarchical RL), and provided behavioral and neural evidence in support of each

model. In the next chapter, we look closer at the behavioral and computational properties of hierarchical RL, and investigate their implications for insensitivity of automatic actions to outcome devaluation, contingency degradation, and action sequence learning.

3 Hierarchical decision-making: learning action sequences

In the previous chapter, we showed that insensitivity to outcome devaluation, and contingency degradation are two important aspects of automatic actions, which are attributed to model-free RL in previous works. In addition, we mentioned that action chunking is also an important aspect of automatic actions, which has been linked to hierarchical RL. In this chapter, we show that these two categories of automatic actions can be explained coherently using a hierarchical RL model. After reviewing some behavioral properties of action sequences, we develop a new normative model for learning action sequences and their interaction with goal-directed processes. Finally, we show how the new model can account for a typical sequence learning experiment, as well as outcome devaluation and contingency degradation experiments.

3.1 Open-loop performance of action sequences

The flexibility of goal-directed actions reflects, the need for immediate, or at least rapidly acquired, solutions to new problems and, indeed, evidence suggests that in novel environments there is a strong tendency for animals to generate behavioral variation and to resist immediately repeating prior actions or sequences of actions (Neuringer, 2004; Neuringer & Jensen, 2010). Of course, the need to explore new environments requires behavioral variation;

once those solutions are found, however, exploiting the environment is best achieved through behavioral stability, i.e. by persisting with a particular behavioral response. It is important to recognize that, with persistence, actions can change their form, often quite rapidly. Errors in execution and the inter-response time are both reduced and, as a consequence, actions previously separated by extraneous movements or by long temporal intervals are more often performed together and with greater invariance (Willingham, Nissen, & Bullemer, 1989; Buitrago, Ringer, Schulz, Dichgans, & Luft, 2004; Buitrago, Schulz, Dichgans, & Luft, 2004). With continuing repetition these action elements can become linked together and run off together as a sequence, i.e. they can become chunked (Terrace, 1991; Graybiel, 1998). Practice appears, therefore, to render variable, flexible, goal-directed actions into rapidly deployed, relatively invariant, components of action sequences, suggesting that an important way in which the form of a goal-directed action can change as it becomes habitual is via the links that it forms with other actions to generate sequences (section 2.4.4).

The most important feature of sequence learning is *the interdependency of actions* (Shah, 2008). Through the process of sequence learning, action control becomes increasingly dependent on the history of previous actions and independent of environmental stimuli, to the point that, given some triggering event, the whole sequence of actions is expressed as an integrated unit. Take, for example, a typical sequence-learning experiment such as the serial reaction time task (SRTT; (Nissen & Bullemer, 1987)). In this task a subject is required to elicit a specific action in response to a specific cue. For example, cues can be asterisks that are presented on a computer screen, and the corresponding responses require the subject to press the keys that spatially match the cues' positions that can appear in either a random or sequential order. In the *sequential trials* condition, the position of the cues is repeated in a pattern such that the position of the next stimulus can be predicted given the previous one. In the *random trials* condition, stimuli are presented in random order and, thus, given the previous stimuli, the next one cannot be predicted.

On a simple conception of the learning process in the SRTT, the subject takes an action and, if it is successful, the association between that action and the cue strengthens. Then, the subject

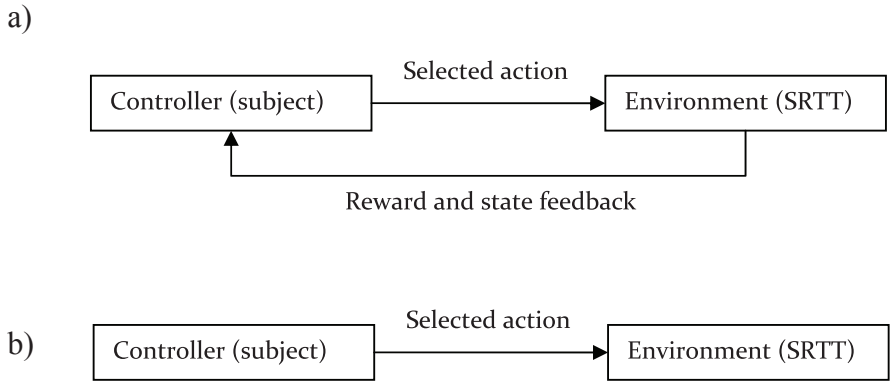


Figure 3.1 – (a) A closed-loop control system. After the controller executes an action it receives cues regarding the new state of the environment and a reward. (b) An open-loop control system in which the controller does not receive feedback from the environment. SRTT, serial reaction time task.

.

waits for the next cue, and takes the action that has the strongest association with the cue. From a control theory point of view, this learning process can be characterized as a *closed-loop control system* in which, after each response, the controller (here the subject), receives a *feedback* signal from the environment to guide future actions. In the SRTT, after taking an action the subject receives two types of feedback: reward feedback and state feedback. Reward feedback is the reward received after each response, for example, a specific amount of juice, and is used for learning stimulus–response associations. State feedback is given by the presentation of the next cue after the response that signals the new state of the environment, and is used by the subject to select its next action. The term closed-loop commonly refers to the loop created by the feedback path (Figure 3.1a). In the absence of this feedback, the control system is called *open-loop* (Figure 3.1b; (Astrom & Murray, 2008)), according to which the actions that the subject takes are not dependent on the presented cues or the received rewards.

Clearly, closed-loop control is crucial for learning. Without feedback, the agent cannot learn which response is correct. It also needs state feedback from the environment because the correct response differs in each state and, as such, without knowing the current state, the agent cannot make the correct response. However, in the sequential trials condition of the SRTT,

the subject could potentially maintain a high level of performance without using the state feedback; indeed, evidence suggests that is exactly what they do in accord with open-loop control. Reaction time, defined as the time between the stimulus onset and the onset of the behavioral response, is the primary measure in the SRTT. Evidence from rodents (Schwarting, 2009), non-human primates (Hikosaka, Rand, Miyachi, & Miyashita, 1995; Matsuzaka et al., 2007; Desmurget & Turner, 2010) and humans (Nissen & Bullemer, 1987; Keele, Ivry, Mayr, Hazeltine, & Heuer, 2003) suggests that, after training, reaction times are shorter in the sequential trials than the random trials condition. In fact, if the subject is permitted to respond during the inter-trial delay, then the reaction time can even be negative, i.e. the next action is often executed before the presentation of the next stimulus (Matsuzaka et al., 2007; Desmurget & Turner, 2010). (Matsuzaka et al., 2007) reported that, with increasing training, the number of these predictive responses increases up to the point that in almost all trials, the subject responds in a predictive manner without the visual cue (see also (Miyashita, Rand, Miyachi, & Hikosaka, 1996), indicating that the number of predictive responses increases as a sequence becomes well learned (see also (Carr & Watson, 1908) for a similar phenomenon in maze navigation).

The occurrence of very short or negative reaction times in sequential trials implies that, after sufficient learning, selection of an action is mostly dependent on the history of previous actions and less dependent on the external stimuli (visual cues). In fact, even if the subject does not respond predictively before stimulus presentation, it is clear that the decision as to which action to take on the next trial is made *before* stimulus presentation. In a sequential button-push task, (Matsumoto, Hanakawa, Maki, Graybiel, & Kimura, 1999) trained a monkey to execute a series of three button pushes in response to illumination of the buttons in a fixed cued sequence. After training, the monkey was tested in a condition in which the third button in the sequence was located in a position different from its position during training. They found that, during the first and sometimes the second trial, the monkeys would continue to push the third button of the learned sequence even if one of the other targets was illuminated. Similarly, (Desmurget & Turner, 2010) reported when the first stimuli of a random trial followed,

by coincidence, the pattern of stimuli from a learned sequence, the animal responded as if the next stimuli will be drawn from the learned sequence.

It appears, therefore, that, during the first stages of training, the subject learns the association between cues and responses. At this stage, action selection is under closed-loop control and relies on the observation of the cues. In the case of random trials, action selection remains closed-loop through the course of learning. In sequential trials, however, with further learning, action selection switches to open-loop control in which the execution of successive actions is not related to the current state of the environment, something that leads to the expression of action chunks (section 2.4.4). When actions are expressed in chunks, both state identification, based on visual cues, and action evaluation appear to be bypassed. (Endress & Wood, 2011), for example, note that successful sequence learning requires view-invariant movement information, i.e. rather than depending on the relation between visual cues and movement information in allocentric space, as goal-directed actions do (Willingham, 1998), sequential movements appear to depend on position-based encoding in egocentric space. Hence, chunked in this way, the performance of sequential movements is largely independent of environmental feedback, allowing for very short reaction times in the open-loop mode.

Another line of evidence consistent with the cue-independency notion of habitual (automatic) behavior comes from place/response learning tasks in animals (Tolman, Ritchie, & Kalish, 1946; Restle, 1957). In this type of task, rats begin each trial at the base of a T-maze surrounded by environmental cues (e.g. windows, doors), and are trained to find food at the end of one arm (e.g. the right, or east, arm). Following this training period, they are given probe tests in which the maze is rotated 180 (with respect to the cues), and thus the start point will be at the opposite side of the maze. Results show that after moderate training, at the choice point the animal turns in the opposite direction to that previously learned (i.e. towards the west arm; place strategy), suggesting that action control is state-dependent and based on the environmental cues (closed-loop action control). However, after overtraining, rats switch and at the test they take the same action that they learned in the initial training (i.e. they turn right; a response strategy), indicating that overtraining renders action selection at the choice point

independent of the environmental cues and the state identification process (open-loop action control; (Ritchie, Aeschliman, & Pierce, 1950; Packard & McGaugh, 1996)).

Similarly, in more complex mazes in which a sequence of actions is required to reach the goal, removal of environmental cues does not affect performance of a learned sequence of egocentric movements (body turns), but disrupts the use of a place strategy (Rondi-Reig et al., 2006). Learning the maze in a cue-deficient environment, but not in a cue-available environment, in which decision-making should minimally rely on state-guided action control is impaired by inactivation of sensorimotor striatum (Chang & Gold, 2004). Few studies have addressed functional differences between the sensorimotor striatum and associative striatum in place/response learning; however, in general it seems that the associative striatum is involved in goal-directed decision-making (the place strategy), and the sensorimotor striatum is involved in habitual responses (the response strategy; (Devan & White, 1999, 1999; Yin et al., 2004; Moussa, Poucet, Amalric, & Sargolini, 2011)), consistent with the role of these striatal sub-regions in instrumental conditioning (section 2.4.6.2) and SRTT.

Based on the mentioned similarities in neural and behavior aspects of increasing automaticity in SRTT (sequential trials), maze learning and instrumental conditioning (insensitivity to outcome devaluation and contingency degradation), we assume that action sequence formation is the underlying process of these modalities of habitual behavior. In order to formalize this assumption, in the next section we use RL to provide a normative approach to modeling changes in performance during action sequence learning. Next, in section 3.5 we show how this model applies to the different forms of habitual behavior.

3.2 Average reward RL

A typical RL agent utilizes the following components for the purpose of learning and action selection: (i) state identification – the agent identifies its current state based on the sensory information received from its environment (e.g. visual cues); (ii) action selection – given its current state, and its knowledge about the environment, the agent evaluates possible

actions then selects and executes one of them; (iii) learning – after executing an action, the agent enters a new state (e.g. receives new visual cues) and also receives a reward from the environment. Using this feedback, the agent improves its knowledge about the environment. This architecture is a closed-loop decision-making process because the action-selection process is guided by the current state of the agent, which is identified based on sensory inputs received from the environment. As we discussed in the previous section, action selection in sequence learning is not guided by environmental stimuli, and so does not require a state identification process. To explain sequence learning, therefore, we need to modify this framework. In the following sections we will introduce a mixed open-loop/closed-loop architecture for this purpose. Before turning to that issue, we shall first take a closer look at the learning process in RL.

Assume that an agent is in state s, in which there is a set of possible actions, and the agent selects one of them for execution, that we denote with a. The agent spends d time steps in state s (commonly referred to as the *state dwell time*) and, after that, by taking actions a, it enters a new state, s' and receives reward r. The next state of the agent, the amount of reward received after taking an action, and the state dwell times, depend on the dynamics of the environment, which are determined, respectively, by the transition function, the reward function and transition time function. The transition function, denoted by $p(s'|s,a)$ indicates the probability of reaching state s' upon taking action a in state s. $R(s)$ denotes the reward function, which is the amount of reward the agent receives in state s. Finally, $D(s)$, the transition time function, is the time spent in state s (dwell time). The time that the agent spends in a state is the sum of the time it takes the agent to make a decision, and the time it takes for new stimuli to appear after executing an action.

The goal of the agent is to select actions that lead to a higher average reward per time step, and this is why this formulation of the RL is called 'average reward RL' (Mahadevan, 1996; Tsitsiklis & Roy, 1999; Daw & Tourtezky, 2000; Daw, 2002). This average reward, denoted by $\bar{R}$, can be defined as the total rewards obtained, divided by the total time spent for acquiring

those rewards:

$$\bar{R} = \frac{r_0 + r_1 + r_2 + \ldots}{d_0 + d_1 + d_2 + \ldots} \tag{3.1}$$

If the environment is *unchain*[1] (e.g., cyclic environment) then the average reward ($\bar{R}$) will be same for all the states. This condition holds for the cyclic environment that we discuss in this chapter.

To choose an action amongst several alternatives, the agent assigns a subjective value to each state–action pair. This subjective value is denoted by $Q(s,a)$, and represents the value of taking action a in state s (Watkins, 1989). These Q-values are learned such that an action with a higher Q-value leads to more reward in a shorter time compared with an action with a lower Q-value.

The first determinant of $Q(s,a)$ is the immediate reward that the agent receives in s, which is $R(s)$. Besides the immediate reward the agent receives, the value of the next state the agent enters is also important: actions through which the agent reaches a more valuable state are more favorable. Thus, assuming that the agent reaches state s' by taking action a, the value of the next state, $V(s')$, is the second determinant of $Q(s,a)$ and is assumed to be proportional to the reward the agent gains in the future by taking its best action in the state s'. In general, for any state s, $V(s)$ is defined as follows:

$$V(s) = \max_a Q(s,a) \tag{3.2}$$

The final determinant of $Q(s,a)$ is the amount of time the agent spends in the state. If the agent spends a long time in a state, then it will lose the opportunity of gaining future rewards. In fact, losing $D(s)$ time steps in state s is equal to losing $D(s)\bar{R}$ that could potentially be accrued in this time. Given these three determinants, the value of taking an action in a state can be

[1] Formally an environment is unchain if every state of the environment will be revisited eventually by probability 1, except for a finite (or empty) set of states which will never be visited after a certain point in time.

computed as follows:

$$Q(s,a) = R(s) - D(s)\bar{R} + E[V'(s)] \tag{3.3}$$

where the expectation in the last term is over s':

$$Q(s,a) = R(s) - D(s)\bar{R} + \sum_{s'} p(s'|s,a)V(s') \tag{3.4}$$

As the above equation implies, computing Q-values requires knowledge of the transition probabilities, the reward functions and the state dwell times, which together constitute a *model of the environment*. However, without a prior model of the environment, the agent can estimate these quantities through its experience with the environment. For example, $R(s)$ can be estimated by averaging immediate reward received in state s. In the same manner, $D(s)$ can be computed as the average of waiting times in state s. An estimation of $p(s'|s,a)$ can be made by counting the number of times taking action a in state s leads to state s'. Given the model of the environment, Q-values can be derived from equation 3.4 using dynamic programming algorithms, such as *value-iteration* (Puterman, 1994; Mahadevan, 1996). Because these methods of value computation rely on the model of the environment, they are called *model-based* value computation methods. Using these state–action pairs, the agent can guide its actions toward ones that lead to a higher average reward rate.

Returning to the overview of the decision-making process in RL, in (i) the agent identifies its current state, s and then feeds state s into equation 3.4, allowing the value of different actions, Q-values, to be computed. These Q-values guide the action-selection process, and the agent takes the appropriate action (ii). By taking an action, the agent enters a new state, receives a reward, and measures the time from entering the previous state, s, to entering the new state, s'. Finally, using these quantities, the model of the environment is updated (iii).

3.3 Action sequence formation

When an agent starts learning in a new environment all the decisions are based on model-based action selection, i.e. after entering a new state, the agent computes Q-values using the process introduced in the previous section and chooses one of the actions that tends to have the higher Q-value. Under certain conditions, however, it may be more beneficial for the agent to execute actions in a sequence without going through the action-selection process. First, we discuss the process of sequence formation and, in the next section (section 3.4), how action sequences interact with the model-based action selection.

We start by reviewing the environmental conditions in which action sequences form. Figure 3.2 shows three environments in which, by taking action A_1 in state S, the agent enters state S' or S'' with equal probability. In states S' and S'' two actions are available, A_2 and A_2'. Figure 3.2a provides an example of the kind of environment in which an action sequence forms, i.e. in states S' and S'', action A_2 is the best action. An example of this environment is the situation in which pressing a lever (action A_1) leads to the illumination of, say, a light (state S') or a tone (state S'') with equal probability, both of which signal that by entering the magazine (action A_2), the rat can gain a desirable outcome, and by not entering the magazine (action A_2') it gains nothing. As a consequence, after taking action A_1 in S the agent does not need to check the upcoming state but can execute A_2 irrespective of the next state, either S' or S'' (light or tone). In this situation, actions A_1 and A_2 form an action sequence consisting of A_1 and A_2. Hereafter in this chapter, we call these action sequences *macro actions*, and denote them, for example, in this situation with $\{A_1 A_2\}$. Actions that are not macro, for example A_1 or A_2, are called *primitive actions*.

Figure 3.2b shows a situation in which an action sequence does not form. In state S', action A_2 is the best action, but in state S'', action A_2' is the best. In the context of the previous example, illumination of the light indicates that by entering the magazine, the animal will gain a desirable outcome; but presentation of the tone indicates that entering the magazine is not followed by a desirable outcome. Here, after taking A_1 in S, the animal cannot select an action

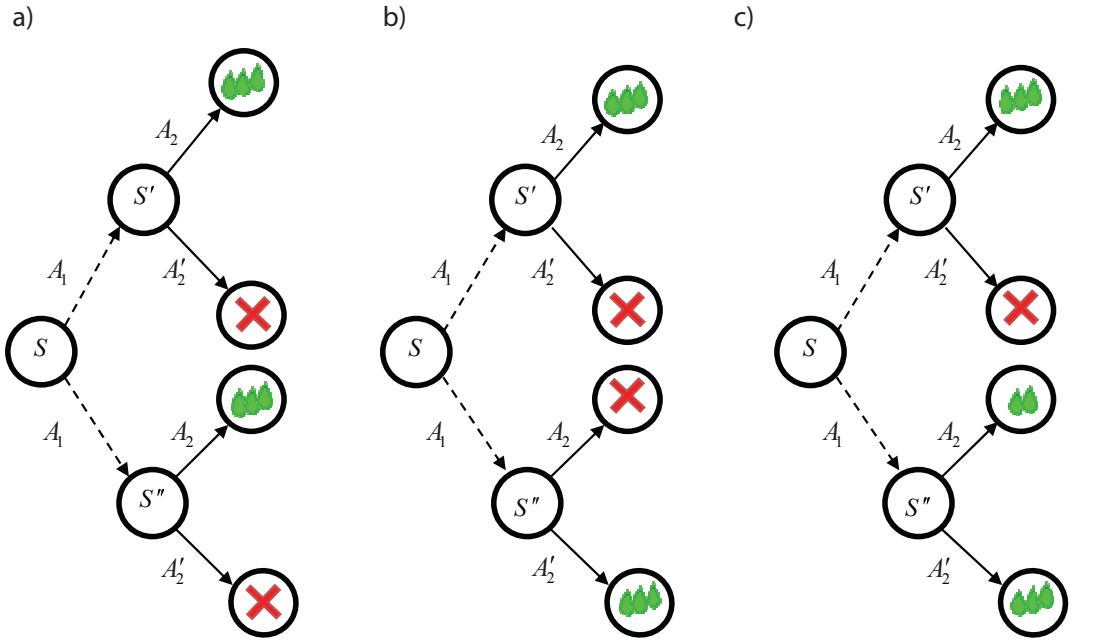


Figure 3.2 – (a) An example of an environment in which action sequences will form. Action A_1 leads to two different states with equal probability, in both of which action A_2 is the best action, and thus action sequence $\{A_1 A_2\}$ forms. (b) An example of an environment in which action sequences do not form. Action A_1 leads to two different states with equal probability, in one of which action A_2 is the best action and, in another, action A'_2 is the best action. As a consequence, an action sequence $\{A_1 A_2\}$ does not form. (c) An example of an environment in which the process of sequence formation is non-trivial. Action A_1 leads to two different states with equal probability, in one of which action A_2 is the best action, but in the other action A'_2 is the best action (although A_2 is a little bit worse than A'_2).

without knowing the upcoming state, and thus a macro action does not form.

Figure 3.2c shows a more challenging example. In state S', A_2 is the best action. In state S'', A_2 is not the best action, but it is slightly worse than the best action, A_2' (e.g. three drops of a liquid reward, vs. two drop of liquid reward). Does a sequence form in this case? To answer this question, we need a cost-benefit analysis, i.e. what the agent gains by executing actions A_1 and A_2 in sequence and what it loses. Assume it decides to always execute A_2 after A_1. If the next state is S', then it loses nothing, because action A_2 is the best action in state S'. But, if the next state is S'', by taking A_2 instead of the best action, A_2', the agent loses some of the future rewards. The amount of these reward losses is equal to the difference between the value of action A_2, $Q(S'', A_2)$, and the value of the best action, $V(S'')$, which will be $Q(S'', A_2) - V(S'')$, that we denote by $A(S'', A_2)$. $A(S'', A_2)$ can be interpreted as the advantage of taking action A_2 in state S'' instead of the best action (Baird, 1993; Dayan & Balleine, 2002). In this example, because the agent enters state S'' after state S only half the time, the total cost of executing A_1 and A_2 in sequence will be $0.5A(S'', A_2)$.

Generalizing from the previous example, the cost of executing the macro action $\{aa'\}$ in state s is equal to:

$$C(s, a, a') = E[Q(s', a') - V(s')] = E[A(s', a')] \tag{3.5}$$

where expectation over the next state, s', given the previous action and the previous state, is:

$$C(s, a, a') = \sum_{s'} P(s'|s, a) A(s', a') \tag{3.6}$$

Using the above equation, the term $C(s, a, a')$ can be computed based on the model-based approaches described. The agent computes $Q(s', a')$ and $V(s')$ using equation 3.4, and then $C(s, a, a')$ is calculated by equation 3.6. However, this means that at each decision point, deciding whether to execute an action sequence, equation 3.6 should be evaluated for all currently possible actions, and all possible subsequent actions. This will likely impose a heavy processing load on the decision-making process, and could considerably increase the latency

of action selection. It turns out, however, that $C(s,a,a')$ can be estimated efficiently using samples of the temporal difference error signal (TD error signal).

Let's assume that an agent is in state s' and takes action a' and reaches state s'' and receives reward r. Then the TD error signal experienced after taking action a' in state s' is defined as follows:

$$\delta = [r - d\bar{R} + V(s'')] - V(s') \tag{3.7}$$

Based on equation 3.4, the term $[r - d\bar{R} + V(s'')]$ is a sample of $Q(s',a')$. Thus, δ will be a sample of $A(s',a')$[2], and hence $C(s,a,a')$ can be estimated using samples of the TD error signal. By taking action a in state s, and action a' in state s', $C(s,a,a')$ will be updated as follows:

$$C(s,a,a') \leftarrow (1-\eta_C)C(s,a,a') + \eta_C \alpha \delta \tag{3.8}$$

where η_C is the learning rate, and α is a factor, which equals 1 when the environment is deterministic (see section 3.8 for more details). As mentioned above, extensive evidence from animal and human studies suggests that the TD error signal is coded by the phasic activities of mid-brain dopamine neurons (Schultz et al., 1997; Schultz, 2000). Thus, besides being more efficient, utilizing the error signal for the purpose of sequence learning provides a neurally plausible way for computing the cost of sequence-based action selection, $C(s,a,a')$.

Up to now, we have only considered one side of the trade-off, which is the cost of sequence-based action selection. What are the benefits of sequence-based action selection? As discussed in the previous section, expression of a sequence of actions is faster than selecting actions one by one, based on the action evaluation process. This can be for several reasons; for example, identification of the current state by processing environmental stimuli can be time consuming, and the evaluation of actions using a model-based process is slower than having solely to select the next action from the sequence. Besides being faster, executing actions without

[2]Note that we assumed that advantages (A) are calculated based on the value of the best action ($V(s)$), and not based on the current policy.

going through the decision-making process makes it possible to perform a simultaneous task that requires decision-making resources. Here, we focus on the first advantage of sequence learning.

Assume that selecting the next action of the current sequence is τ time steps faster than selecting an action based on the action evaluation process. Saving τ time steps is equivalent to gaining $\bar{R}\tau$ more reward in the future (Niv, 2007). This provides the other side of the trade-off: if the benefit of sequence-based action selection, $\bar{R}\tau$, exceeds its costs, $-C(s,a,a')$, then the macro action $\{aa'\}$ replaces action a in state s. Otherwise, if the macro action is already formed, it decomposes to its constitutent actions, and action a replaces the macro action $\{aa'\}$:

if $-C(s,a,a') < \bar{R}\tau$ **then**
  replace action a with macro action $\{aa'\}$ in state s
else
  replace the macro action $\{aa'\}$ with action a in state s
end if

After a macro action is added, it can be concatenated with other actions to form a longer macro action. For example, macro action $\{aa'\}$ can be concatenated with another action, say a'', and form the macro action $\{aa'a''\}$. It is important to recognize that, during execution of a macro action, primitive actions are not evaluated and thus the TD error signal is not computed, which means the cost of the action sequence, $C(s,a,a')$, is not updated after it is formed. This implies that a sequence should only form after the agent is certain about the estimated costs and benefits of the sequence; otherwise, the agent could stick to a sub-optimal sequence for a long period of time. This implies that action sequences should not form during early stages of instrumental learning because a high degree of certainty requires sufficient experience of the environment and hence more training time. In the current example, we did not model the agent's certainty about its estimations. Instead we assumed a large initial value for the cost of sequence formation, and chose a slow learning rate for that cost η_C, something that ensures sequences form only after the environment is well learned.

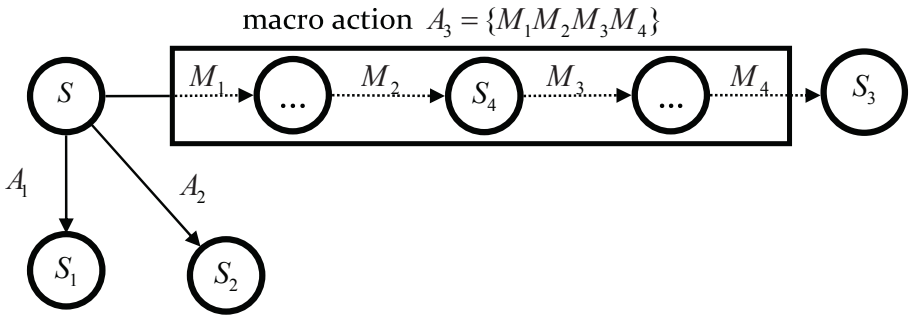


Figure 3.3 – At state S, three actions are available: A_1, A_2, A_3, where A_1 and A_2 are primitive actions, and A_3 is a macro action composed of primitive actions $M_1 \dots M_4$. If at state S, the macro action is selected for execution, the action control transfers to the sequence-based controller, and actions $M_1 \dots M_4$ become executed. After the termination of the macro action control returns back to model-based decision-making at state S_3.

3.4 A hierarchical model-based architecture

Assume that the agent is in state S in which several choices are available (Figure 3.3), two of which are primitive actions (A_1 and A_2), and one of which is a macro action (A_3). In state S the agent uses model-based action evaluation, and selects one of the available actions for execution, which can be either a primitive action or a macro action. After completion of a primitive action the agent enters a new state in which it again uses a model-based evaluation for selecting subsequent actions. However, if the selected action is the macro action, A_3, its execution is composed of taking a sequence of primitive actions ($M1 \dots M4$). Nevertheless, upon completion of the macro action, the agent identifies its new state (S_3), and uses model-based action evaluation again for selection of the next action.

The above scenario involves hierarchical model-based and sequence-based action control. At the choice points, actions are selected based on the model-based evaluations. However, during the execution of a sequence of actions, they are selected based on their sequential order, without going through the evaluation process. As discussed in the previous section, this sequence-based action selection can lead to higher average reward rates, in comparison to a pure model-based decision-making system, which is the benefit of sequence-based action selection. However, it can lead to a maladaptive behavior if the environment changes after

action sequences have formed. For example, assume that after the macro action $\{aa'\}$ has formed, the environment changes so that the execution of action a' after action a no longer satisfies the cost–benefit analysis presented in the previous section - say the change causes the value of the state to which action a' leads to decrease significantly - as a consequence, taking action a' after action a will no longer be the best choice. If action control is sequence-based, it is the previous action that determines the next action and not the consequences of the action. Hence, the agent will continue to take action a' even though it is not the most appropriate action.

Ultimately, in this situation, we expect the macro action to decompose to its components so that the agent can consider other alternative actions other than action a'. However, this does not happen instantly after the environment has changed and, thus at least for a while, the agent will continue behaving maladaptively. As mentioned in the previous section, after the macro action has formed, its cost, $C(s,a,a')$, is not updated, because the system is working on the open-loop mode and the TD error signal is not computed to update $C(s,a,a')$. As a consequence, the cost side of the sequence formation trade-off is relatively insensitive to environmental changes. The other side of the trade-off, $\bar{R}\tau$, however, is sensitive to environmental changes: if the environment changes so that executing the macro action leads to a decrease in the average reward the agent experiences, $\bar{R}$, then this circumstance motivates decomposition of the macro. In fact, this model predicts that, if the environmental changes do not alter the average reward, then the agent will continue to take action a' after a, even if the change introduces better alternatives other than taking action a'. Nevertheless, if the change causes a decrease in the average reward, then the macro action will decompose, and the responses will adapt to the new situation. However, this cannot happen instantly after the change because it takes several trials before the average reward adapts to the new condition.

The above feature of the model implies different sensitivity of sequence-based responses of the model after an environmental change compared with the situation where responses are under model-based control. As an example, in Figure 3.3 assume that the action A_1 is the best action, i.e. it has the highest Q-value among actions A_1, A_2 and A_3, and so the agent takes action A_1

more frequently than the others. Now, assume that the environment changes, and the value of the state that action A_1 leads to (state S_1) dramatically decreases. The next time that the agent is making a decision in state S, it evaluates the consequences of action A_1 using equation 3.4, and finds out that A_1 is no longer the best action, and adapts its behavior instantly to the new conditions. Evidence for the effect of this type of environmental change on the behavior comes from an experiment (Ostlund, Winterbauer, & Balleine, 2009) in which rats were trained on two action sequences for two outcomes, i.e. R1→R2→O1 and R2→R1→O2. After this training, either O1 or O2 was devalued, and performance of the two sequences (macro actions {R1R2} and {R2R1}) were assessed in extinction. Results show that the performance of the sequence that leads to the devalued outcome decreases, which implies that performance of a macro action (e.g. {R1 R2}) is immediately sensitive to the value of the states to which it leads (O1).

Compare the above situation with one in which an environmental change causes a decrease in the value of one of the states visited during a sequence, for example state S_4. Here, a change in the value of state S_4 will not affect the value of action A_3 (when evaluated at S). In fact, one effect of chunking actions together and turning them into a single response unit is that the representation of the action sequence is independent of its embedded individual actions and their outcomes. Although this higher level representation of actions makes decision-making easier and faster, it also renders the evaluation of action sequences insensitive both to offline changes in individual action–outcome contingencies and to changes in the value of any outcomes (states) delivered within the sequence boundaries (section 2.4.5). As such, although the selection of an action other than A_3 would be more optimal, A_3 will be selected at state S. After several trials, because taking action M_2 does not lead to reward, the average reward that the agent experiences decreases, and the sequence should then decompose into its elements. At this point, the control of actions will return to the model-based system and choices adapt to the new environmental conditions. In the next section we show that insensitivity to reinforcer devaluation and contingency degradation is due to this type of environmental change.

3.5 Simulations

Having described the model, we are now in a position to establish whether it can provide an accurate account of: (i) sequence learning, such as that observed in SRTT; and (ii) instrumental conditioning, particularly the shift in sensitivity of instrumental actions to reinforcer devaluation and contingency degradation during the course of overtraining (see section 3.8 for implementation details).

3.5.1 Sequential and random trials of sequence learning

As already noted, when a sequence of stimuli is predictable, such as in the sequential trials of the SRTT, along with the progress of learning as a result of sequence-based action selection, reaction times decline. In contrast, when the sequence of stimuli is random, as it is in the random trials condition of SRTT, reaction times do not decrease substantially during the course of learning. Here, we simulated the model described previously in a task similar to SRTT. After each correct button press the model receives one unit of reward. In the sequential trials condition, after each correct button press the next stimulus in a fixed sequence is presented, otherwise the sequence restarts and the first stimulus is presented. Here, we assume that the order of stimuli is S_0 to S_3, where the correct button press in S_0 is B_0, B_1 in S_1, etc. In the random trials condition, the next stimulus is selected randomly. It is assumed that it takes 400 milliseconds (ms) to make a decision using the model-based method, and 100 ms to elicit a response under sequence-based action control.

The temporal dynamics of sequence formation is depicted in Figure 3.4. After several learning trials, the agent learns the model of the environment (the rewards in states, delays in states and the consequences of each action) and, as a consequence, the probability of taking the correct actions increases, which implies that the agent gains more rewards and, thus, the average reward that the agent receives, $\bar{R}$, increases. This increase in the average reward implies that a significant number of the rewards that could have been gained in the future are being lost due to time taken for model-based action selection, and this favors the transition of

action control to the sequence-based method, which is faster. At the same time, the cost of sequence-based action selection, $C(S_0, B_0, B_1)$ decreases (Figure 3.4a), which means that the agent has learned B_1 is always the action that should be taken after B_0. Eventually, the benefit of sequence-based action selection becomes larger than its cost and, at that stage, the macro action $\{B_0 B_1\}$ replaces the B_0 action (Figure 3.4b). Later, actions B_2 and B_3 form the macro action $\{B_2 B_3\}$ and, finally, the two previously formed macro actions concatenate, and the macro action $\{B_0 B_1 B_2 B_3\}$ is formed. In addition, as shown in Figure 3.4a, after a macro action is formed the average reward that the agent gains increases due to faster decision-making.

Figure 3.4c shows the reaction times. As the figure shows, by forming new action sequences, reaction times decrease up to the point that only selection of the action sequence is based on model-based action control, and all subsequent button presses during the sequence are based on sequence-based action control. Figure 3.4c also shows the reaction times in the case of random trials, which remain constant largely because the sequence of stimuli is not predictable and the cost of sequence-based action selection remains high so that no action sequence forms (Figure 3.4d).

3.5.2 Instrumental conditioning

In this section, we aimed to validate the model in instrumental conditioning paradigms. Figure 3.5a depicts a formal representation of a simple instrumental conditioning task. The task starts in state S_0 and the agent has two options: the press lever (PL) action; and enter magazine (EM) action. By taking action PL, and then action EM, the agent enters state S_0 in which it receives one unit of reward ($r = 1$). All other actions, for example taking action EM before action PL, leads to no reward. Entering state S_1 is cued for example by a 'click' produced by the pellet dispenser, or 'buzz' of the pump if sucrose solution is the reward. After several learning trials, the agent learns the model of the environment; the value of the PL action exceeds the value of action EM, and the probability of taking action PL increases. As the probability of taking action PL increases, the agent gains more rewards and, hence, the average reward $\bar{R}$ increases. Simultaneously, the cost of sequence-based action selection

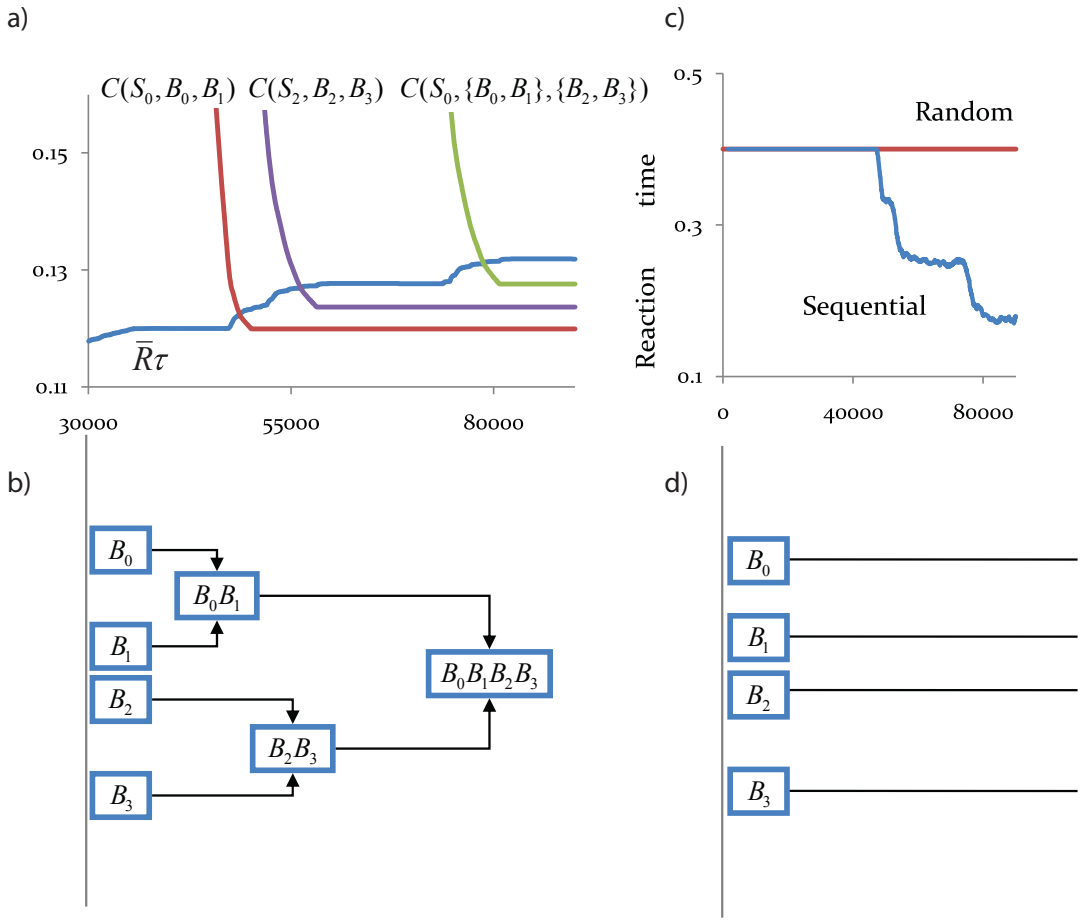


Figure 3.4 – The dynamics of sequence learning in sequential and random trials of SRTT. (a, b) As the learning progresses, the average reward that the agent gains increases, indicative of a high cost of waiting for model-based action selection. At the same time, the cost of sequence-based action selection decreases (top panel), which means that the agent has discovered the correct action sequences. Whenever the cost becomes less than the benefit, a new action sequence forms (bottom panel). The abscissa axis shows the number of action selections. (c) Reaction times decrease in sequential trials as a result of sequence formation but they remain constant in the random trials of SRTT because, (d) no action sequence forms. Data reported are means over 10 runs.

decreases, which means the agent has learned that action PL is always the action that should be taken after the EM action and, as a consequence, the macro action $\{EM, PL\}$ replaces the EM action (Figure 3.6a). From that point, the action PL is always taken after EM. In addition to the $\{EM, PL\}$ action sequence, the other action sequence, i.e., $\{PL, EM\}$ is also presumably being formed by over-training, which together with $\{EM, PL\}$ will eventually form a cyclic pattern of EM actions followed by PL actions. In the following section we will show that the $\{EM, PL\}$ part of the cycle plays an important role in behavioral data observed after over-training, and therefore we will focus on this component in the future sections.

The schematic representation in Figure 3.5a corresponds to a continuous reinforcement schedule, in which each lever press is followed by a reinforcer delivery (e.g. (C. D. Adams, 1982)). However, in most experimental settings, animals are required to execute a number of lever presses (in the case of ratio schedules), or press the lever after an amount of time has passed since the previous reward delivery (in the case of interval schedules) in order to obtain reward. One approach to analysing these kinds of experiments using the paradigm illustrated in Figure 3.5a is through application of the 'response unit hypothesis' (Skinner, 1938; Mowrer & Jones, 1945), according to which the total set of lever presses required for reward delivery is considered as a single unit of response. For example, in the case of fixed ratio schedules, if 10 lever presses are required to produce reinforcement, the whole 10 lever presses are considered as a single unit, corresponding to the action PL in Figure 3.5a. In ratio schedules this hypothesis is supported by the observation that, early in learning, the animal frequently takes the EM action, which, with the progress of learning, tends to occur only after the last lever press (Denny, Wells, & Maatsch, 1957; Overmann & Denny, 1974; Platt & Day, 1979).

Following this training, animals are given an extinction test, in which rewards are not delivered. During the course of extinction, the response unit mainly preserves its form and only the last response is likely to be followed by the EM action (Denny et al., 1957). Further, the average number of executed response units in extinction is independent of the number of lever presses required for reinforcer delivery, indicating that the whole response unit is being extinguished,

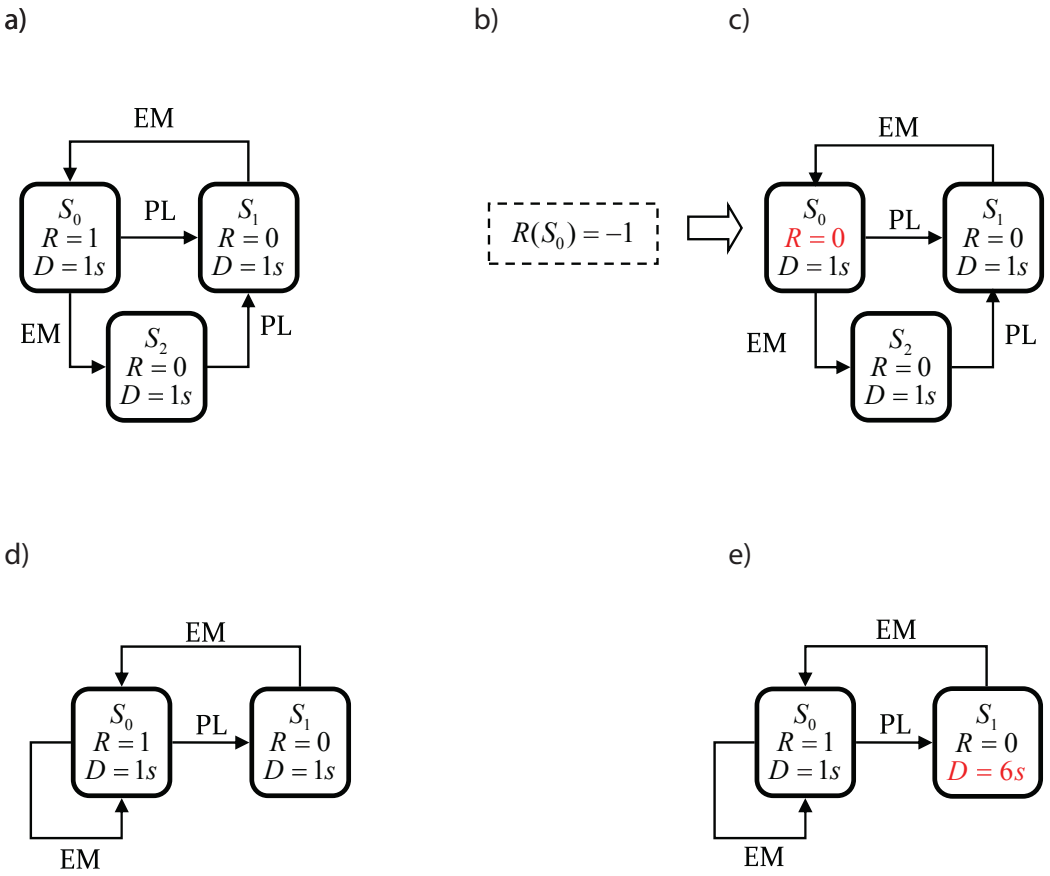


Figure 3.5 – Formal representation of instrumental conditioning tasks. (a) Instrumental learning: by taking the press lever (PL) action, and then enter magazine (EM) action, the agent earns a reward of magnitude one. By taking EM action in state S_2 and PL action in state S_1, the agent remains in the same state (not shown in the figure). (b, c) Reinforcer devaluation: the agent learns that the reward at state S_0 is devalued, and then is tested in extinction in which no reward is delivered. (d) Non-contingent training: unlike as in (a), reward is not contingent on the PL action, and the agent can gain the reward only by entering the magazine. (e) Omission training: taking the PL action causes a delay in the reward delivery, and the agent should wait 6 seconds before it can gain the reward by entering the magazine.

and not individual lever presses (Denny et al., 1957; Overmann & Denny, 1974; Platt & Day, 1979). In the case of interval schedules, because reinforcement of a lever press depends on the time that has passed since the previous reward delivery, the 'response unit hypothesis' has to be generalized to temporal response unit structures in which the animal continues to lever press for a certain amount of time (instead of for a certain number of times), in the form of 'bouts' of lever pressing (Shull & Grimes, 2003), or nose poking (Shull, Gaynor, & Grimes, 2002). In fact, in the previous section, in the course of analysing SRTT, we applied the 'response unit hypothesis' by treating the action of pressing the button as a single response unit, which of course can be broken into smaller units. Similarly, in the case of maze learning, the action of reaching the choice point from starting point can be thought as a sequence of steps. Is the formation of such response units (e.g. PL action composed of homogenous set of responses) through the action sequence formation method proposed in the previous section, or do they form in a level different from that in which macro action $\{EM, PL\}$ forms? We leave the answer to this question for future works.

In the next two sections, we investigated the behavior of the model, based on Figure 3.5a, when an environmental change occurs both before and after sequence formation.

3.5.2.1 Reinforcer devaluation before vs. after action sequence formation

As described above, in reinforcer devaluation studies, the value of the outcome of an action is reduced offline through some treatment (such as specific satiety or taste aversion learning) and the performance of the action subsequently assessed in extinction. There is a considerable literature demonstrating that, with moderate training, instrumental performance is sensitive to this change on value, whereas after more extended training it is not (cf. Figure 2.4).

To assess the accuracy of the model it was simulated using the procedure depicted in Figure 3.5. The procedure has three steps. The first step (Figure 3.5a) is the instrumental learning phase, described in the previous section. The next step (Figure 3.5b) models the devaluation phase in which the model learns that the reward obtained in state S_0 is devalued ($r = -1$). The third

phase is the test conducted under extinction conditions, i.e. reward is not delivered in state S_0 ($r = 0$). The critical question here is whether the model chooses action PL in state S_0. As noted previously, experimental evidence shows that after moderate training, the agent chooses action PL, whereas after extended training it does not. Figure 3.6b shows the probability of taking action PL after moderate training (in this case after 3000 action selections). As the figure shows, because action selection is under model-based control, when the reward in state S_0 is devalued, the value of taking PL action in state S_0 is immediately affected and, as such, the probability of taking action PL decreases.

The same is not true of overtrained actions. Figure 3.6c shows the sensitivity of responses to devaluation after extended training (9000 action selections). At this point the action sequence $\{EM, PL\}$ has been formed (Figure 3.6a) and, as the figure shows, unlike moderate training the agent continues taking action PL after reinforcer devaluation. This is because action selection is under sequence-based action control, and the outcome is delivered in the middle of action sequence $\{EM, PL\}$, which implies that the value of the action sequence will remain unchanged after the offline devaluation of the value of the outcome, and therefore, the agent will continue selecting action sequence $\{EM, PL\}$. After several learning trials, because the experiment is conducted in extinction and no reward is received, the average reward decreases, which means deciding faster is not beneficial, and causes the macro action $\{EM, PL\}$ to decompose to action EM and action PL. At this point, behavioral control should return to the model-based system, and the probability of taking action PL should adjust to the new conditions induced by devaluation.

3.5.2.2 Contingency degradation before vs. after action sequence formation

In section 2.4.1 we pointed out that habits are not just insensitive to reinforcer devaluation but also to the effects of degrading the instrumental contingency. A good example of this is the failure of habits to adjust to the imposition of an omission schedule, as shown in figure 2.4B. Having learned that lever pressing delivers food, the omission schedule reverses that relationship such that food becomes freely available without needing to lever press

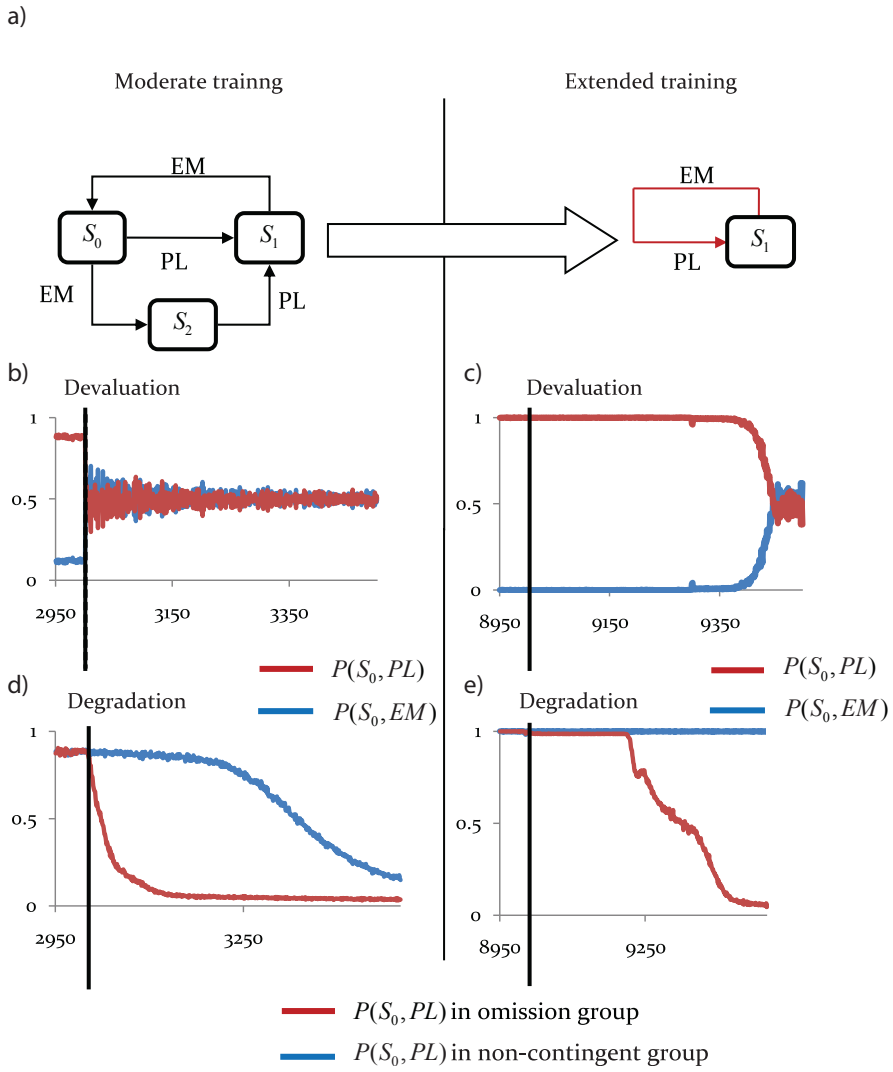


Figure 3.6 – Sensitivity of the model to reinforcer devaluation and contingency manipulations before and after sequence formation. (a) In the moderate training condition, actions are selected based on the model-based evaluation (left panel) but, after extended training, the selection of the press lever (PL) action is potentiated by its previous action [here enter magazine (EM); right panel]. (b) After the devaluation phase (shown by the solid-line), the probability of pressing the lever decreases instantly if the model is moderately trained. The abscissa axis shows the number of action selections. (c) After the devaluation phase behavior does not adapt until the action sequence decomposes and control returns to the model-based method. (d) In a moderately trained model the probability of selecting action PL starts to decrease in the contingency degradation condition, although the rate of decrease is greater in the case of omission training. (e) When training is extensive, behavior does not adjust and the non-contingent and omission groups perform at the same rate until the sequence decomposes. Data reported are means over 3000 runs.

to receive it. However, in the experimental group, lever pressing delays free food delivery; hence, the rats now have to learn to stop lever pressing to get the reward. The ability to stop responding in the omission group is compared with rats given exposure to a zero contingency between lever pressing and reward delivery (the non-contingent control). As shown previously (e.g. (Dickinson et al., 1998); Figure 2.4B) in this situation, rats who are given moderate instrumental training are able to withhold their lever press action to get food; the omission group responds less than the control group when the omission schedule is introduced. When lever pressing has been overtrained, however, the rats are insensitive to omission and cannot withhold their lever press responses compared with the control group.

For the simulation of non-contingent reward delivery in the control condition, the model underwent the instrumental conditioning procedure described in the previous section (Figure 3.5a), and then the model was simulated in the task depicted in Figure 3.5d. The agent can obtain reward by taking action EM without performing action PL, and hence reward delivery is no longer contingent upon action PL. For the simulation of the omission schedule, after instrumental training (Figure 3.5a), the agent was exposed to the schedule depicted in Figure 3.5e. The difference between this schedule and the non-contingent schedule is that, after pressing the lever, the model must wait 6 seconds before obtaining the reward by taking action EM, which models the fact that rewards are delayed if the animal chooses action PL under the omission schedule. The behavior of the model after moderate training is depicted in Figure 3.6d. As the figure shows, after the introduction of the degradation procedure, the probability of taking action PL starts to decrease. The rate of decrease is faster in the case of the omission schedule, in comparison to the non-contingent schedule. This is because the final value of action PL in the omission condition is lower than the final value of the non-contingent condition and, therefore, the values adjust faster in the omission case.

Figure 3.6e shows the effect of omission and degradation after more extended training when action selection is under sequence-based control. As the figure shows, in both the control and omission conditions, the probability of selecting the action fails to adjust to the new conditions and the agent continues selecting action PL. However, in the omission condition, because

rewards are delayed as a result of pressing the lever, the average reward starts to decrease and, after sufficient omission training, the action sequence decomposes, and behavior starts to adapt to the new conditions. In the case of non-contingent reward delivery, because the average reward remains constant, the model predicts that the agent will continue pressing the lever, even after extended exposure to the non-contingent schedule. This prediction has not been assessed in the literature largely because exposure to non-contingent reward tends to have a generally suppressive effect on performance, as competition from EM responding increases and action PL begins to extinguish.

Although somewhat idealized relative to the effects observed in real animals, it should be clear that, in contrast to simple RL, sequence learning and habitual actions are both accurately modeled by this mixed model-based and sequence-based architecture. The implications of this model for the behavior of real animals and for theories of goal-directed and habitual action control are described below.

3.6 Discussion

A number of investigators have noted the apparently competitive nature of these forms of action control, and some have suggested that these processes may compete for access to the motor system. Using this general approach, previous computational accounts have successfully explained the effect of reinforcer devaluation on instrumental responses in different stages of learning (Daw et al., 2005; Keramati et al., 2011), the effect of habit formation on reaction times (Keramati et al., 2011), and the effect of the complexity of state identification on the behavioral control (Shah & Barto, 2009). All these approaches shares a common *flat* architecture in which the goal-directed and the habitual systems work in parallel at the same level, and utilize a third mechanism, called an arbitration mechanism (section 2.4.3), to decide whether the next action will be controlled by the goal-directed or the habit process. They differ from the hierarchical architecture used here in which the goal-directed system stands at a higher level, with the role of the habit process limited to efficiently implementing decisions

made by the goal-directed system in the form of macro actions. This structural difference itself raises some important behavioral predictions. For example, in the hierarchical structure, the goal-directed system treats macro actions as integrated units, and thus the action evaluation process is blind to the change in the value of states visited during execution of the macro action. Thus, in Figure 3.3, goal-directed action evaluation in state S depends only on the value of states S_1, S_2, S_3 and the total reward obtained through executing the macro action. As a consequence, changing the value of state S_4 has no immediate impact on the decisions made at state S. In contrast, in a flat architecture, goal-directed action selection is conducted by searching all the consequences of possible actions and, thus, a change in the value of state S_4 should immediately affect action selection in state S.

Another prediction of a sequence-based conception of habits is that, if an agent starts decision-making in a state in which an action sequence has been learned (state S), it will immediately show habit-like behavior, such as insensitivity to outcome devaluation. In contrast, if it starts decision-making somewhere in the middle of a sequence (e.g. in the test phase the task starts in an intermediary state such as S_4), it will not show habit-like behavior. This is because in the current conception of the model, an action sequence can be launched only in the state in which is has been learned.

The problem of mixed closed-loop and open-loop decision-making has been previously addressed and discussed in the literature, but the solutions proposed differ from that suggested here. In the model proposed here, the inputs for the mixed architecture come from fast action control in the open-loop mode, whereas in previous work they have come from a cost associated with sensing the environment (Hansen, Barto, & Zilberstein, 1996), or the complexity of modeling the environment (Kolter, Plagemann, Jackson, Ng, & Thrun, 2010). From a control point of view, in most situations (especially stochastic environments), open-loop action control is considered to be inappropriate. As such, in hierarchical approaches to RL, the idea of macro actions is usually generalized to closed-loop action control (Barto & Mahadevan, 2003).

Behavioral and neural signatures of this generalized notion have been found in previous studies (Haruno & Kawato, 2006; Botvinick, 2008; Botvinick et al., 2009; Badre & Frank, 2012; Ribas-Fernandes et al., 2011). Here, instead of using this generalized hierarchical RL, we utilized a hierarchical approach with a mixed open-loop and closed-loop architecture that allows us to incorporate the role of action sequences in habit-learning. Further, to examine the potential neural substrates of this architecture, we developed a method based on the TD error for learning macro actions, though various alternative methods have also been proposed for this purpose (Korf, 1985; Iba, 1989; Randløv, 1998; Mcgovern, 2002).

With regard specifically to sequences, the effect of overtraining on reaction times has been addressed in instrumental conditioning models (Keramati et al., 2011), which often interprets them as the result of transition to habitual control, which, because it is fast, results in a reduction in reaction times. However, that model predicts a decrease in reaction times in both random and sequential trials of SRTT. This is because, on that approach, a habit forms whenever the values of the available choices are significantly different, irrespective of whether the sequence of states is predictable. As such, because in both sequential and random trials of SRTT the values of correct and incorrect responses are different, the model predicts that habits will form and reaction times decrease in both cases, which is not consistent with the evidence. In the sequence-learning literature, the issue of learning sequences of responses and stimuli using TD error has been addressed (Berns & Sejnowski, 1998; Bapi & Doya, 2001; Nakahara et al., 2001; Bissmarck, Nakahara, Doya, & Hikosaka, 2008). However, because these models are generally developed for the purpose of visuo-motor sequence learning, it is not straightforward to apply them to instrumental conditioning tasks. Likewise, the effect of the predictability of stimuli (i.e. random vs. sequential trials in the SRTT) on reaction times is not directly addressed in these models, which makes it hard to compare them in SRTT.

Based on a number of studies, it appears that the acquisition of goal-directed actions is controlled by a circuit involving the medial prefrontal cortex and dorsomedial striatum (DMS). For example in one study animals were trained to take two different actions (two different levers) to earn two different outcomes. A test conducted after the devaluation of one of the

outcomes, revealed that the control animals (with intact DMS) showed a preference for the action which its outcome was not devalued, but the animals in which DMS was inactivated selected both levers equally. Here an important question is how to interpret the lack of insensitivity to outcome devaluation in animals with an inactivated DMS. One interpretation can be that the inactivation of DMS resulted in the emergence of habits and such responses are the result of the automatic action selection. However, this interpretation is not entirely consistent with the framework proposed in this chapter, because in such experiments animals are generally moderately trained, and therefore, the action sequences still won't have had a chance to develop. Another interpretation of such results is that the equal preference for both actions observed during the test is actually the result of the random responses generated by the disrupted goal-directed process (as a result of inactivating DMS), which cannot evaluate and select actions appropriately. This latter interpretation is more consistent with the framework suggested in this chapter, in which all the actions are dependent on the goal-directed processes.

There is some evidence from previous studies indicating that although over-training makes lever presses insensitive to changes in the outcome values, magazine entries remain sensitive to the changes of the value of outcomes (Killcross & Coutureau, 2003) (although this effect is not always observed). That is, animals for which the outcome has been devalued press the lever at the same rate as the animals for which the outcome is has not been devalued, but they enter the magazine less than the other group. This observation seems inconsistent with the framework proposed here, since here it is suggested that animals perform a sequence of magazine entries followed by lever presses. It is important to recognize that each initial magazine entry is usually followed by further several magazine checks, i.e., animals check the magazine multiple times before taking the next press. Such extra magazine entries can be guided by Pavlovian processes, which are presumably sensitive to the changes in the value of outcomes. As such, the observed decrement in magazine entries after outcome devaluation can be because of the elimination of the extra magazine entries that animals make after the first magazine entry. That is, the proposed action sequence is in place and animals take an

EM action after each *LP* action, however, the *EM* action in non-devalued animals consists of several magazine entries, while in the devalued animals it consists of a lower number of magazine entries, which is consistent with the proposed framework and the above observation.

A recent study has investigated the role of dorsolateral striatum (DLS) and infralimbic cortex in automatic actions measured by both the development of action sequences and the insensitivity of actions to changes in outcome values (Smith & Graybiel, 2013, 2014). Animals were trained in a T-maze, in which at the end of each arm a separate outcome (O1 and O2; chocolate milk and sucrose solution in this experiment) was delivered to the them. At the start of each trial there was a cue which told animals which arm they needed to enter in order to earn a reward (the outcome specific to that arm). The study included two groups of animals: one group of animals (CT) were trained moderately and only until they reached the criterion of statistically significant performance, and the other group (OT) were trained for at least ten additional sessions. Animals were then given a post-devaluation test, in which one of the outcomes was devalued. Results showed that animals in the OT group kept running to the arm that would have delivered the devalued outcome, while animals in the CT group significantly reduced their runnings to the arm which would have delivered the devalued outcome. Neuronal recording showed that task-bracketing activity in DLS, that marked the beginning and end of each trial, was emerged in the OT group, and also interestingly in the CT group that according to the behavioral results showed sensitivity to outcome devaluation. Such task-bracketing activity, therefore, was unrelated to the sensitivity to outcome devaluation, however it was negatively correlated with deliberative head movements at the choice point of the maze (at the junction in which the animals needed to select an arm to run to). That is, higher DLS activity at the beginning of each trial marked less amounts of deliberative head movements. Such deliberative head movements, however, were not related to the sensitivity or insensitivity to outcome values. Finally, The task-bracketing DLS activity once established, persisted even after delivering the devalued outcome, which was accompanied by marked behavioral changes.

In the above study the disassociation between deliberative head movements and sensitivity

to outcome values can be understood within the hierarchical model-based RL framework proposed here: the goal-directed process evaluates both O1 and O2 and chooses to run the action sequence that leads to the non-devalued outcome, which implies sensitivity to outcome values but lack of deliberative head movements (since actions will be selected according to the sequence). As such one would expect to see the bracketing activity in DLS and at the same time sensitivity to outcome values, as observed in the CT group. As a result of over-training, according to the theory, insensitivity to outcome values developed in the OT group because the boundaries of action sequences had expanded to include the outcome of each action sequence, which made the evaluation of action sequences insensitive to changes in outcome values.

One important restriction of the proposed model relates to the fact that it may seem implausible to assume that, after an action sequence has been formed, the individual actions are always executed together. To address this issue it can, for example, be assumed that, occasionally, the model-based controller interrupts the execution of actions during the performance of an action sequence in order to facilitate learning new action sequences. Along the same lines, it can be assumed that action sequences do not replace primitive actions (as proposed in this paper), but are added as new actions to the list of available actions (see section 2.4.5 for a discussion). Finally, investigating why some kinds of reinforcement schedules lead to habits (action sequences) whilst others do not, is an interesting issue and each of these will be addressed in future work.

3.7 Summary

In this chapter, we presented a new computational model for hierarchical goal-directed decision-making, and developed a potential role for dopamine in learning action sequences. Furthermore, we showed that this model can explain behavioral properties that have been previously attributed to model-free RL. In the next chapter, we focus on the exclusive properties of the proposed hierarchical account that model-free RL is unable to explain, and compare

these two alternative account of automatic actions more directly.

3.8 Appendix

We aim to estimate $C(s,a,a')$ by the samples of the error signal experienced by taking action a', after action a in state s. $C(s,a,a')$ is defined as:

$$C(s,a,a') = E[\delta(s',a')|s,a] \tag{3.9}$$

we maintain that:

$$C(s,a,a') = E\left[\frac{P(a'|s,a)}{P(a'|s')}\delta(s',a')|s,a,a'\right] \tag{3.10}$$

this follows from:

$$\begin{aligned} C(s,a,a') =& E\left[\frac{P(a'|s,a)}{P(a'|s')}\delta(s',a')|s,a,a'\right] \\ &= \sum_{s'}\frac{P(a'|s,a)P(s'|s,a,a')}{P(a'|s')}\delta(s',a') \\ &= \sum_{s'}P(s'|s,a)\delta(s',a') \\ &= E[\delta(s',a')|s,a] \end{aligned} \tag{3.11}$$

$P(a'|s,a)$ and $P(a'|s')$ can be estimated directly by counting number of times action a' has been taken after s, a and after s', respectively. Given these, $C(s,a,a')$ can be estimated by averaging over the samples of the error signal multiplied by the factor α:

$$\alpha = \frac{P(a'|s,a)}{P(a'|s')} \tag{3.12}$$

α is in fact the percentage of taking action a' after s and a is due to being in state s'. If action a' is not available in state s', we assume this factor is infinity. Given α, the cost function is estimated using equation 3.8.

The implementation of the mixed architecture is similar to model-based hierarchical reinforcement learning (RL). After execution of a macro action finished, the characteristics of the underlying semi-Markov Decision Process (SMDP) are updated. That is, the total reward obtained through executing the macro action, and the total time spent for executing the macro action, are used for updating the reward and transition delay functions. Here, because we assumed that reward function depends on the states, and not state–action pairs, the total reward obtained through the macro action cannot be assigned to the state in which the macro action was launched. This is because the total reward obtained through the macro action can be different from the reward of the state. For addressing this problem, when an action sequence was formed, a temporary extended auxiliary state is added to the SMDP, to which the agent enters when the macro action starts, and exits when the macro action finished. The transition delay and reward function of this auxiliary state is updated using the rewards obtained through executing the macro action, and the time spent to take the macro action.

Because according to the Bellman equation for average reward semi-Markov RL, the Q-values satisfy equation 3.4 (Puterman, 1994), when a non-exploratory action is taken (the action with highest value is taken), we update the average reward as follows:

$$\bar{R} \leftarrow (1-\sigma)\bar{R} + \sigma \left[\frac{r + V(s') - V(s)}{d} \right] \tag{3.13}$$

where σ is the learning rate of the average reward. For the action selection (in the model-based system), soft-max rule is used:

$$P(s|a) = \frac{e^{\beta Q(s,a)}}{\sum_{a'} e^{\beta Q(s,a')}} \tag{3.14}$$

where parameter β determines the rate of exploration. For computing model-based values, a tree-search algorithm was applied with the depth of search of three levels. After this level, goal-directed estimations are replaced by model-free estimations.

Due to the fluctuations of $C(s,a,a')$ and $\bar{R}\tau$ at the point they meet, a series of sequence formation/decomposition may happen before the two curves become separated. To address this

Table 3.1 – Free parameters of the model, and their assigned values

Parameter	Value
Update rate of the reward function	0.05
Update rate of the average reward (σ)	0.002
Update rate of the cost of sequence-based control (η_C)	0.001
Initial value of the cost of sequence-based control	-2
Rate of exploration (β)	4

issue, an asymmetric rule for sequence decomposition is used, and a sequence decomposes only if $-C(s,a,a') > 1.6\bar{R}\tau$. Also, for simplicity, we assumed that macro actions could not be cyclic.

Internal parameters of the model are shown in Table 3.1.

4 Hierarchical decision-making in humans

In the previous chapter we showed that the outcome devaluation and contingency degradation experiments can be explained using the proposed model-based hierarchical RL. In this chapter, we develop a two-stage decision-making task in humans (based on the task developed by (Daw et al., 2011)), and then we make a direct comparison between model-free and hierarchical accounts of automatic actions.

4.1 Introduction

There is now considerable evidence from studies of instrumental conditioning in rats and humans that the performance of reward-related actions reflects the involvement of two learning processes, one controlling the acquisition of goal-directed actions and the other of habits (C. D. Adams, 1982; Dickinson et al., 1998; Dickinson, 1994; Balleine & O'Doherty, 2010). This evidence suggests that goal-directed decision-making involves deliberating over the consequences of alternative actions in order to predict their outcomes after which action selection is guided by the value of the predicted outcome of each action. In this respect, action evaluation relies on the representation of contingencies between actions and outcomes as well as the value of the outcomes, which in sum constitute a model of the environment. In contrast, habitual actions reflect the tendency of individuals to repeat behaviors that have

led to desirable outcomes in the past and respect neither their causal relationship to, nor the value of their consequences. As such, they are not guided by a model of the environment, and are relatively inflexible in the face of environmental changes (Daw et al., 2005; Keramati et al., 2011; Doya, 1999).

Although these features of goal-directed and habitual action are reasonably well accepted, the structure of habitual control, and the way in which it interacts with the goal-directed process in exerting that control, is not well understood. Two types of architecture have been proposed: a hierarchical architecture and a flat architecture. In chapter 3, we described a version of the hierarchical structure in the context of advancing a new theory of habits. Although habits are usually described as single step actions, their tendency to combine or chunk with other actions (Graybiel, 2008; Book, 1908; Lashley, 1951; Pew, 1966) and their insensitivity to changes in the value of, and the causal relationship to, their consequences (Balleine & O'Doherty, 2010; Dickinson & Balleine, 2002), suggests that they may best be viewed as action sequences. On this view habit sequences are represented independently of the individual actions and outcomes embedded in them such that the decision-maker treats the whole sequence of actions as a single response unit. As a consequence, the evaluation of action sequences is divorced from offline environmental changes in individual action-outcome contingencies or the value of outcomes inside the sequence boundaries and, as they are no longer guided by the model of the environment (chapter 3), are executed irrespective of the outcome of each individual action (Pew, 1966; Keele, 1968); i.e., the actions run off in an order predetermined by the sequence, without requiring immediate feedback.

On this hierarchical view, such action sequences are utilized by a global goal-directed system in order to efficiently reach its goals. This is achieved by learning the contingencies between action sequences and goals and assessing at each decision point whether there is a habit that can achieve that goal. If there is, it executes that habit after which control returns to the goal-directed system. In essence, the goal-directed system functions at a higher level and selects which habit should be executed whereas the role of habits is limited to the efficient implementation of the decisions made by the goal-directed process (Ostlund et al., 2009)

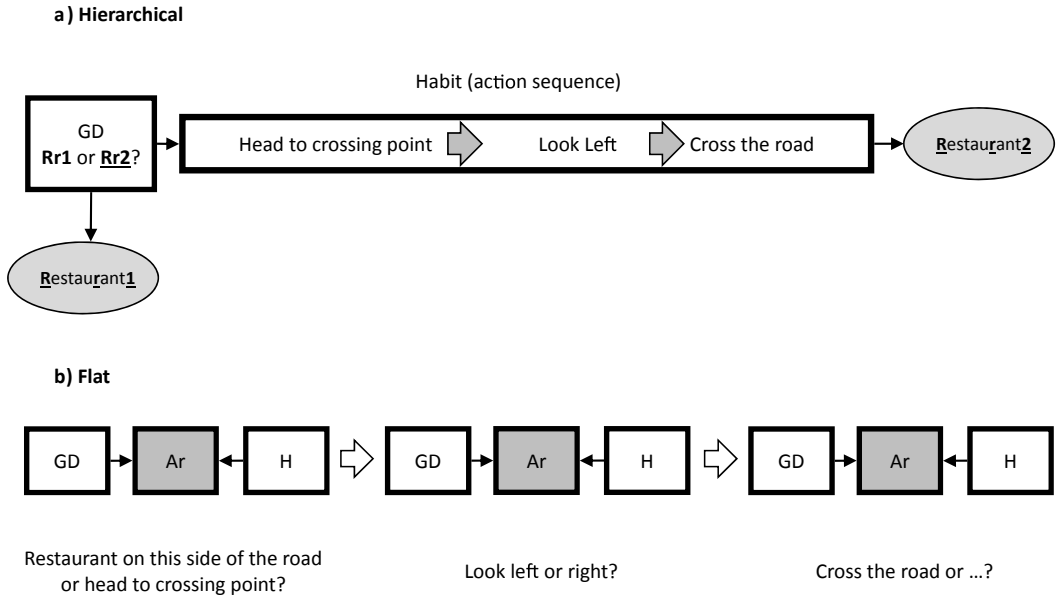


Figure 4.1 – An example illustrating the difference between the hierarchical and flat organizations. (a) Hierarchical interaction. The goal- directed system (GD) selects goals and decides whether to go to a restaurant on this side of the road (Rr1) or on the other side of the road (Rr2). If it chooses to go to the restaurant on the other side of the road, then it triggers the habit of crossing the road and control transfers to the habitual process. After execution of the habit finishes, control returns to the goal-directed system. (b) Flat interaction. At each decision point, the arbitration mechanism (Ar) decides whether the next action should be controlled by the goal-directed system or the habitual system (H).

(chapter 3) (see (Botvinick et al., 2009) for a review of other schemas).

Assume, for example, you are deciding whether to go to a restaurant on this side of the road or on the other side of the road (Figure 4.1a). The goal-directed system evaluates both options, and decides to go to the restaurant across the road. It thus triggers a 'crossing the road' habit, and transfers the control to the habitual system. The habit is an action sequence composed of several individual actions: (1) head to the crossing point, (2) look left, and (3) cross the road. Individual actions are executed one after another, and after they finish, the control transfers back to the goal-directed system to make the next decision such as, for example, choosing from the menu in the restaurant.

In contrast to the hierarchical architecture, the flat architecture treats habits as single step

actions rather than action sequences (e.g. (Daw et al., 2005)). At each step, an arbitration mechanism decides whether the next action should be controlled by the goal-directed system or the habitual system. In the context of the above example, at the beginning the arbitration mechanism selects one of the systems to decide whether to go to the restaurant on this side of the road or to the crossing point. Again, at the crossing point, the arbitration mechanism selects one of the systems to decide whether to look left, or right, and similarly at each future step the arbitration mechanism selects one of the systems to control behavior (Figure 4.1b). It should be clear, therefore, that, in the flat approach, both systems are at the same level and action evaluation happens in both processes; both systems evaluate available alternatives, and the arbitration mechanism determines how these two evaluations combine to make the final decision.

From the flat perspective, another difference between goal-directed and habitual processes lies in how they evaluate actions. The goal-directed process obeys the same principles sketched earlier: learning the model of the environment, and making predictions based on that model (model-based evaluation). In contrast, the habitual system is model-free and evaluates actions based on their 'cached' reward history without searching through the action-outcome contingencies (Daw et al., 2005; Doya, 1999).

More recently, Daw et al (Daw et al., 2011) have exploited the difference between model-free and model-based evaluation to investigate the interaction of goal-directed and habit processes in a flat structure reasoning that, because model-free evaluation is retrospective, chaining predictions backward across previous trials, and model-based evaluation is prospective, directly assessing available future possibilities, it is possible to distinguish the two using a sequential, multistage choice task. In this task subjects first make a binary choice (stage 1) then transition to stage 2 in which they make a second choice to earn a reward. The best choice at stage 2 varies depending on the first choice and, to maintain a constant trade-off between habitual and goal-directed systems, the reward probabilities in stage 2 are continually varied. By examining stage 1 choices, Daw et al were able to find evidence of mixed goal-directed and habitual predictions.

Here we show that stage 1 habitual actions, explained by the model-free evaluation in previous work, can also be explained by assuming that stage 1 actions chunk with stage 2 actions, reducing the source of habitual actions to the formation of action sequences. Based on this finding we next examined specific predictions of each account. With regard to the two-stage task, the flat account predicts that feedback received after the execution of an action will affect subsequent decisions and, therefore, that arbitration between goal-directed and habit controllers will recur anew at each stage. As a consequence, action-control at each stage of the task should be independently established; in particular it should be noted that action control in stage 2 should not depend on stage 1. In contrast, because our hierarchical account treats habits as action sequences, and because the execution of habits is open-loop (section 3.1), it predicts that, during the execution of a habit, actions will be executed one after another without considering feedback from the environment during the sequence and, therefore, that, when habitual, the action taken at stage 2 is already determined when starting the habit sequence at stage 1. We made two further predictions from the hierarchical account: first, because of their relative freedom from feedback, action sequences should be elicited more quickly than single actions predicting that, when habitual, reaction times between stage 1 and stage 2 actions will be faster than when non-habitual. Second, and based on these predictions, we anticipated that the hierarchical model would better fit the performance of subjects working on this two-stage task than the flat model.

4.2 Material and Methods

4.2.1 Participants and behavioral task

Fifteen English speaking subjects (seven females; eight males; mean age 23.8 years [SD 4.3]) completed a two-stage decision-making task. After a description of the study, written consent was obtained. This study was approved by the Sydney University Ethics Committee.

Each subject completed 270 trials, with a break after the first 120 trials (Figure 4.2). Each trial started with the presentation of a black square and subjects could choose between pressing

either 'Z' (using left hand) or '/' (using right hand). After pressing the key, a slot machine appeared on the screen, and the subject could make the next response, which would result in either a monetary reward or no reward. The outcome was shown for two seconds and after that an inter trial interval started and lasted for one second, after which the next trial began.

The probability of earning money at each choice was randomly set to either 0.2 or 0.7 at the beginning of the session, and in each trial, with the chance of 1/7, they were again randomly set to 0.2 or 0.7. This later step was to encourage searching for the best keys throughout the session.

Subjects were instructed that the chance of reaching each slot machine by pressing each key will not change throughout the task, but the goodness of the keys in terms of leading to rewards will change over time.

If a stage 1 action is the best action (the maximum probability of receiving reward on the keys of the slot machine that it commonly leads to is greater than the other action), and slot machines reset in the next trial, the probability that the action remains the best action is 3/16. Based on this, and given that probability of resetting is 1/7, the average number of trials for which a stage 1 action remains the best action is as follows:

$$\sum_{i=1\ldots\infty} i\,(6/7 + 1/7 \times 3/16)^{(i-1)}\,(13/16 \times 1/7) \approx 8.6 \tag{4.1}$$

The fact that a stage 1 action remains the best for a few numbers of trials ensures that reward-transition interaction does not emerge as the result of developing bias toward the best action.

4.2.2 Behavioral analysis

For all the analyses, we used R (R Core Team, 2012), and the R package lme4 (Bates & Maechler, 2009).

In the analysis presented in the section 4.3.1, we used mixed-effects logistic regression in which whether the previous first stage action is repeated was a dependent variable, and the

transition type (rare or common), and reward received in the previous trial were explanatory variables. We treated all the explanatory variables as random effects.

In the analysis in the section 4.3.2, staying or switching to the other stage 2 action is the dependent variable, and the reward received in the previous trial and staying on the stage 1 action were the explanatory variables. Only trials in which the stage 2 states were different from previous trials were included in this analysis. All the explanatory variables were used as random effects. In the second analysis of this section, staying on the same stage 2 action is dependent variable, and whether stage 2 state is the same, and whether previous trial was rewarded, are explanatory variables, and also random effects. Only trials in which stage 1 action is the same as the previous trial were included in this analysis. The third analysis is similar to the third one, except that trials in which stage 1 action is not the same as the previous trial are included in the analysis.

For analysis of the model behavior in the section 4.3.2, each model was simulated 3000 trials in the task with the best fitting parameters of each individual (see the section 4.2.3 below for more information). Then we analyzed data using linear mixed-effects regression in which the probability of selecting the same second stage action by the model was taken as the dependent variable, and the reward received in the previous trial and staying on the first stage actions were explanatory variables. The intercept was treated as the random effect, and reported p-values are MCMC-estimated using R package LanguageR (Baayen, 2011).

In the analysis in the first part of the section 4.3.3, staying on the same stage 2 action was a dependent variable, and the reaction time was an explanatory and random effect. Only trials in which the previous trial was rewarded (first analysis) or not rewarded (second analysis), the stage 1 action was repeated, and the stage 2 state was not the same, were included in this analysis.

In the second analysis of this section, we applied a recursive partitioning method by taking (i) whether the previous trial is rewarded, (ii) whether the same first stage action is being taken, and (iii) reaction time as covariates, and staying on the same second stage action

as response. We used R package 'party' (Hothorn, Hornik, & Zeileis, 2006) for the analysis which employs conditional inference trees for recursive partitioning. In short, the partitioning method works as follows: at each stage of partitioning the algorithm performs a significance test on independence between any of covariates and the response using permutation tests. If the hypothesis is rejected (in the current analysis p-value less than 0.05), it selects the covariate which has strongest association with the response, and performs a split on that covariate.

4.2.3 Computational modeling

4.2.3.1 Simulation environment

We assumed that the environment has five states; the initial state denoted by S_0, (the black screen in Figure 4.2), slot machine states denoted by S_1 and S_2, the reward state denoted by S_{Re} and no-reward state denoted by S_{NR}.

4.2.3.2 Model-based, model-free RL hybrid

For modeling the flat interaction, a family of hybrid models similar to the previous works was used (Daw et al., 2011; Gläscher et al., 2010; Otto, Gershman, et al., 2013). A model-based RL (Sutton & Barto, 1998) model was used for modeling goal-directed behavior; and a Q-learning model (Watkins, 1989) was used to model the habitual behavior. We assumed that actions $A1$ and $A2$ are available in states S_0, S_1 and S_2.

Model-based RL- we denote the transition function with $T(s'|a,s)$ which is the probability of reaching state s' after executing action a in state s. We assume that the transition function at stage 1 is fixed ($T(S_1|A_1,S_0)=0.7$ and $T(S_2|A_2,S_0)=0.7$) and it will not change during learning. For other states, after executing action a in state s and reaching state s', the transition function updates as follows:

$$\forall s'' \in \{S_{Re}, S_{NR}\} : T(s''|s,a) = \begin{cases} (1-\eta)T(s''|s,a)+\eta & : s' = s'' \\ (1-\eta)T(s''|s,a) & : s' \neq s'' \end{cases} \tag{4.2}$$

Where η ($0 < \eta < 1$) is the update rate of the state-action-state transitions.

We assumed that the reward at state S_{Re} is one ($R(S_{Re}) = 1$), and zero in all other states. Based on this, the goal-directed value of taking action a in state s is as follows:

$$\forall s \in \{S_0, S_1, S_2\} : V^G(s,a) = \sum_{s'} T(s'|s,a) V^G(s') \tag{4.3}$$

Where:

$$V^G(s) = \begin{cases} \max_a V^G(s,a) & : s \in \{S_0, S_1, S_2\} \\ R(s) & : s \in \{S_{Re}, S_{NR}\} \end{cases} \tag{4.4}$$

Model-free RL- After taking action a in state s, and reaching state s', model-free values update as follows:

$$Q^H(s,a) \leftarrow Q^H(s,a) + \alpha(V^H(s') - Q^H(s,a)) \tag{4.5}$$

Where α ($0 < \alpha < 1$) is the learning rate, which can be different in stage 1 and stage 2 actions. For stage 1 actions (actions executed in S_0), $\alpha = \alpha_1$, and for stage 2 actions $\alpha = \alpha_2$. Also

$$V^H(s) = \begin{cases} \max_a Q^H(s,a) & : s \in \{S_0, S_1, S_2\} \\ R(s) & : s \in \{S_{Re}, S_{NR}\} \end{cases} \tag{4.6}$$

In the trials in which the best action is executed in $s \in \{S_1, S_2\}$ the model-free value of the action executed in state S_0 also updates according to the outcome. If a was to be the action which was taken in S_0, a' the action taken in s, and s' the state visited after executing a', values update as follows:

$$Q^H(S_0,a) \leftarrow Q^H(S_0,a) + \alpha_1 \lambda(V^H(s') - Q^H(s,a')) \tag{4.7}$$

Where $\lambda(0 < \lambda < 1)$ is the reinforcement eligibility parameter, and determines how stage 1 action values are affected by receiving the outcome after executing stage 2 actions.

Final values are then computed by combining the values provided by the habitual and goal-directed processes:

$$V(s,a) = wV^G(s,a) + (1-w)Q^H(s,a) \tag{4.8}$$

Were $w(0 < w < 1)$ determines the relative contribution of habitual and goal-directed values into the final values.

Finally, the probability of selecting action a in state s will be determined according to the soft-max rule:

$$\pi(s,a) = \frac{e^{\beta(s)V(s,a)+\kappa(s,a)}}{\sum_{a'} e^{\beta(s)V(s,a')+\kappa(s,a')}} \tag{4.9}$$

Where $\kappa(s,a)$ is the action preservation parameter and captures the general tendency of taking the same action as the previous trial (H. Kim, Sul, Huh, Lee, & Jung, 2009; Lau & Glimcher, 2005). Assuming $s = S_0$ and a being the action taken in the previous trial in the S_0 state, then $k(s,a) = k$, otherwise it will be zero. The $\beta(s)$ parameter controls the rate of exploration, and $\beta(s) = \beta_1$ if $s = S_0$ and $\beta(s) = \beta_2$ if $s \in \{S_1, S_2\}$.

In the most general form, all the free parameters are included in the model: $\beta_1, \beta_2, \eta, \alpha_1, \lambda, k, w$ (we assumed that $\alpha_2 = \eta$). We generated eight simpler models by setting $\lambda = 0$, $\alpha_1 = \alpha_2$, and $\beta_2 = \beta_1$.

4.2.3.3 Hierarchical model-based, sequence-based RL

Implementation of the hierarchical structure is similar to hierarchical RL (Barto & Mahadevan, 2003; Dietterich, 2000; Sutton et al., 1999), with action sequences (A_1A_1, A_1A_2, etc) as options (Sutton et al., 1999). We assumed in state S_0, actions A_1, A_2, A_1A_1, A_1A_2, A_2A_2, and A_2A_1 are

available. In states S_1 and S_2, actions A_1 and A_2 are available. After reaching a terminal state (S_{Re} or S_{NR}), transition functions of both the action sequence, and the single action that led to that state update according to equation 4.2. In the case of single actions, the transition function will be updated by the $\eta = \eta_1$ update rate, and in the case of action sequences, the transition function will be updated by the $\eta = \eta_2$ update rate. Based on the learned transition function, value of action a in state s is calculated by the goal-directed system using equation 4.3.

The probability of selecting each action will be as follows:

$$\pi(s,a) = \frac{e^{\beta\omega(a)V^G(s,a)+\kappa(s,a)}}{\sum_{a'} e^{\beta\omega(a)V^G(s,a')+\kappa(s,a')}} \tag{4.10}$$

Where $\omega(a)$ determines the relative preference for single actions instead of executing action sequences. If action a is a single action $\omega(a) = w$, and if action a is an action sequence, $\omega(a) = 1 - w$. As before, $\kappa(s,a)$ captures action perseveration. We assumed that $\kappa(s,a) = k_1$ if action a is a single action, and $\kappa(s,a) = k_2$ if action a is an action sequence. $V^G(s,a)$ is calculated using Equation 4.3.

For calculating the probability of selecting actions in stage 2, given the first choice of the subject, we need to know whether that action is a part of an action sequence selected earlier, or is it under goal-directed control. Assume we know action A_1 has been executed in state S_0 by the subject, the probability of this action being due to performing the A_1A_2 action sequence is:

$$P(A_1A_2|S_0,A_1) = \frac{\pi(A_1A_2|S_0)}{\pi(A_1|S_0)+\pi(A_1A_2|S_0)+\pi(A_1A_1|S_0)} \tag{4.11}$$

Similarly, the probability of observing A_1 due to selecting the single action A_1 is:

$$P(A_1|S_0,A_1) = \frac{\pi(A_1|S_0)}{\pi(A_1|S_0)+\pi(A_1A_2|S_0)+\pi(A_1A_1|S_0)} \tag{4.12}$$

Based on this, the probability that the model assigns to action a in state $s \in \{S_1, S_2\}$, given that

action a' is being observed in S_0 is:

$$P(s,a) = P(a'|S_0,a')\pi(s|S) + P(a'a|S_0,a') \tag{4.13}$$

Where $P(a'a|S_0,a')$ and $P(a|S_0,a')$ are calculated using equations 4.11 and 4.12 respectively. In the most general form, all the free parameters are included in the model: β, η_1, η_2, k_1, k_2, w. We generated eight simpler models by setting $\eta_2 = \eta_1$, $\omega(a) = 1$, and $k_2 = k_1$.

In the analyses in the section 4.3.3, we assumed that reaction times in stage 2 are inversely related to the probability of executing an action sequence in stage 1. As such, if subject has taken action A_1 in stage 1, and action A_2 in stage 2, then model prediction of the reaction time of A_2 will be:

$$RT^{-1} = \frac{\pi(A_1A_2|S_0)}{\pi(A_1|S_0) + \pi(A_1A_2|S_0)} \tag{4.14}$$

For the second analysis in the section 4.3.3, we aimed to remove the effect of action sequences in stage 2 choices. We used eight models same as above, but the probability that the model assigns to action a in state $s \in \{S_1, S_2\}$, was defined as:

$$P(a|s) = \pi(a|s) \tag{4.15}$$

Which indicates probability of taking each action in each slot machine is guided only by the rewards earned on that slot machine, and not by the action sequences in stage1.

4.2.3.4 Model selection

Since the two families of models that we are comparing are not nested in each other, we can't use classical model selection. Instead, we use a Bayesian model selection for comparing these two families of models (Penny et al., 2010). We first calculated the model evidence for each model using the Laplace approximation (Daw, 2011; MacKay, 2003), and then calcu-

lated the exceedance probability favoring each family, (taking model identity as a random effect) using the 'spm_compare_families' routine in the spm8 software. Within each family, exceedance probabilities were calculated using the 'spm_BMS' routine (Stephan, Penny, Daunizeau, Moran, & Friston, 2009).

The Laplace approximation requires a prior assumption of probability distributions over the free parameters of models. Similar to the previous study (Daw et al., 2011), for parameters between zero and one (learning rates, reinforcement eligibility, weight parameter), we assumed a Beta(1.1, 1.1) distribution; for exploration-exploitation parameters we assumed a Gamma(1.2, 5) distribution, and for perseveration parameters, a Normal(0, 1) distribution was assumed. The Laplace approximation includes finding the maximum a posteriori (MAP) parameter estimates. For this purpose, we used the IPOPT software package (Wächter & Biegler, 2005) for nonlinear optimization, and the DerApproximator package (Kroshko, n.d.) in order to estimate the Hessian at the MAP point.

4.3 Results

Fifteen subjects completed a two-stage decision-making task (Figure 4.2), in which each trial started with a choice between two key presses (stage 1 actions; A1 vs. A2). Each key press resulted in the appearance of either of two slot machines (denoted by S1 and S2 and distinguished by their colors) in a probabilistic manner. Next, at the slot machines, subjects again chose between two key presses (stage 2 actions; A1 versus A2), and, as a result, received an outcome; i.e., either a monetary reward or a neutral outcome. At stage 1, A1 most commonly led to S1, and A2 to S2 (common transitions; 70% of the time). In a minority of trials, A1 led to S2, and A2 to S1 (rare transitions; 30% of the time). This relationship was kept fixed throughout the test. Each of the stage 2 responses at the slot machines earned a reward either at a high probability (0.7) or a low probability (0.2). In order to ensure the subjects kept searching for the best keys and slot machines during the test, at each trial, with a small probability (1:7), the rewarding probability of each key changed randomly to either the high or low probability.

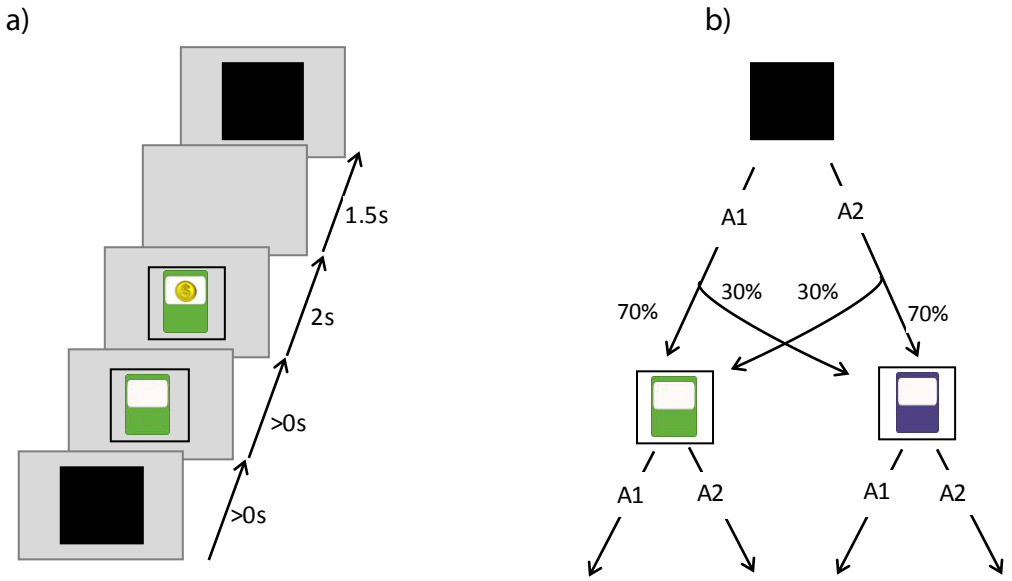


Figure 4.2 – (a) Illustration of the timeline of events within a trial. Initially a black screen is presented, and the subject can choose between pressing A1, or A2 (stage 1 choice). After a key is pressed, one of the slot machines is presented, and the subject can again choose between pressing A1, and A2 (stage 2 choice). Choices at stage 2 are reinforced by monetary reward. (b) Structure of the task. One of the key presses commonly leads to one of slot machines (70% of the time), and the other key commonly leads to the other slot machine. Choices at stage 2 are reinforced either by a high probability (0.7) or a low probability (0.2). With a small probability (1/7), the rewarding probability of each key changes randomly to either the high or low probability.

.

Each participant completed 270 trials.

4.3.1 Goal-directed and habitual performance

In the analysis, we first sought to establish whether decision-making in this task is goal-directed, habitual or a mixture of both and, if both, to assess whether goal-directed and habitual control interact according to a flat structure or a hierarchical structure.

The first question can be answered by looking at the likelihood of the subjects repeating the same stage 1 action on each trial based on feedback received on the previous trial (Daw et al., 2011). Take for example a trial in which a subject presses A1 and transfers to the S2 slot machine (which is rare result of choosing A1). If the participant presses a button of that slot

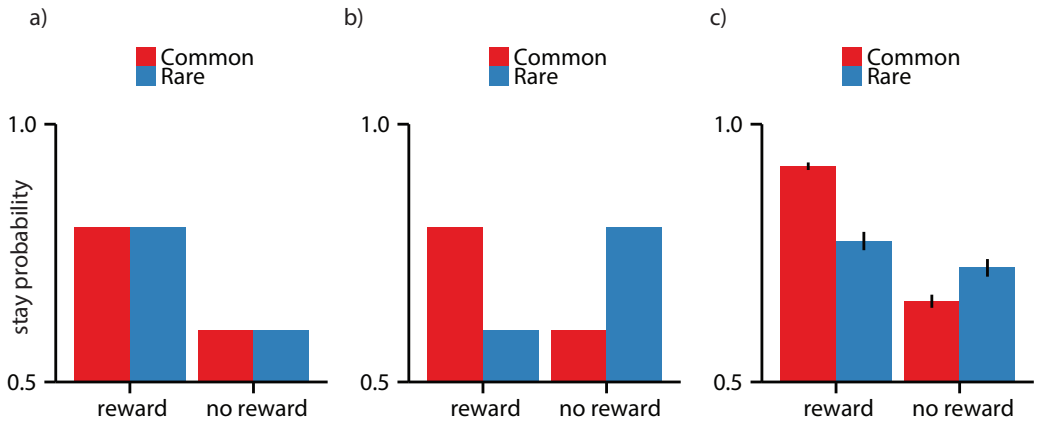


Figure 4.3 – (a) Modeled habitual action control on the two–stage task: Under habitual control a stage 1 action that has been eventually reinforced (reward) in the previous trial is more likely to be repeated (higher stay probability), regardless of whether the repeated action commonly leads to the same slot machine (common) or not (rare). (b) Modeled goal-directed action control on the two–stage task: Under goal-directed action control a reinforced action is repeated if it commonly leads to the same slot machine in which reward is received, otherwise the other action is selected. (c) Data from the experiment: Actual stay probabilities averaged over all subjects and trials. When the previous trial was rewarded, stay probability was generally higher (as in habitual control), and was also higher when the previous trial was a common transition (as in goal-directed control). Thus, the responses of the subjects in the experiment were found to be a mixture of both habitual and goal-directed action control.

.

machine and receives a reward, this implies S2 is probably a good slot machine and, if the decision-making is goal-directed, in stage 1 of the next trial the subject should try to reach this S2 slot machine again. It is expected therefore, that the probability that the subject will press A2 will increase because it is this key that (in this example) commonly leads to S2 (cf. Figure 4.3b). In contrast, if decisions are habitual, subjects should not be guided by contingencies between the responses and slot machines, and should tend to stay on the previously rewarded action, A1 (Figure 4.3a).

The results are presented in Figure 4.3c, which shows the probability of repeating the same action computed across all subjects and trials. We analyzed the data using mixed-effects logistic regression analyses by taking all coefficients as random effects across subjects (see section 4.2.2). Results show that being rewarded in the previous trial increased the chance of staying on the same action, irrespective of whether it was a rare or a common transition (main

effect of reward; coefficient estimate = 0.61; SE = 0.09; p<3e-11), which suggests that habits constitute a component of the behavior. On the other hand, this increase was higher if the previous trial was a common transition (and lower after an unrewarded trial), suggesting that subjects also utilized their knowledge about the task structure (reward-transition interaction; coefficient estimate = 0.41; SE = 0.11; p<5e-4). Therefore, the subjects' behavior was a mixture of both goal-directed and habitual actions. Also, as the figure shows, the probability of staying on the same action is generally higher than not staying on it, irrespective of reward and transition type in the previous trial (the intercept term is significantly positive; estimate = 1.52; SE = 0.20; p<10e-14), which reflects a general tendency of animals and humans to repeat previous actions (Ito & Doya, 2009; Lau & Glimcher, 2005).

In previous studies, a hybrid model of model-free and model-based reinforcement learning (RL) was advanced to explain the behavior of subjects on this task based on the flat structure (Daw et al., 2011; Gläscher et al., 2010; Otto, Gershman, et al., 2013) (section 2.4.3). According to this model, action values learned in model-free RL, roughly, reflect the frequency of the action rewarded on previous trials irrespective of the action-outcome contingency (i.e., in the current task, which key generates which slot machine) and, as such, these values underlie the habitual component of the model. These model-free values are then mixed with the values provided by the goal-directed system (modeled by a model-based RL) to produce the final values which guide action selection. As a consequence, and consistent with the above results, we should expect to see a combination of both habitual and goal-directed actions. The prediction from this hybrid model is illustrated in Figure 4.4a. A hierarchical structure can, however, also be used to explain these results. For example, assume that a subject presses A1 in stage 1, and A2 in stage 2 and receives a reward. As a result, the goal-directed system learns that contingency between the A1A2 action sequence and the reward is increased and so it should be more likely to repeat the action sequence in the next trial, whether or not the reward was received from the S1 or S2 slot machine (i.e., the common or rare transition). As the evaluation and performance of an action sequence is not guided by the task structure (i.e. the key-slot machine association), from this perspective it constitutes the habitual component

Figure 4.4 – The probability of staying on stage 1 action in simulations of: (a) the flat architecture; and, (b) the hierarchical architecture. Both architectures can model the pattern of data observed in stage 1 stay probabilities on the task: i.e., a higher stay probability after being rewarded on the previous trial and an interaction between reward and transition.

of the behavior (see (Dezfouli, Lingawi, & Balleine, 2014) for the relation of phenomenon to outcome devaluation experiments). All actions - either single action (e.g., A1) or action sequences (e.g., A1A2)-, will be subject to the goal-directed action selection process, such that actions with higher values will be selected with a higher probability. As a consequence, this implies that the behavior will be a mixture of habitual (when action sequences are selected) and goal-directed (when single actions are selected) actions and that this mix of actions can be generated without the need for the model-free component or an explicit arbitration mechanism used in the flat structure. This prediction is illustrated in Figure 4.4b.

4.3.2 The interaction of goal-directed and habitual actions

Although both approaches are able to explain the mixture of behavioral control in the stage 1, they make different predictions about stage 2 choices. This is because, if the observed habitual behavior is due to the execution of an action sequence, rather than cached values as the model-free account supposes, then we expect the subject to repeat the whole action sequence in the next trial, not just stage 1 action.

Staying on the same stage 1 action in the next trial after being rewarded implies that this is probably a habitual response and so we expect the subject to repeat stage 2 action as well,

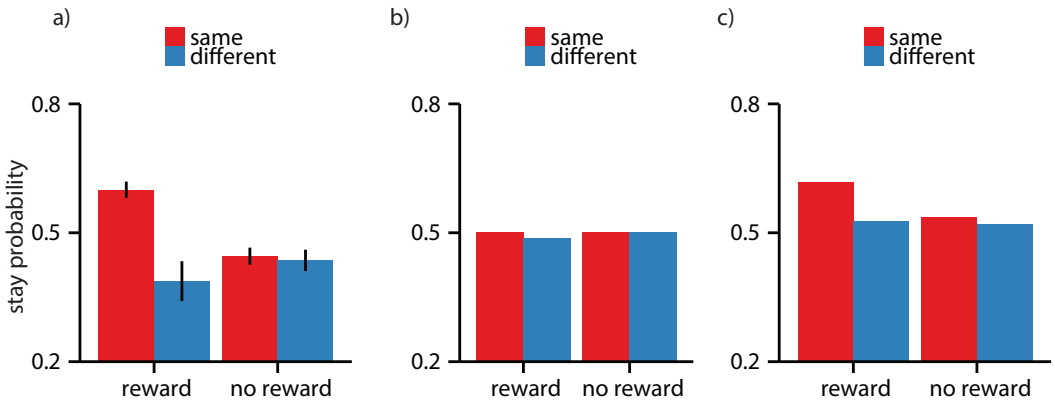


Figure 4.5 – The probability of staying on stage 2 action on trials for which the slot machine differs from the on one in the previous trial: (a) The observed stay probabilities. When the subjects are rewarded and stay on the same stage 1 action (same), the probability of staying on the same stage 2 action is higher. (a) Simulation of the flat architecture. Note that this is not consistent with the pattern in panel (a). (c) Simulation of the hierarchical architecture, which is consistent with the pattern observed in actual stay probabilities.

even if the slot machine is different from the one in the previous trial. In contrast, if the subject switches to the other stage 1 action, the previous action sequence is not repeated, and thus stage 2 action is not expected to be repeated if the subject ends with a different slot machine in the next trial. In order to test this prediction, we looked at the trials that had a different slot machine to the one in their previous trial.

Figure 4.5 shows the probability of repeating the same stage 2 action as a function of whether this action was rewarded on the previous trial and the subject had subsequently taken the same stage 1 action. Logistic regression conducted on stage 2 choices using factors of reward, separating rewarded and non-rewarded trials, and action, separating trials on which stage 1 action was the same from those on which it differed, found neither an effect of reward ($p>0.05$), nor of action ($p>0.05$) but found a significant interaction between these factors (coefficient estimate = 1.02; SE = 0.38; $p<0.008$), indicating that, during the execution of habitual responses, subjects tended to repeat stage 2 action. This interaction remained significant even when we restricted the analysis either to trials after rare transitions (coefficient estimate = 1.33; SE = 0.60; $p<0.05$) or after common transitions (coefficient estimate = 0.93; SE = 0.38; $p<0.05$).

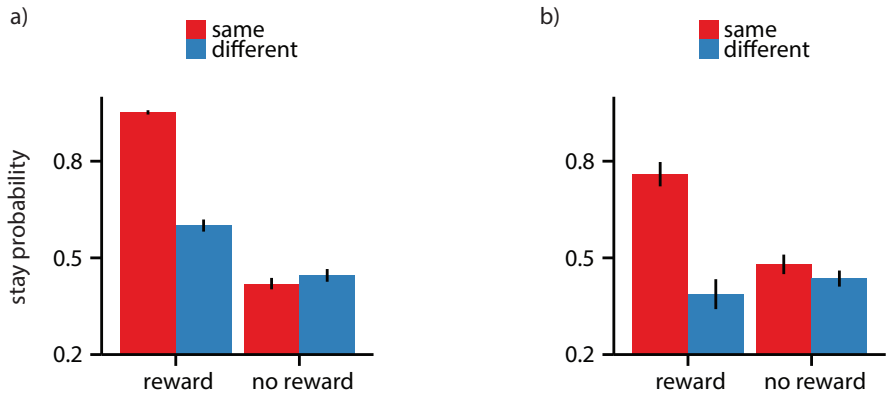


Figure 4.6 – The probability of staying on the stage 2 action when the same (a) or different (b) stage 1 action is taken, as a function of whether the previous trial is rewarded, and whether the stage 2 state is the same or different from the previous trial

Importantly, the fact that the effect of the reward was not significant rules out the possibility that the effect was due to the generalization of the values across slot machines.

Simulations of the flat and hierarchical models are presented in Figure 4.5b and c, respectively. As predicted, the hierarchical structure captures the pattern of the subjects' stage 2 actions (the interaction between the reward and the same stage 1 action; $p<0.001$), whereas the flat structure is not consistent with repeating the same action in stage 2 ($p>0.05$).

Previously, we focused on trials with a different slot machine to the one in the previous trial. This was because, in this condition, flat and hierarchical accounts provide different predictions. When the slot machine is the same, both accounts (flat and hierarchical) predict that being rewarded in the previous trial increases the probability of staying on the same stage 2 action. In addition to this prediction, the hierarchical account predicts that when the slot machine is the same as the one on the previous trial, this increase should be higher than the increase when the slot machine is different. This is because, when the slot machine is different, staying on the same stage 2 action is driven by execution of the previous action sequence whereas, when the slot machine is the same, executing either the previous action sequence or a goal-directed decision at stage 2 can result in staying on the same stage 2 action.

As a consequence we looked at the effect of being rewarded in the previous trial, and whether

the slot machine was the same as the one in the previous trial, on the probability of staying of the same stage 2 action (in the trials in which stage 1 action was the same as the previous trial).

Figure 4.6a shows the results. A significant main effect of reward was found (coefficient estimate = 0.69; SE = 0.21; $p<0.002$) indicating that being rewarded in the previous trial increases the probability of taking the same stage 2 action, irrespective of whether the slot machine was the same as the previous trial or not, which is consistent with the hierarchical account. In addition, we found a significant interaction between the effect of reward and whether the slot machine being the same (coefficient estimate = 3.46; SE = 0.51; $p<3e-11$), consistent with the finding that the probability of staying on stage 2 action was higher when stage 1 action was the same.

Figure 4.6b shows the probability of staying on the same stage 2 action when the subject takes a different stage 1 action. As predicted, because the subject did not execute the previous action sequence, the main effect of reward was not significant ($p>0.05$) but the interaction between reward and stage 2 state being the same was significant (coefficient estimate = 1.72; SE = 0.40; $p<3e-5$) which means that subjects tend to take the same action on the same slot machine after being rewarded, as predicted by both accounts.

4.3.3 Reaction times during habit execution

In the previous section we showed that if, after being rewarded, the subject repeats the same stage 1 action, they are probably repeating the previous action sequence and, as such, they tend to repeat stage 2 action as well. However, even in the situation in which the subject is executing an action sequence there will be trials on which they might not repeat the same stage 2 action. In such conditions, we should suppose that either (i) the subject took an exploratory goal-directed action in stage 1, or (ii) the subject started an action sequence but its performance was inhibited and control returned to the evaluation system in stage 2. In both cases, the hierarchical account predicts that reaction times on trials in which the same

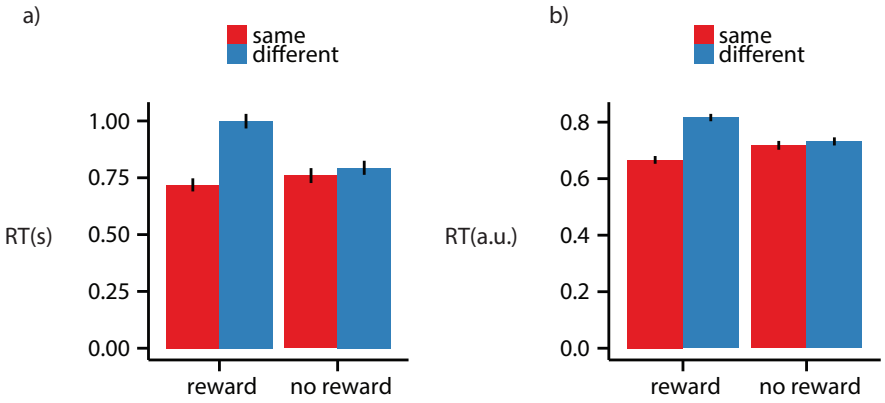


Figure 4.7 – (a) Reaction time (RT) in stage 2 action (when the same stage 1 action was taken) as a function of whether the same stage 2 action was taken and whether previous trial is rewarded (only calculated for trials on which stage 2 state was different from the previous trial). (b) Predicted reaction times by the model (a.u. : arbitrary unit).

stage 2 action is not taken should be higher.

Figure 4.7a illustrates these reaction times as a function of whether the previous trial was rewarded and the subject takes the same stage 2 action (only in trials on which the slot machine is different from that on the previous trial and the subject subsequently takes the same stage 1 action). If the previous trial is rewarded, reaction times were lower when a subject completes an action sequence than when stage 2 action was not executed as a part of a sequence (coefficient estimate = -1.66; SE = 0.45; $p<3e\text{-}4$). Importantly, the effect was not significant when the previous trial was not rewarded ($p>0.05$), which rules out the possibility that the observed increase in the reaction times was because of the cost of switching to the other stage 2 action.

We further asked whether the model can predict the reaction times in stage 2. As mentioned above, at stage 1, the goal-directed process more frequently selects actions that have a higher contingency to reward (either single actions, or action sequences). As such, if an action sequence has a high value, it is likely to be selected for execution, and so we expect a low reaction time in stage 2. For example, assume the subject has executed action A1 in stage 1, and A2 in stage 2 and the aim is to predict whether A2 has a high or low reaction time. It can be argued, if the value of the A1A2 action sequence is high, then it was probably executed in

stage 1, and thus the execution of A2 is part of an action sequence (A1A2) started in stage 1, implying the subject should show a low reaction time. In general we assume that the reaction time in stage 2 is inversely related to the value of the action sequence that contains that action (see section 4.2.3.3). In the case of this example we will have:

$$RT^{-1} \propto \texttt{ probability of executing action sequence A1A2}$$

Based on this, we calculated the predicted reaction time of the action taken by the subject in the conditions shown in Figure 4.7a. The results are shown in Figure 4.7b. As the figure shows, the predicted reaction times by the model are consistent with the pattern of reaction times observed in the data.

In general, the above analysis of stage 2 performance and this analysis of reaction times implies that (i) when the previous trial is rewarded, (ii) the same stage 1 action is taken, and (iii) the reaction time is low, then the subject is most likely performing an action sequence. As a consequence it is expected to repeat the same stage 2 action, even on a different slot machine to the one in the previous trial. In order to more closely examine this relationship we used conditional inference trees and partitioned stage 2 actions into whether they involved staying or switching to the other action based on the above three factors (see section 4.2.2). The results are shown in Figure 4.8. As the figure shows, when the previous trial was not rewarded (node # 1 'no reward' condition), staying on the same stage 2 action was independent of either whether stage 1 action was repeated or the reaction time was low ($p>0.05$; permutation test). If the previous trial was rewarded (node # 1 'reward' condition) then, if the reaction time was high (node #2 RT>0.437s) or the reaction time was low but the subject doesn't repeat stage 1 action (node #3 'different' condition), then again stage 2 action was not repeated. Only when: (i) the previous trial was rewarded, (ii) the subject took the same stage 1 action, and (iii) their reaction time was low (node #3 'same' condition), did the subject repeat stage 2 action, consistent with the prediction of the hierarchical account.

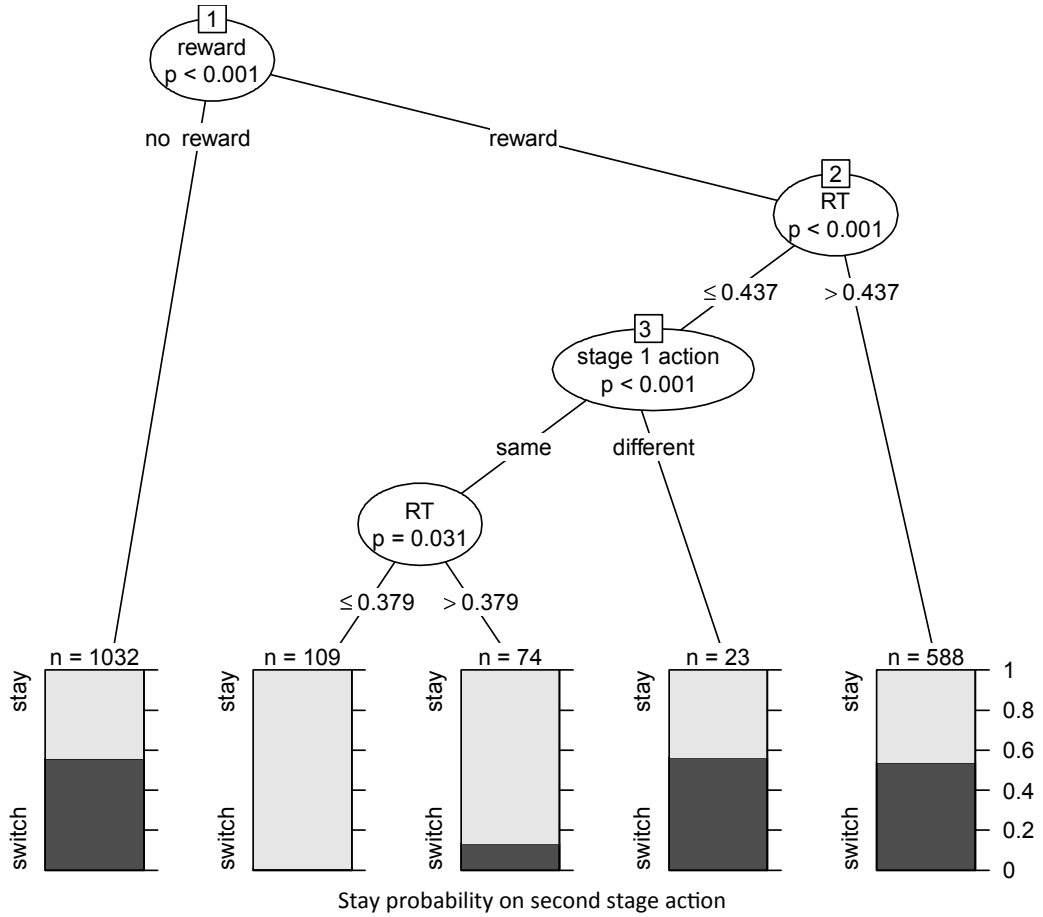


Figure 4.8 – Partitioning the probability of staying on the same stage 2 action (stay: staying on the same stage 2 action; switch: switching to the other stage 2 action) as a function of (i) reward on the previous trial (node #1), (ii) whether the same stage 1 action is taken (stage 1 action; node #3), and (iii) reaction times (RT). 'n' represents the number of data points; p-values are calculated using a permutation test.

4.4 Behavioral modeling: Bayesian model selection

The results described in the previous sections suggest that a hierarchical structure better characterizes the effect of feedback from the previous trial on performance on the subsequent trial. However, choices are generally guided by the feedback from all previous trials, not just the immediately prior trial. As such, it is still to be established which framework better captures behavior in this more general condition.

We used a Bayesian model selection method to establish which framework produces choices that are the most similar to the subjects' actions. Both flat and hierarchical architectures have different variants with different degrees of freedom. As such, we compared a family of flat models with a family of hierarchical models (Penny et al., 2010), where each family consists of a complex model, and its nested simpler models. The results (Table 4.1) show that, given the subjects' data, the hierarchical family is more likely than the flat family to produces choices similar to those made by the subjects. We found that the exceedance probability in favor of the hierarchical family was 0.99 meaning, roughly, that we can be 99% confident that the hierarchical family generated the observed data.

In the hierarchical family, the probabilities of taking actions in stage 2 are partially based on the probability of taking an action sequence in stage 1. As these stage 2 choices are the canonical difference between the two families, we expected that removing the effect of action sequences on stage 2 choices would reduce the fit of the hierarchical account to data. Thus we generated a family of hierarchical models similar to Table 4.1. but with the effect of action sequences on stage 2 actions removed, and compared the generated family with the family of hierarchical models presented in Table 4.1 (see section 4.2.3.3). Results indicated that the exceedance probability in favor of the family in which the performance of action sequences was reflected in stage 2 choices was 0.99, confirming that the selection of an action sequence in stage 1 increased the probability of taking the second element of the action sequence in the next stage.

Table 4.1 represents the model comparison results within each family. The parameter estimates

Table 4.1 – Model comparison between hierarchical and flat families. H: Hierarchical; F: Flat. Shown for each model: negative log model evidence $-\log(P(D|M))$; a pseudo-r statistic (p–r^2) which is a normalized measure of the degree of variance accounted for in comparison to a model with random choices; the number of subjects favoring the best fitting model based on the model evidence; The exceedance probability which represents the probability that each model (or family) is most likely among alternatives over the population. *best fitting model in each family. ** best fitting model.

Family	Free parameters	-log $p(D\|M)$	psudo-r^2	In each family		In total		
				number favoring best model	Exceedance probability	Exceedance probability	number favoring best model	Exceedance probability
H	β,η_1,k_1	4219.6	0.26	12	0.000	0.004	12	0.993
	β,η_1,k_1,w	4092.7	0.29	13	0.000	0.003	13	
	β,η_1,k_1,k_2	4189.3	0.27	12	0.001	0.005	12	
	β,η_1,k_1,η_2	4127.9	0.29	11	0.003	0.013	11	
	β,η_1,k_1,w,k_2	4074.9	0.30	11	0.002	0.011	11	
	$\beta,\eta_1,k_1,w,\eta_2$	4078.0	0.30	10	0.004	0.011	10	
	$\beta,\eta_1,k_1,k_2,\eta_2$	4110.3	0.29	11	0.000	0.004	11	
	$\beta,\eta_1,w,k_1,k_2,\eta_2$	4058.7	0.3	-	0.986*	0.911**	-	
F	β_1,η,k,w	4212.0	0.27	14	0.004	0.002	14	0.006
	β_1,η,k,w,β_2	4212.5	0.27	12	0.004	0.004	13	
	β_1,η,k,w,λ	4168.6	0.28	-	0.697*	0.006	12	
	$\beta_1,\eta,k,w,\alpha_1$	4201.1	0.27	12	0.005	0.002	14	
	$\beta_1,\eta,k,w,\beta_2,\lambda$	4173.1	0.28	9	0.032	0.006	11	
	$\beta_1,\eta,k,w,\beta_2,\alpha_1$	4198.4	0.27	11	0.007	0.003	13	
	$\beta_1,\eta,k,w,\lambda,\alpha_1$	4174.2	0.28	11	0.049	0.002	12	
	$\beta_1,\eta,k,w,\beta_2,\lambda,\alpha_1$	4169.4	0.28	10	0.199	0.005	12	

for the best fitting model from each family in terms of the exceedance probabilities (Stephan et al., 2009) are presented in Table 4.2. The best fitting models from each family were simulated in the task conditions to produce Figure 4.4 and Figure 4.5.

Table 4.2 – Best fitting parameter estimates for each family across subjects.

Hierarchical				Flat			
Parameter	First Quartile	Median	Third Quartile	Parameter	First Quartile	Median	Third Quartile
β	4.24	5.80	6.96	β_1	1.64	2.33	3.44
η_1	0.82	0.89	0.95	η	0.65	0.84	0.93
k_1	1.25	1.66	2.25	k	0.84	0.94	1.40
w	0.25	0.46	0.62	w	0.40	0.59	0.68
k_2	-0.50	0.30	0.70	λ	0.69	0.93	0.95
η_2	0.14	0.29	0.77	-	-	-	-

4.5 Discussion

Although prior research has suggested that goal-directed and habitual actions should be conceived as single step actions organized according to a flat architecture (e.g. (Daw et al., 2005, 2011)), the results of the current experiment found that: (i) human subjects combined actions together to form action sequences, as revealed by the open-loop execution of sequences of actions and reaction times in the current task, and, therefore, that action sequences constituted a necessary component of behavior; (ii) the use of action sequences by human subjects was sufficient to explain habitual decisions on this task, meaning choices that were not guided by action-outcome contingencies; and, (iii) a goal-directed system assessing both actions and action sequences in a hierarchical manner explained behavior better than a flat model attributing habits to model-free evaluation.

Furthermore, although hierarchical models have had a longstanding role in decision-making (Lashley, 1951; Miller et al., 1960; Newell & Simon, 1963; Botvinick & Plaut, 2004; Estes, 1972; Schneider & Logan, 2006; Cooper & Shallice, 2006, 2000), here we provide direct experimental evidence for the role of these models in understanding the operation and interaction of goal-directed and habitual actions. We used a version of the two-stage discrimination task described by Daw et al (Daw et al., 2011) in which the ambiguity of stage 1 predictions by both actions and stimuli was reduced by removing the explicit predictive cues of previous versions. Using this task we found, as previously described, that action selection in stage 1 reflected a mixture of goal-directed and habitual strategies. The two accounts diverge with respect to the status of stage 2 actions; whereas the flat architecture/single step action perspective predicts that the status of action selection in stage 2 should be independent of the first, we found that this was not true; habitual action selection in stage 1 predicted continued habitual selection in stage 2 as a sequence of actions, a finding predicted by a hierarchical goal-directed/habit sequence account (Dezfouli & Balleine, 2012). According to this account, at the top of the hierarchy the goal-directed system evaluates and selects goals and then habits efficiently implement decisions made by the goal-directed system in the form of action sequences. In

comparison to the other accounts, which posit a flat interaction between these two systems, we found that the hierarchical account provides more accurate predictions both in terms of the choices of the subjects, and in terms of their reaction times during action selection. When performing according to a habitual sequence of actions, subjects tended to repeat both previously reinforced sequences and to perform these sequences at significantly lower reaction times than when their actions were goal-directed.

4.5.1 Hierarchical decision-making and the two-stage task

A number of studies have previously investigated the relationship between hierarchical RL and decision-making (Doya, 1999; Botvinick et al., 2009; Ribas-Fernandes et al., 2011; Diuk, Tsai, Wallis, Botvinick, & Niv, 2013; Frank & Badre, 2012; Badre & Frank, 2012; Reynolds & O'Reilly, 2009; Holroyd & Yeung, 2012). We extended these studies by showing how the formation of action sequences can lead to decisions that are insensitive to (i) the values of the outcomes (Dezfouli & Balleine, 2012) and (ii) the contingency between specific actions and their outcomes (i.e. the key press–slot machine associations in this study), the two defining characteristics of the habitual behavior.

The other difference between the hierarchical RL model that we used here and previous work is that we assumed that performance of action sequences is insensitive to the feedback received during execution (Pew, 1966; Keele, 1968), whereas, in general, previous work based on hierarchical RL theory has assumed that action selection is based on the state of the environment (Barto & Mahadevan, 2003; Dietterich, 2000; Sutton et al., 1999). Within this latter framework, one can posit that habits are hierarchically organized actions but that their performance is sensitive to the feedback received after execution of each individual action. Although this class of models can explain habitual behavior executed in stage 1 of the current task, this approach predicts that stage 2 actions will, ultimately, be similar to those of the flat architecture discussed earlier, which is not consistent with the data observed in this study.

In the hierarchical account advanced here we assumed, based on the previous findings in

rodents (Balleine, Garner, Gonzalez, & Dickinson, 1995; Ostlund et al., 2009), that, similar to single actions, action sequences are also under goal-directed control. Alternatively, it is possible that the value of any action sequence is learned in a model-free manner (for example using *Q*-learning) without learning the identity of the particular outcome that it predicts. Our results are silent with respect to this latter assumption; nevertheless, whatever the case, the conclusion that habitual responses in stage 1 were due to the execution of an action sequence still holds. One way to study this issue is to add another choice to the end of the task, making it a three stage task, and then asking whether performance of for example A1A2 action sequence is goal-directed or habitual, which can be answered by devaluation of outcome of A1A2, or using the same task structure that we used here to distinguish habitual and goal-directed actions. However, again, if it were found that the selection of the A1A2 sequence was not sensitive to environmental contingencies, or outcome values, this could be due either to the formation of A1A2A3 action sequence (since outcome of A1A2 falls within sequence boundaries (Dezfouli & Balleine, 2012)), or it could be because action sequences are open to model-free evaluation. Similar to the study here, these accounts can be distinguished by examining whether the subject selects A3 during habitual selection of A1A2 irrespective of the outcome of A1A2 performance. If so, it can be concluded that the observed habitual behavior is due to the formation of an action sequence, not model-free RL. Along the same lines, it is possible to assume that, in the current study, stage 2 habitual responses were guided by a flat model operating in parallel to the hierarchical model we propose here. Again, although the task results are neutral with respect to this assumption, adding a parallel model increases the model's complexity, is not required to account for the current data, and so it necessity should be motivated by additional behavioral data.

It might also be argued that, although the current predictions apply to the modified two-stage discrimination task used here, they may not apply to previous versions of the task. In previous versions, subjects at each stage chose between two symbols instead of two fixed actions and the symbols moved from side to side at each trial ensuring there was no consistent mapping between the button presses and the symbols. There are two points to make here: First, the

fact that specific (e.g. left- or right-hand) actions are degraded in their contingency with the outcome on this version of the task raises the issue of stimulus control; either the stimuli exclusively mediate the predictions of stage 2 outcomes or the concept of action needs to be made more liberal to the selection of a symbol. The former approach would, of course, render the task Pavlovian, rather than instrumental, and the applicability of model-based control problematic. Second, and relatedly, in order to apply our hierarchical model to the earlier task, we also need to extend the concept of an 'action' from pressing a button (as in our task), to selecting a symbol; if this is accepted then, using the logic laid out earlier, the hierarchical goal-directed/habit sequences model can explain the results of the task. In the prior version of the task, symbols in stage 2 were different from each other, for example in one of stage 2 states subject could choose between symbols 'C' and 'D', but in the other stage 2 state, the choice was between symbols 'E' and 'F'. As such, we cannot directly assess the probability of staying on the same stage 2 action if the subjects end up in a different stage 2 state. Nevertheless, the hierarchical theory predicts that if the subject selects same stage 1 action, and ends up with the same stage 2 state and selects the same stage 2 action, then the reaction time will be faster than when they end up with in a different stage 2 state.

Here we assumed that action sequences were available to the subjects from the beginning of the training, and we ignored that such action sequences needed to be acquired through a learning process, as described in the previous chapter. This was mainly because we assumed that the human subjects that participated in the study were generally familiar with taking a sequence of button pressing on a computer keyboard, and so they could develop such action sequences very quickly. Therefore, we didn't include the process of action sequence learning here.

4.5.2 Deviations from prediction and the interpretation of the two-stage task

Predictions of both models (flat and hierarchical) were found to deviate from the behavior of the subjects in two cases. In the first case, if, after being rewarded, the subject switches to the other action then both accounts predict that the probability of staying on stage 2 action

should be on average 0.5 (Figure 5b,c). However, in the actual data it is below 0.5 (Figure 5a). In the second case, both accounts predict that the difference between stay probability in common and rare transitions should be equal in both the reward and no-reward conditions (Figure 4a,b), however, as Figure 3c shows, the difference is larger in the reward condition. It is possible to capture these two deviations by adding more free parameters to the models; however, since the deviations exist for both the flat and hierarchical families and so do not affect the comparison between them, we didn't add further parameters to account for these two deviations.

As in previous work, we interpreted the interaction between being rewarded and the type of transition in the previous trial (rare or common) as the evidence for goal-directed behavior. It should, however, be noted that, if there is a strong initial bias in total possible reward for one action vs. the other at the first-stage, and reward transitions are slow, then it is possible to observe an interaction between reward and transition type without engaging a goal-directed system. As a consequence of the higher overall probability of reward for taking, say, action 'A1' in stage 1, the subject can establish that action has a higher value (without relying on the task structure) and so will take that action, i.e. 'A1', more frequently than the other, i.e. action 'A2', which means that the probability of staying on action 'A1' will be higher than action 'A2' in general. At the same time, because action 'A1' is better than the other action, most of the rewarded common transitions and unrewarded rare transitions result from taking action 'A1'. Likewise, most of the unrewarded common transitions and rewarded rare transitions will be the result of taking action 'A2'. This fact, and the fact that stay probability on action 'A1' is generally higher, will produce a reward-transition interaction, without having a goal-directed system, at least in the period that action 'A1' is better than the other action. This bias is proportional to how fast the bias in stage 1 values changes and cannot account for the current data. It should also be noted that, as the comparison between the flat and hierarchical model families was based on model fit, those results don't suffer from this problem.

4.5.3 Inhibitory interactions between goal-directed and habitual control

Although, on the hierarchical goal-directed/habit sequence model advanced here, habits are integrated with the goal-directed process to reach the goals selected by this latter system, competition can also occur between these two systems when the further execution of an ongoing habit sequence is found to be inappropriate by the goal-directed system and it attempts to take back control. This type of competition resembles the situation in an inhibitory control task, such as the stop-signal task, in which subjects must respond quickly when a 'go' signal appears but must stop the action if a stop-signal appears (Verbruggen & Logan, 2008). In the context of our task, seeing a slot machine at stage 2 is the 'go' signal, which causes the execution of the next action in the sequence. The stop signal comes from the goal-directed system when the pending response is identified as inappropriate. Consistent with this conception in conditions in which sequence performance is inhibited, reaction times are slower. In the stop-signal task, subjects are typically able to inhibit their responses when the stop signal is temporally close to the 'go' signal. Although the stop signal task is more global in terms of response inhibition, whereas in the current task the inhibition is specific to one as opposed to an alternative action, this implies that the ability of the goal-directed system to override habits depends on how fast it calculates the correct action: the faster it calculates, the higher the chance of taking control back before action execution.

4.5.4 Habit sequences vs. stimulus-response habits

It is also interesting to consider the relationship between habit sequences and stimulus-response (S-R) theories of habit learning. The S-R theory of habit learning maintains that habits are responses that are elicited by antecedent stimuli rather than their consequences (Guthrie, 1935; Hull, 1943). Such S-R theories maintain that stimuli trigger their associated behavioral responses due to an association between the stimulus and the response. According to the habit sequence theory, however, the stimulus instead signals that the next action in the sequence should be executed; i.e., in the context of our task, seeing a slot machine signals that it is time for the next action to be executed. Although the next action to be executed

is determined by the sequence and the stimulus does not play a role in the *selection* of the next response, the response is still stimulus-bound to some extent (e.g., for timing and motor coordination) and is elicited only when the next expected stimulus is encountered. Nevertheless, these two theories provide different predictions. For example, S-R theory predicts that, in the presence of the appropriate stimulus the response will be performed, irrespective of whether that stimulus was encountered as part of the habit sequence or not. In contrast, habit sequence theory predicts that the individual will respond to the stimulus only when the appropriate habit sequence has already been launched by the goal-directed system.

5 Hierarchical decision-making in rats

As mentioned in chapter 2, evidence indicates that goal-directed actions, i.e., actions that are taken by an agent to attain a certain goal, are crucial components of decision-making processes in animals and humans. The source of these actions is suggested to be a model-based reinforcement learning system in the brain, which learns the contingencies between actions and the states of the environment. Such a decision-making process, however, requires an agent to firstly learn the correct representation of the task space, and secondly, hierarchically organize actions in order to make goal-directed decision-making scalable to multi-stage and complex environments. Here, using a sequential decision-making task in rats, we show that the profile of choices made by animals reveals a gradual shift from a simple representation of the task space, to a more complex form, which matches the true sequential structure of the task. Subsequently, we show that within this multi-stage representation, animals engage in a hierarchical model-based decision-making process. Furthermore, results of a Bayesian model comparison procedure, consistent with the behavioral results, confirm that animals are using hierarchical model-based reinforcement learning. Therefore, we provide evidence supporting the hypothesis that the decision-making processes in the brain can be conceptualized using model-based hierarchical reinforcement learning and we also provide a new experimental protocol in rats, that can be used to measure the operation of these decision-making processes.

5.1 Introduction

According to experimental evidence, there is a combination of multiple reinforcement-learning (RL) systems behind the value-based decisions made by humans and other animals (Balleine & O'Doherty, 2010; Daw et al., 2005; Dolan & Dayan, 2013; Doya, 1999; Keramati et al., 2011). These systems have partly overlapping neural and computational substrates and choices made by each system exhibit dissociable behavioral characteristics. Recently, it has been suggested that these multiple systems are organized in a hierarchical structure in which higher-level systems make high-level plans that are then delegated to the lower-level processes for the purpose of implementation (Botvinick et al., 2009; Dezfouli & Balleine, 2013; Frank & Badre, 2012; Holroyd & Yeung, 2012; Ito & Doya, 2011). As for example, one might decide between making tea or coffee (high-level planning) and, if tea is chosen, the lower-lever processes repeat the sequence of actions that are used for making tea.

The above description of the hierarchical organization of planning systems broadly maps onto hierarchical model-based RL (HMB RL) (Barto & Mahadevan, 2003; Botvinick & Weinstein, 2014)(see section 2.4.5). The model-based component of this framework is a computational description of what is known in the animal learning literature as goal-directed instrumental conditioning (C. D. Adams, 1982; Dickinson & Balleine, 1994; Tolman, 1948) (section 2.3). For action selection in a goal-directed manner, an agent deliberates over the consequences of available actions and selects the actions that are predicted to lead to the desirable outcomes (or goals). Such an action selection process, however, is known to suffer from the 'curse of dimensionality', i.e., model-based RL does not scale to complex environments in which the agent needs to iterate over a large number of future states and actions for decision-making. HMB RL (Barto & Mahadevan, 2003) offers a solution to this problem by introducing temporally extended actions. This notion is related to psychological studies of motor control and skill learning according to which humans and other animals form action chunks or action sequences by concatenating actions together (Lashley, 1951; Rosenbaum, 2009; Smith & Graybiel, 2014) (section 2.4.4). After formation, these action sequences will serve the agent to

reach its goals without the need to make decisions for each component of the action sequence separately. In this way, a HMB RL evaluates all the actions, including actions sequences, and selects one of them for execution (Dezfouli & Balleine, 2012).

Conceptually, predictions from HMB RL regarding how decisions are made and executed can be divided into four dimensions: Firstly, HMB RL predicts that actions concatenate and make action sequences that are performed as a single response unit. This prediction is supported by a body of evidence that shows that humans (Rosenbaum, Cohen, Jax, Weiss, & van der Wel, 2007; Wymbs et al., 2012), non-human primates (Tanji, 2001), and rodents (Jin & Costa, 2010) are able to form and execute sequences of actions in this way. The markers of such response units are usually taken to be (i) faster reaction times between the elements of the action sequences and (ii) feedback-free operation of the sequences, meaning that, once started, the sequence will run to its end irrespective of the feedback received after the performance of each individual action (Keele, 1968; Pew, 1966) (section 3.1).

Secondly, HMB RL implies that an action sequence is evaluated as a whole, not based on its individual components. Providing evidence for this prediction requires a choice task in which the disjoint evaluation of actions predicts different choices rather than evaluating a sequence of actions as a whole. Here, direct studies are scarce (Dezfouli & Balleine, 2013; Ostlund et al., 2009), although there is evidence from spatial navigation literature (Solway et al., 2014; Wiener & Mallot, 2003).

Thirdly, according to HMB RL single actions are evaluated in a goal-directed manner. This prediction is supported by outcome revaluation experiments in humans and rodents (Balleine & O'Doherty, 2010), which show that the choice of actions is sensitive to changes in the value of the outcome earned by a specific action.

Fourthly, HMB RL requires that, as with single actions, action sequences are also evaluated in a goal-directed manner, meaning that, for example, a person used to starting a sequence of actions in order to attain a certain goal will no longer start that sequence if the goal is no longer desirable. This property extends the second point above such that an action sequence is not

only evaluated as a whole, but also with respect to its ultimate outcome, which is supported by outcome devaluation experiments in rats (Ostlund et al., 2009).

Finally, there is a hidden assumption for using HMB RL as a model of goal-directed actions (or any other model-based RL system): an agent should be able to build a sequential representation of the task space, i.e., the several intermediary needed to reach the goal such that outcomes are not necessarily the immediate consequence of actions. Such ability has been demonstrated in humans (e.g., (Daw et al., 2011)), rodents (Ostlund et al., 2009), and monkeys (e.g., (Mushiake et al., 2006)).

As such, there is independent evidence regarding different aspects of a HMB RL in prior studies, however, these aspects need to be demonstrated in a single decision-making task to confirm the integrated operation of HMB RL as the backbone of the decision-making processes in humans and animals. Here, we target the first three properties of HMB RL, the aim of the current series of experiments is, therefore, to provide concurrent evidence for the operation of these properties of HMB RL in rats. This aim requires a multi-stage decision-making task, which can potentially engage action sequences and, on the other hand, the task must involve changes in the value of the outcomes, in order to evaluate the operation of the model-based component of the structure. Such a task has previously been developed in humans (Daw et al., 2011; Dezfouli & Balleine, 2013) (chapter 4), and here we translated it to an experimental protocol in rats.

We firstly show that, without any explicit instructions about the structure of the task (which obviously cannot be provided in rats), early in training subjects made decisions based on the assumption that the environment is composed of a single stage. Later in training they learned the true state-space of the task, and made decisions accordingly. After this stage, we show that subjects executed (first prediction), and evaluated (second prediction) action sequences as a single response unit and, furthermore, the selection of the actions was goal-directed (third prediction).

5.2 Materials and methods

5.2.1 Subjects and apparatus

Eight experimentally naive male Hooded Wistar rats served as subjects in each experiment (4 x 8 rats in total). All animals were housed in groups of three and handled daily for one week before training. Training and testing took place in 32 Med Associates operant chambers housed within sound- and light-resistant shells. The chambers were also equipped with a pellet dispenser that delivered one 45 mg pellet when activated (Bio-Serve). The chambers contained two retractable levers that could be inserted to the left and the right of the magazine. The chambers contained a white noise generator, a Sonalert that delivered a 3 kHz tone, and a solenoid that, when activated, delivered a 5 Hz clicker stimulus. All stimuli were adjusted to 80 dB in the presence of a background noise of 60 dB provided by a ventilation fan. A 3 W, 24 V house light mounted on the wall across from the levers and magazine illuminated the chamber. Microcomputers equipped with MED-PC software (Med Associates) controlled the equipment and recorded responses. Animals were food deprived one week before starting behavioral procedures. Feeding was such that rats were maintained at 90% of their free-feeding weight. The animals were fed after the training sessions of the day and had ad libitum access to tap water while in the home cage. Each training session (except magazine training sessions) started with insertion of the levers, and ended with the retraction of the levers.

5.2.2 Statistical analysis

We used R (R Core Team, 2014) and lme4 packages (Bates & Maechler, 2014) to perform a linear mixed effects analysis. In all the analyses, all the fixed effects (including intercepts) were treated as random effects varying across subjects. For the analysis that included more than one session, random effects were assumed to vary across sessions and subjects in a nested manner. Confidence intervals (CI) of the estimates were calculated using the 'confint' method of lme4 package with the 'Wald' parameter.

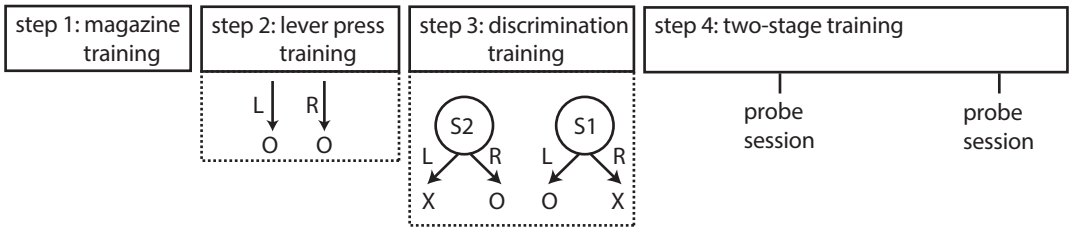


Figure 5.1 – Each experiment started with two magazine training sessions (step 1), followed by several lever training sessions (step 2), in which animals learned that pressing each lever (left and right levers corresponding to 'L' and 'R' in the figure) would delivered a reward. The next step was discrimination training (step 3), in which animals learned that when stimulus S1 was presented action 'L' should be taken to earn a reward, and when S2 was presented action 'R' should be taken to earn a reward. The results of this phase of the experiments is presented in section 5.3.1. The final step of the experiments was two-stage training, in which animals were trained on a two-stage decision-making task. This training phase comprised multiple training sessions, and in the middle or at the end of these training sessions several *probe sessions* were inserted. The result of the training sessions and probe sessions are presented in two separate sections. The result of the two-stage training is presented in section 5.3.2 and the result of the probe sessions is presented in sections 5.3.4, 5.3.5, 5.3.6 and 5.3.7. The structure of the two-stage task and the probe sessions is explained in Figure 5.2a,b,c.

5.2.3 Behavioral procedures

We conducted four experiments, each of which had several steps. The general structure of the experiments was similar and it is depicted in Figure 5.1. Below we explain each step and highlight the differences between the experiments.

Each experiment started with two sessions of magazine training in which grain-based food pellets were delivered into the food magazine. A total of 30 pellets were delivered, on average one pellet every 60 seconds. After this stage, animals were trained on a continuous reinforcement schedule, in which each lever press led to the delivery of a pellet. Each lever was trained in a separate session, and the total number of outcomes was limited to 60 per session. The total duration of the sessions was limited to 60 minutes. Animals were trained one session on the left lever, and one session on the right lever. In Experiments 1&3, animals received one more session on each lever, and in Experiments 2&4 they received one more session in which the levers were presented in alteration every ten minutes (each lever two times).

Next, animals were trained on the discrimination phase. Each session started with the presentation of a stimulus. The stimulus remained presented until the animal took an action (either pressing the left or right lever). Taking an action caused the stimulus to turn off, and it could lead to the delivery of the outcome or it could have no consequence depending on the action taken and presented stimulus. For one of the stimuli, taking the left action led to reward, and for the other stimulus taking the right action led to reward. Levers and stimuli were counterbalanced across subjects. After an action was chosen, there was a 60 second inter-trial interval (ITI), and after that, the next trial started with the presentation of the next stimulus, again chosen randomly. The duration of each session was 90 minutes, with no limit on the maximum number of earned rewards. In Experiments 1-3, the stimuli were tone and clicker, and in Experiment 4, the stimuli were constant and blinking house lights (5 Hz).

Next, subjects received training on the two-stage task depicted in Figure 5.2a in which animals first made a binary choice at stage 1 (signaled by the illumination of the house light in Experiments 1-3), after which they transitioned the stage 2 states, in which again they made another binary choice that could lead to reward delivery, or no-reward. After this, there was an ITI started, and then the next trial stared with the presentation of the house light (Experiments 1-3). The ITI in Experiments 1&2 was 20 seconds, in Experiment 3, 5-sec, and in Experiment 4 it was 0-sec. Stage 2 states were signaled by the stimuli trained on the previous phase of the experiment. In each trial, only one of the stage 2 states led to reward, whereas the other state would not lead to a reward irrespective of the choice of actions. The stage 2 state that earned a reward frequently switched between the states during the course of a session (Figure 5.2b). In Experiment 1, this switch occurred every time four outcomes were received since the last switch (with a maximum 40 outcomes in a session), which later in the training increased to every eight outcomes (with a maximum 48 outcomes in a session), as depicted in Figure 5.4b.

In Experiment 2, the switch occurred whenever a randomly selected number of outcomes were received since the last switch. This random number was uniformly drawn from range 8-16 (maximum 48 outcomes in a session). In Experiment 3, the switch occurred every fourth outcome received (maximum 50 outcomes in a session). In Experiment 4, the switch occurred

with a probability of 0.14 whenever the subject received an outcome (maximum 60 outcomes in a session). Furthermore, in Experiments 1&2 because the ITI was long, animals received a pre-training phase on the two-stage task in which the reward in the stage 2 state was fixed during a session, and was changed across sessions. Subjects received 10 training sessions in this condition. Similarly, in Experiment 2, subjects received two pre-training sessions in which they could earn a reward in both stage 2 states.

Animals were trained on the two-stage task (Figure 5.2b) for 69 sessions in Experiment 1, 57 sessions in Experiment 2, 60 sessions in Experiment 3, and 40 sessions in Experiment 4. In the middle of, or at the end of these training sessions, animals were given *probe sessions* in which stage 1 actions led to stage 2 states in a probabilistic manner. One of the stage 1 actions commonly led to one of the stage 2 states (80% of the time), and the other stage 1 action commonly led to the other stage 2 state (Figure 5.2c). For the last probe session in Experiments 1&3, the probability that stage 1 actions led to stage 2 states was 50%. The exact positions of probe sessions for Experiments 1-4 are depicted in Figure 5.4b, Figure 5.5b, Figure 5.6b, and Figure 5.7b respectively. The probe sessions are marked with '*' close to their labels. Hereafter, we will refer to training sessions by their labels. For example 'sx', refers to the x*th* session of training (starting from the very first training session, which was magazine training).

5.3 Results

We conducted four experiments, each of which had several phases. The general structure of the experiments was similar and is depicted in Figure 5.1. The aim of the first phase of training was to train animals to discriminate between the states of the environment (corresponding to step 3 in Figure 5.1); the results of this phase are presented in section 5.3.1 for all experiments. In the second phase, animals were trained on the two-stage task, which comprised multiple-training sessions. In the middle, or at the end of these training sessions several probe sessions were inserted. The results of the training sessions and probe sessions are presented in two separate sections. In the first section (section 5.3.2), we present the overall profile of the

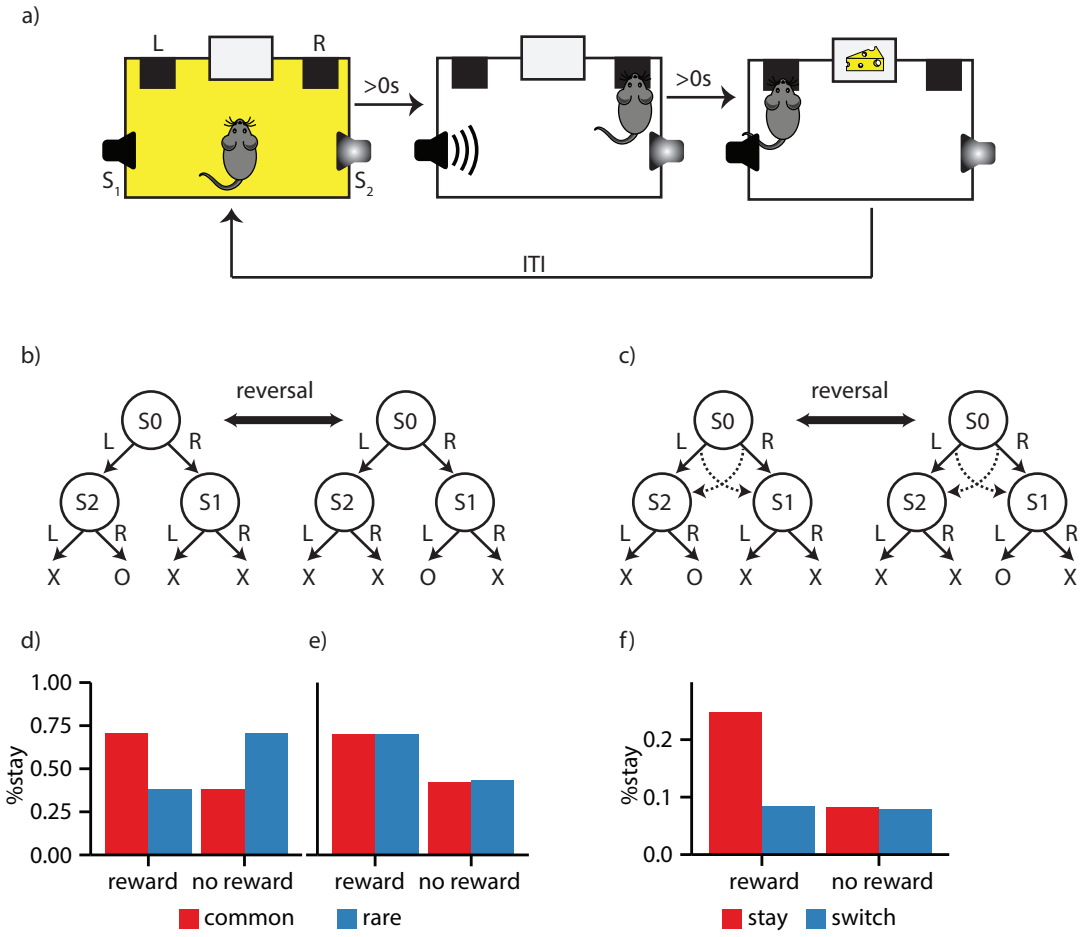


Figure 5.2 – (a) Flow of events in the two-stage task. Trials started with the illumination of the house light which signaled state S0 (Experiments 1-3). After an action was taken ('L' or 'R'), the house light turned off, and a stimulus started (S1 or S2). Next, subjects could take another action ('L' or 'R'), which could lead to the delivery of the outcome, or it might have no outcome. (b) Stage 1 actions led to the stage 2 stimuli in a deterministic manner. The rewarding stage 2 state changed frequently (reversal). 'O' represents outcome, and 'X' represents no-outcome. (c) In probe sessions, stage 1 actions led to the stage 2 states in a probabilistic manner. Taking action 'L' led to state 'S2' commonly (80% of the times), and to state 'S3' rarely (dashed lines). (d) Prediction of a pure model-based RL, which decides between choosing actions 'L' and 'R' at stage 1. This model predicts an interaction between staying on the same stage 1 action, and earning a reward in the previous trial. (e) Prediction of a pure hierarchical model-based RL which chooses between different action-sequences at stage 1 (LL, LR, RL, RR). This model predicts the main effect of whether reward was received in the previous trial. (f) Prediction of hierarchical model-based RL regarding the stage 2 choices, when the stimulus in the next trial is different from the current trial. This model predicts that when the previous trial is rewarded, if the subject stays on the same stage 1 action (stay), it will also stay on the same stage 2 action, even if the other stage 2 action should be selected. In panels (d,e,f) numbers are arbitrary and for the purpose of illustration.

animals' choices over the entire training period of all four experiments, and we establish that over the course of training, rats learned that the task is a two-stage decision-making process (not a single stage task that turns out to be the rats' a priori assumption). Next, we focus on the probe sessions which provided specific predictions regarding HMB RL framework (section 5.3.3), and we present the results of Experiments 1-4 in sections 5.3.4, 5.3.5, 5.3.6 and 5.3.7 respectively.

5.3.1 Learning to discriminate between states

In this phase of training, subjects learned to discriminate between the states of the environment (S1 and S2). States are signaled by the stimuli, e.g., presence of the tone signaled S1, and the clicker signaled S2. The basis of the discrimination was the action that should be taken in each state to earn a reward, i.e., in S1 pressing the left lever (action 'L' hereafter) was rewarded, and in S2 action 'R' was rewarded. The probability of taking the correct action for each session is depicted in Figure 5.4a, Figure 5.5a, Figure 5.6a, and Figure 5.7a respectively for Experiments 1-4, which shows that animals learned to take the correct action. This is indicated by the results of a logistic regression analysis over all the sessions of each experiment, which revealed a significant effect of intercept on taking the correct action (Experiment 1 :β=2.1 (CI: 1.51, 2.71), SE=0.30,p<1e-11; Experiment 2: β=1.79 (CI: 1.09,2.50), SE=0.35,p<1e-6; Experiment 3: β=1.58 (CI: 0.80, 2.37), SE=0.40,p<e-4; Experiment 4: β=0.89 (CI: 0.60, 1.18) SE=0.14,p<1e-8). In Experiments 1-3, the tone and clicker were used as stimuli, but in Experiment 4, a blinking house light, and a constant house light were used as stimuli (for the reasons described below).

5.3.2 Learning a two-stage representation of the task

This phase of the training process had two aims. The first aim was for the subjects to learn the two-stage logic of the task, i.e., the outcomes are not only the result of their immediate actions (stage2 actions), but stage 1 actions also matter. As we will show, without training, animals attribute outcomes to the most recent actions, ignoring the two-stage structure of the environment. The second aim was to train subjects to select actions flexibly based on the

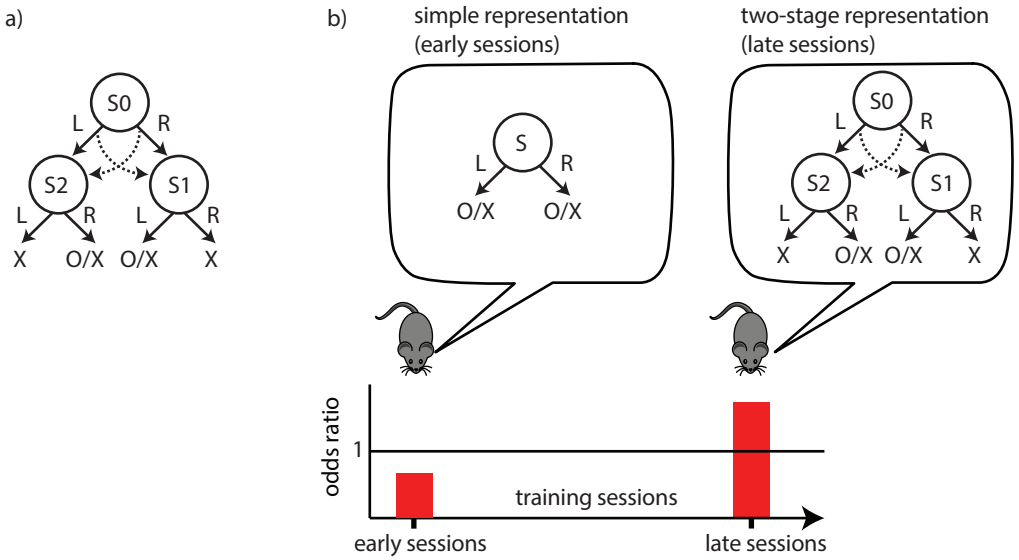


Figure 5.3 – a) The task has two stages, b) however, animals are not aware of that and they need to learn this by experiencing the task. Data shows that early in training animals assume the task has only one stage. As a result they repeated the most rewarded action in the next trial. It can be shown that this kind of representation predicts that the odds ratio of the probability of staying on the same stage 1 action will be below one (please see the text for an example). Later in the training after the animals built a correct representation of the task, then the odds ratio of the probability of staying on the same stage 1 action went above 1.

feedback they received in recent trials, i.e., promoting high learning rates. To achieve this aim, we designed a dynamic environment that deprived animals of rewards if they were selecting actions solely based on the most recent reinforced actions (the first aim), or when actions were rigid and not guided by feedback (the second aim). The flow of events in the task is depicted in Figure 5.2a. Subjects first made a binary choice at stage 1 (stage S0), and then they transferred to either the S1 or S2 states. Then they made another binary choice, which led to either reward delivery, or no-reward. After an inter-trial interval, the next trial started with presenting the stimulus that signaled state S0.

Reward of each action and transition between states is depicted Figure 5.2b. Taking action 'L' at stage 1 led to state S2 at the second stage, and taking action 'R' at the first stage led to state S1. Taking action 'L' in S2, or 'R' in S1 was never rewarded (consistent with the discrimination training phase), however the two other actions ('R' in S1, and 'L' in S2) could be rewarded. In each trial, only one of the states led to reward, and the other state would not lead to a reward

irrespective of the choice of actions. The reward-earning state changed frequently after several outcomes were received. As a result of this design, taking a particular stage 1 action earned outcomes for certain number of times, but after that, the rat needed to switch to the other action, otherwise it would not have earned any reward at all.

For the illustration of the dynamic of subjects' choices across training sessions, we first looked at the effect of reward on staying on the same stage 1 action. In all the subsequent analysis, we only included trials in which subjects in the previous trial made a correct discrimination at stage 2, because it was unclear how subjects updated the value of state and actions when they did not receive a reward because of an incorrect discrimination at stage 2 (which was not rewarded in any case).

Results of Experiments 1-4 are depicted in Figure 5.4b, Figure 5.5b, Figure 5.6b, and Figure 5.7b respectively. For each session of training, the figures show the odds ratio of staying on the same stage 1 action after earning a reward on the previous trial. Odds ratios were calculated using logistic regression analysis on the effect the reward had on staying on the same stage 1 action in the next trial. At the beginning of the training, as Figure 5.5b, Figure 5.6b, and Figure 5.7b show, subjects not only did not show a tendency to take the same action after earning a reward in the previous trial, but they showed a tendency to switch to the other action. This effect is statistically significant in the first five sessions of Experiment 2 (sessions s20 to s24; β=-0.25 (CI: -0.37, -0.13), SE=0.06, p<1e-4), Experiment 3 (sessions s15 to s19; β=-0.14 (CI: -0.29, -0.007), SE=0.07, p=0.039), and Experiment 4 (sessions s31 to s35; β=-0.68 (CI: -0.92, -0.45), SE=0.11, p<1e-8); but, for Experiment 1, although the effects were in the same direction, they are not statistically significant (sessions s26 to s30; β=-0.06, SE=0.04, p=0.161), probably because of the pre-training that animals received in this experiment. In fact, this pattern of choices (Experiments 2-4) is exactly what we expect to see when subjects are treating the task as having only a single stage (Figure 5.3): subjects repeat the stage 2 action that was rewarded in the previous trial, in stage 1 of the next trial, i.e., they attribute the reward to the most recent action. This resembles to switching to the other stage 1 action after receiving a reward. For example, assume that a subject has performed action 'L' at stage 1 and 'R' at stage 2 and has

received a reward. In the next trial, since they are attributing the reward to the most recent action ('R'), they repeat this action in the next trial, which makes it seems as though they have switched to the other action at stage 1.

Another indicator of whether actions are selected under a single-stage, or a two-stage representation of the task, is the pattern of magazine entries. Figure 5.4c, Figure 5.5c, Figure 5.6c, and Figure 5.7c show the probability of magazine entry depending on whether subjects stayed on the same stage 1 action and whether the previous trial was rewarded. As the figures show, whenever the previous trial is rewarded, if the subject takes the rewarded stage 2 action of the previous trial in stage 1 of the next trial (or equivalently switch to the other stage 1 actions), then they enter the magazine after that action, suggesting that they treated the task as a single stage decision process. This is indicated by the significant interaction between staying on the same stage 1 action and earning a reward in the previous trial. In the first five sessions of training, this effect was significant in all of the experiments: Experiment 1 (sessions s26 to s30; β=-0.25 (CI: -0.38, -0.12), SE=0.06, p=1e-4), Experiment 2 (sessions s20 to s24; β=-0.14 (CI: -0.25, -0.03), SE=0.05, p=0.012), Experiment 3 (sessions s15 to s19; β=-0.25 (CI: -0.37, -0.14), SE=0.05, p<1e-4), and Experiment 4 (sessions s31 to s35; β=-0.19 (CI: -0.33, -0.06), SE=0.06, p=0.004).

5.3.3 Predictions of HMB RL at stage 1 and stage 2 choices

In the previous phase of training, stage 1 actions led to stage 2 actions in a deterministic manner, i.e., taking 'L' at stage 1 always led to state S2, and taking 'R' at stage 1 always led to state S1 (Figure 5.2b). These are called 'common' transitions. Both during and at the end of these training sessions, we inserted several probe sessions that also included some 'rare' transitions in which a stage 1 action led to the other stage 2 action, i.e., taking 'L' led to state 'S1', and taking 'R' led to state 'S2' (Figure 5.2c)(Daw et al., 2011). These are called rare transitions because they constitute 20% of the trials. Nevertheless, they are important because they allow us to detect the operation of the model-based and hierarchical decision-making by analyzing choices at stage 1 and stage 2 of the task.

Following (Daw et al., 2011), the main analysis regarding stage 1 choices is to look at the probability of staying on the same stage 1 action, depending on whether the previous trial was rewarded and whether the transition in the previous trial was common or rare. Pure model-based RL, which is choosing between actions 'L' and 'R' at stage 1, predicts that whenever subjects receive a reward in the previous trial, if the transition in the previous trial was common, then they will stay on the same stage 1 action in the next trial, but if the previous trial was a rare transition, then they will switch to the other stage 1 action in the next trial to reach the state in which they earned a reward. As such, the operation of model-based RL predicts an interaction between earning a reward in the previous trial and the transition type of the previous trial (Figure 5.2d).

The other side of the spectrum is a pure hierarchical model-based RL in which, at stage 1, model-based RL chooses between performing a sequence of actions: LL, LR, RL, RR. In this condition, if the performance of an action was rewarded in the previous trial (e.g., performance of the RR action sequence), then the subject takes the same action sequence in the next trial (RR), which predicts that the subject will also stay on the same stage 1 action in the next trial irrespective of the transition type of the previous trial. As such, involvement of pure hierarchical model-based RL predicts a main effect of reward (Figure 5.2e). Given these two extremes, a HMB RL, which chooses between all the actions and action sequences (i.e., L, R, LL, LR, RL, RR), predicts a main effect reward as well as an interaction between reward and transition.

In addition to the above pattern at stage 1, the HMB RL makes another prediction regarding the pattern of choices at stage 2 (Dezfouli & Balleine, 2013). Imagine that a subject has performed a LR action sequence and has received a reward. This means that after performing action 'L' at stage 1, the subject has transitioned to state S2, in which it has taken action 'R'. Hierarchical model-based RL predicts that, in the next trial, the subject will take the same action sequence LR, implying that it will take action 'L' at stage 1, and 'R' at stage 2, irrespective of which stage 2 state subjects end up in after taking action 'L'. This predicts that, even if after taking action 'L', the subject transitions to state S1 (by a rare transition) in which action 'L' should be taken,

the subject still has a tendency to take action 'R' to complete the action sequence. Thus, the prediction will be: whenever a trial is rewarded, if the subject stays on the same stage 1 action, it is probably repeating the previously rewarded action sequence and thus will repeat the same stage 2 action even if that is a different stage 2 state from the previous trial (Figure 5.2f). This pattern, predicts, therefore, an interaction between staying on the same stage 1 action, and whether a reward was earned. In the following, we present analyses of stage 1 and stage 2 choices in each experiment (Dezfouli & Balleine, 2013). Note that for the stage 2 results, only trials in which the stage 2 state differed from the previous trial are included for the analysis. Similar to the previous phase, only trials in which subjects made a correct discrimination in the previous trial are included in the subsequent analysis.

5.3.4 Experiment 1

Eight subjects were trained in a two-stage decision-making task that only included common transitions (Figure 5.2b, Figure 5.4b). This experiment involved eight probe sessions (Figure 5.2c), and in each test session the probability of rare transitions was 80%, except for the last session in which the probability of common and rare transitions was equal (%50) for the reasons that we will explain below. For each session, we analyzed the pattern of choices at stage 1 and stage 2 of the task, which are depicted in Figure 5.4d-h, and their statistical analyses are presented in Table 5.1.

At the beginning of training, in session s40, there is a significant interaction between reward and transition type (β=-0.28 (CI: -0.50, -0.05), SE=0.11, p=0.013), however with a negative coefficient, which seems to indicate that subjects are staying away from reward in a goal-directed manner (Figure 5.4d). However, it is straightforward to see that such a pattern emerges when subjects treat the task as a single stage decision process and so tend to repeat the most recently rewarded action. In contrast to this pattern, at the end of the training in session s87 (Figure 5.4g), there is a significant interaction between reward and transition, with a positive coefficient (β=0.62 (CI: 0.26, 0.98), SE=0.18, p=6e-4), which indicates that (1) subjects have learned the two-stage logic of the task, (2) they have learned the contingency between

Table 5.1 – Results of logistic regression analysis of stage 1, and stage 2 choices of Experiment 1. For the stage 1 choices, the analysis is focused on staying on the same stage 1 action in the next trial based on whether the previous trial was rewarded (rew) and whether the previous trial was common or rare (trans). 're:tr' is the interaction between reward, and transition type, and 'inter' refers to the intercept term. For stage 2 choices, the analysis focuses on staying on the same stage 2 actions, based on staying on the same stage 1 action (stay), and earning a reward in the previous trial (rew). 're:st' is the interaction between 'reward', and 'stay'. 'p' refers to p-value.

Stage 1 actions												
session		s32	s40	s49	s57	s66	s78	s87	s94	s94:1	s94:2	s94:3
inter	p	0.773	0.312	0.017	0.299	0.177	0.162	0.885	0.428	0.460	0.131	0.194
	SE	0.15	0.11	0.16	0.12	0.13	0.11	0.22	0.12	0.24	0.26	0.29
	β	0.04	0.11	0.38	0.12	0.18	0.15	-0.03	0.10	0.18	-0.39	0.37
rew	p	0.938	0.085	0.090	0.067	0.021	0.001	0.904	<0.001	0.356	0.003	0.007
	SE	0.13	0.11	0.13	0.14	0.11	0.12	0.20	0.12	0.20	0.29	0.23
	β	-0.01	0.20	0.22	0.26	0.27	0.39	0.02	0.46	0.19	0.85	0.62
trans	p	0.764	0.628	0.098	0.016	0.825	0.383	0.993	0.498	0.407	0.644	0.335
	SE	0.14	0.10	0.19	0.11	0.12	0.13	0.24	0.14	0.18	0.29	0.18
	β	0.04	-0.05	-0.31	-0.28	-0.02	-0.11	0.00	0.09	0.15	-0.13	0.17
re:tr	p	0.882	0.013	0.362	0.397	0.485	0.417	6.00e-04	<0.001	0.011	0.003	0.668
	SE	0.12	0.11	0.14	0.12	0.10	0.09	0.18	0.11	0.18	0.29	0.21
	β	0.01	-0.28	-0.12	0.10	-0.07	0.08	0.62	0.40	0.45	0.85	0.09
Stage 2 actions												
session		s32	s40	s49	s57	s66	s78	s87	s94	s94:1	s94:2	s94:3
inter	p	<1e-7	<1e-12	<1e-11	0.957	<1e-8	<1e-14	<1e-9	<1e-16	0.981	<1e-4	0.999
	SE	0.44	0.41	0.43	169.6	0.47	0.35	0.45	0.29	329	0.57	>100
	β	-2.44	-2.93	-2.99	-9.07	-2.84	-2.82	-2.96	-2.67	-7.86	-2.47	-17.63
rew	p	0.933	0.938	0.148	0.981	0.383	0.137	0.169	0.010	0.987	0.924	1
	SE	0.32	0.37	0.45	169.6	0.43	0.32	0.45	0.26	329	0.51	>100
	β	0.02	0.02	0.66	3.95	0.37	0.48	0.63	0.62	5.33	0.04	5.82
stay	p	0.057	0.852	0.074	0.969	0.401	0.386	0.891	0.79	0.987	0.387	1
	SE	0.40	0.52	0.48	169.6	0.40	0.33	0.46	0.27	329	0.50	>100
	β	0.77	-0.09	0.86	-6.56	0.33	0.29	0.06	0.07	-5.21	0.43	2.95
re:st	p	0.105	0.515	0.957	0.979	0.089	0.200	0.443	0.85	0.990	0.809	1
	SE	0.32	0.35	0.45	169.6	0.39	0.37	0.49	0.28	329	0.50	>100
	β	0.51	0.23	-0.02	4.38	0.67	0.47	0.37	-0.05	4.29	-0.12	5.88

stage 1 actions and stage 2 states. In between these two sessions, there are two sessions (s66 and s78) in which subjects treat the task as a two-stage task, but actions are not yet guided by the contingency between stage 1 actions and stage 2 states (Figure 5.4e,f). This is indicated by the significant main effect of reward in these two sessions (Table 5.1/stage1/s66 and s78).

Such a pattern of responses in sessions s66, and s78 can be a product of hierarchical decision-making. However, the hierarchical account, as mentioned earlier, predicts an interaction between stay and reward at stage 2 choices, which is not observed in this dataset (Figure 5.4i; Table 5.1/stage2/s66, s78), which could reflect inhibition of the performance of action sequences (we will elaborate this effect in the next experiments). Equally likely, this pattern of choices could be the result of model-free RL, which predicts staying on the stage 1 choice after earning reward (a main effect of reward at stage 1) without necessarily staying on the stage 2 choice. In any case, these two sessions seem to be intermediary stages in the development of model-based actions, rather than a signature of other RL systems. Following this logic, one can assume that, because the model-based RL is uncertain about the contingencies at this stage, it updates them with each experience, i.e., with each rare transition, the contingencies of stage 1 actions will be updated as if they lead to their rare outcomes, which predicts a main effect of reward at stage 1 and the absence of a reward x transition interaction.

For further investigation of the above issue, at the end of the training sessions we performed a probe session (s94) in which each stage 1 action led to each stage 2 action with equal probability. The rationale for this manipulation was to make the contingencies ambiguous for the subject and so render actions sensitive to their final outcomes. Figure 5.4h shows the result of this session, and Figure 5.4j-l shows the result of the break down of the trials to early, mid, and late trials (there were 48 outcomes in total). Over the whole session, the effects of reward and also the interaction between reward and transition type were significant (Table 5.1/stage1/s94). However, consistent with our expectation, early in training the behavior was only model-based (Table 5.1/stage1/s94:1), whereas in the middle of the session there was also an effect of reward (Table 5.1/stage1/s94:2), which is indicated by a reward by session period interaction (period 1: number of outcomes < 16, period 2: 32 >number of outcomes

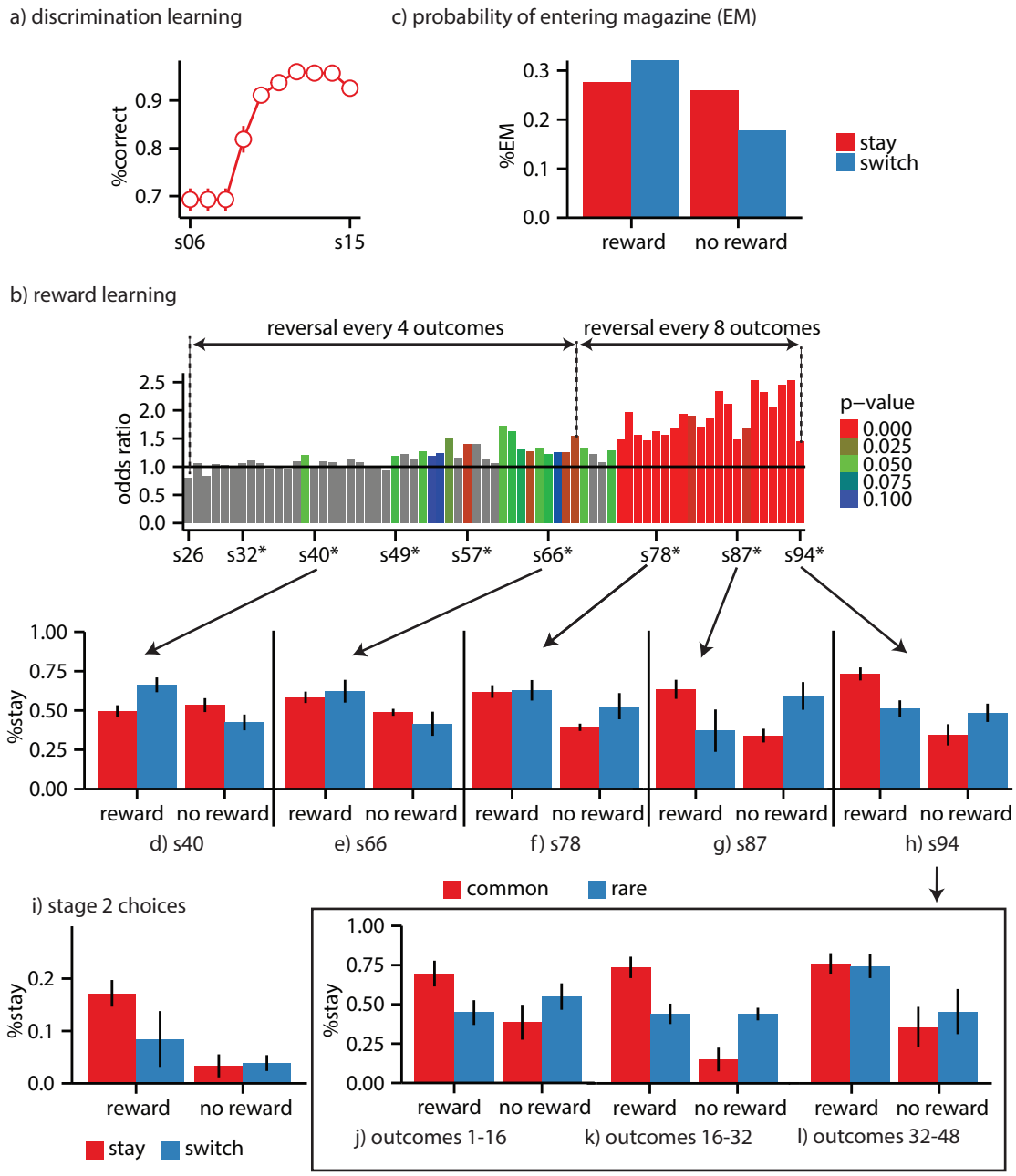


Figure 5.4 – Experiment 1. (a) Results of the discrimination training showing the percentage of the correct responses averaged over subjects. (b) Odds ratio of the probability of staying on the same stage 1 action after getting rewarded on the previous trial. Odds ratio=1 implies an equal preference for both actions. Sessions marked with '*' are the probe sessions which included both rare and common transitions. (c) The probability of entering the magazine as a function of staying (stay) or switching (switch) to the other stage 1 action, and whether the previous trial was rewarded (averaged over subjects and the first five sessions of the training; 8 * 5 data points contributed to each mean). (d-h, j-l) The probability of staying on the same stage 1 action, averaged over subjects, as a function of whether the previous trial was rewarded (reward/no reward), and whether the transition in the previous trial was common or rare. (i) Stage 2 choices in session s78. The graph shows the probability of staying on the same stage 2 action averaged over subjects, as a function of whether the previous trial was rewarded (reward/no reward), and whether subjects stayed on the same stage 1 action (stay/switch). Only trials in which the stage 2 state is different from the previous trial are included. Error-bars 1SEM.

>=16) (β=0.35 (CI: 0.29, 1.19), SE=0.16, p=0.030). Late in the session the model-based effect was removed from choices (Table 5.1/stage1/s94:3), which is indicated by a reward by transition by session period interaction (period 1: 32 >number of outcomes >=16, period 2: number of outcomes >=32) (β=-0.38 (CI: -0.74, -0.03), SE=0.18, p=0.033).

The results of this experiment revealed the development of model-based action, and how they are affected by manipulating contingencies in the environment. However, we were not able to see evidence for the engagement of HMB RL, in contrast to a previous study using a similar task in humans (Dezfouli & Balleine, 2013). As discussed earlier, in order for the hierarchical decision-making to emerge, animals need to consider the performance of action sequences LL, LR, RL, RR at stage 1. This requires that these action sequences be already formed by the concatenation of single actions, which takes place when the consecutive performance of single actions gets rewarded. However, the training conditions of this experiment only included common transitions, which means that only the performance of the sequences L->R, and R->L will be rewarded, whereas and the performance of the sequence L->L and R->R is never rewarded. As such, in order to promote the formation of these action sequences, in the next set of experiments the subjects were explicitly trained on a condition in which the performance of the LL and RR sequences earned reward.

5.3.5 Experiment 2

Eight subjects were trained in a two-stage decision-making task that only included common transitions (Figure 5.2b, Figure 5.5b). In the middle of these training sessions, animals were trained in a condition that only included rare transitions[1] (sessions s62 to s64; Figure 5.5b), which means that animals needed to perform the LL and RR sequences in order to earn a reward. At the end of the training sessions, the animals were given a probe session which included both common and rare transitions (80% common transitions).

[1] Note that in these sessions the definition of the rare transitions is similar to the previous sessions (i.e., action 'L' leading to state S1 and action 'R' leading to state S2), even though in these sessions rare transitions are more common.

Table 5.2 – Results of logistic regression analysis of stage 1, and stage 2 choices in Experiments 2-4. For the stage 1 choices, the analysis is focused on staying on the same stage 1 action in the next trial, based on whether the previous trial was rewarded (rew), and whether the previous trial was common or rare (trans). 're:tr' is the interaction between reward, and transition type, and 'inter' refers to the intercept term. For stage 2 choices, the analysis is focused on staying on the same stage 2 actions, based on staying on the same stage 1 action (stay) and earning a reward in the previous trial (rew). 're:st' is the interaction between 'reward', and 'stay'. 'p' refers to p-value. 'Expr' stands for 'Experiment'.

Stage 1 actions						
		Expr2	Expr3	Expr4		
session		s76	s74	s57	s58	s70
inter	p	<1e-8	0.84	0.149	0.019	0.375
	SE	0.13	0.12	0.14	0.20	0.18
	β	0.78	-0.02	0.21	0.47	-0.16
rew	p	<0.001	<1e-8	0.565	0.020	0.003
	SE	0.13	0.08	0.18	0.13	0.20
	β	0.45	0.49	0.10	0.32	0.59
trans	p	0.467	0.78	0.299	0.616	0.111
	SE	0.15	0.14	0.26	0.28	0.25
	β	-0.11	-0.04	0.27	-0.14	0.39
re:tr	p	0.190	0.56	0.174	0.432	<1e-4
	SE	0.14	0.08	0.36	0.16	0.21
	β	0.18	0.05	0.49	0.12	0.92
Stage 2 actions						
		Expr2	Expr3	Expr4		
session		s76	s74	s57	s58	s70
inter	p	<1e-13	<1e-15	0.456	0.830	<0.001
	SE	0.35	0.09	0.24	0.36	0.28
	β	-2.61	-0.86	-0.18	-0.07	0.96
rew	p	0.145	0.001	0.004	<0.001	<0.001
	SE	0.38	0.09	0.15	0.26	0.22
	β	0.56	0.28	0.42	0.98	0.85
stay	p	0.887	0.02	0.072	0.155	0.017
	SE	0.33	0.09	0.14	0.19	0.24
	β	-0.04	0.20	0.25	0.27	0.58
re:st	p	0.600	0.24	0.007	0.513	0.011
	SE	0.34	0.08	0.16	0.17	0.20
	β	0.18	0.09	0.43	0.11	0.52

Stage 1 choices are depicted in Figure 5.5d, and the results of the logistic regression analysis are presented in Table 5.2/stage1/expr2/s76, which indicates only the main effect of reward. This pattern of choices at stage 1 is consistent with the engagement of pure hierarchical decision-making. However, this account also predicts an interaction between staying on stage 1 actions, and receiving a reward in the previous trial. The results of this stage 2 action are depicted in Figure 5.5e, in which the mentioned interaction is not significant (Table 5.2/stage2/expr2/s76).

For a closer inspection of this effect, we broke down the trials into two kinds: (1) trials in which animals earned a reward from a common transition in the previous trial (i.e., they have performed a LR or RL action sequences), and therefore repeating that sequence in the next trial involved moving from one side of the box to the other, (2) trials in which animals earned a reward from a rare transition in the previous trial (i.e., they have performed a LL or RR action sequence in the previous trial) and, therefore, repeating that sequence involved pressing the same lever twice, without a need to move to the other side of the box. One can assume that, because repeating the action sequence in the first kind of trial takes longer, subjects are more likely to be able to inhibit an inappropriate ongoing action sequence, whereas in the later case, performance of the action sequence is faster, and is more likely to be executed. Figure 5.5f shows the probability of staying on the same stage 2 action, depending on whether the action sequence to be executed required switching to the other side of the training box or not. As the figure shows, when the previous trial was rewarded, and subjects stayed on the same stage 1 action, then in the second kind of trials, subjects were less able to inhibit the ongoing action sequence. This is indicated by a result of logistic regression on the probability of staying on the same stage 2 action as a function of whether the sequence performance required switching to the other side of the training box (β=-1.06 (CI: -1.87,-0.24), SE=0.41, p=0.010). Note in Figure 5.5f, some of the conditions occur only when there are two consecutive rare transitions (since only trials in which the stage 2 state is different from the previous trial are included in the analysis), which are sparse. As such, in the next experiment, in addition to other changes, we gave a test in which the probability of common and rare transitions was equal (50%).

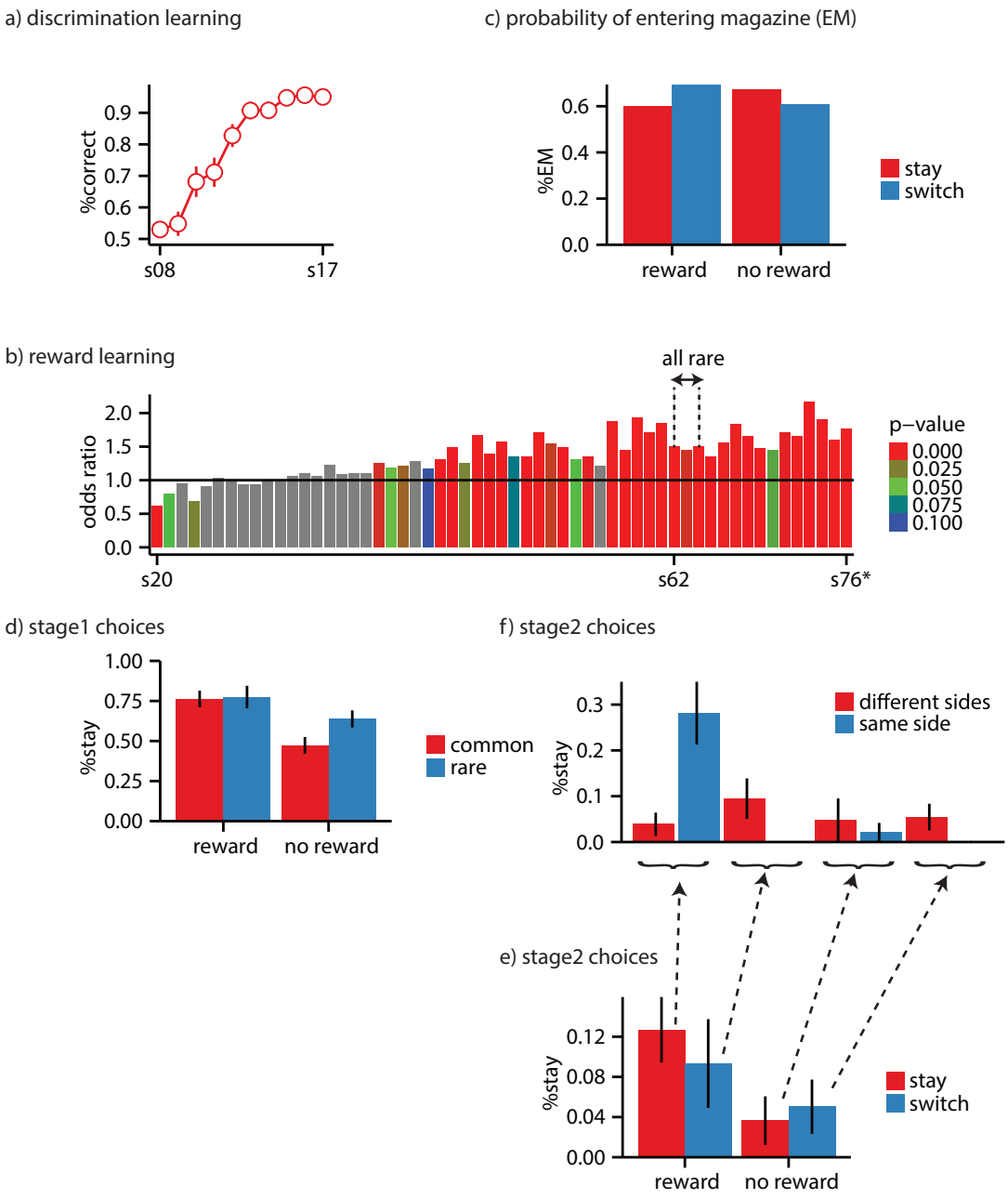


Figure 5.5 – Experiment 2. (a) Results of the discrimination training, showing the percentage of the correct responses. (b) Odds ratio of the probability of staying on the same stage 1 action after getting rewarded in the previous trial. (c) The probability of entering the magazine as a function of staying (stay) or switching (switch) to the other stage 1 action, and whether the previous trial was rewarded (first five sessions of the training; 8 * 5 data points contributed to each mean). (d) The probability of staying on the same stage 1 action (session s76) as a function of whether the previous trial was rewarded (reward/no reward) and whether the transition in the previous trial was common or rare. (e) The probability of staying on the same stage 2 action (session s76), as a function of whether the previous trial was rewarded (reward/no reward) and whether subjects stayed on the same stage 1 action (stay/switch). (f) The probability of staying on the same stage 2 action as a function of whether the previous trial was rewarded (reward/ no reward), whether the same stage 1 action was taken in the current trial, (same/switch), and whether the performance of the action sequence required switching to the other side of the box or not (same side/different sides). Only trials in which the stage 2 state was different from the previous trial are included in panels e,f. Error-bars 1SEM.

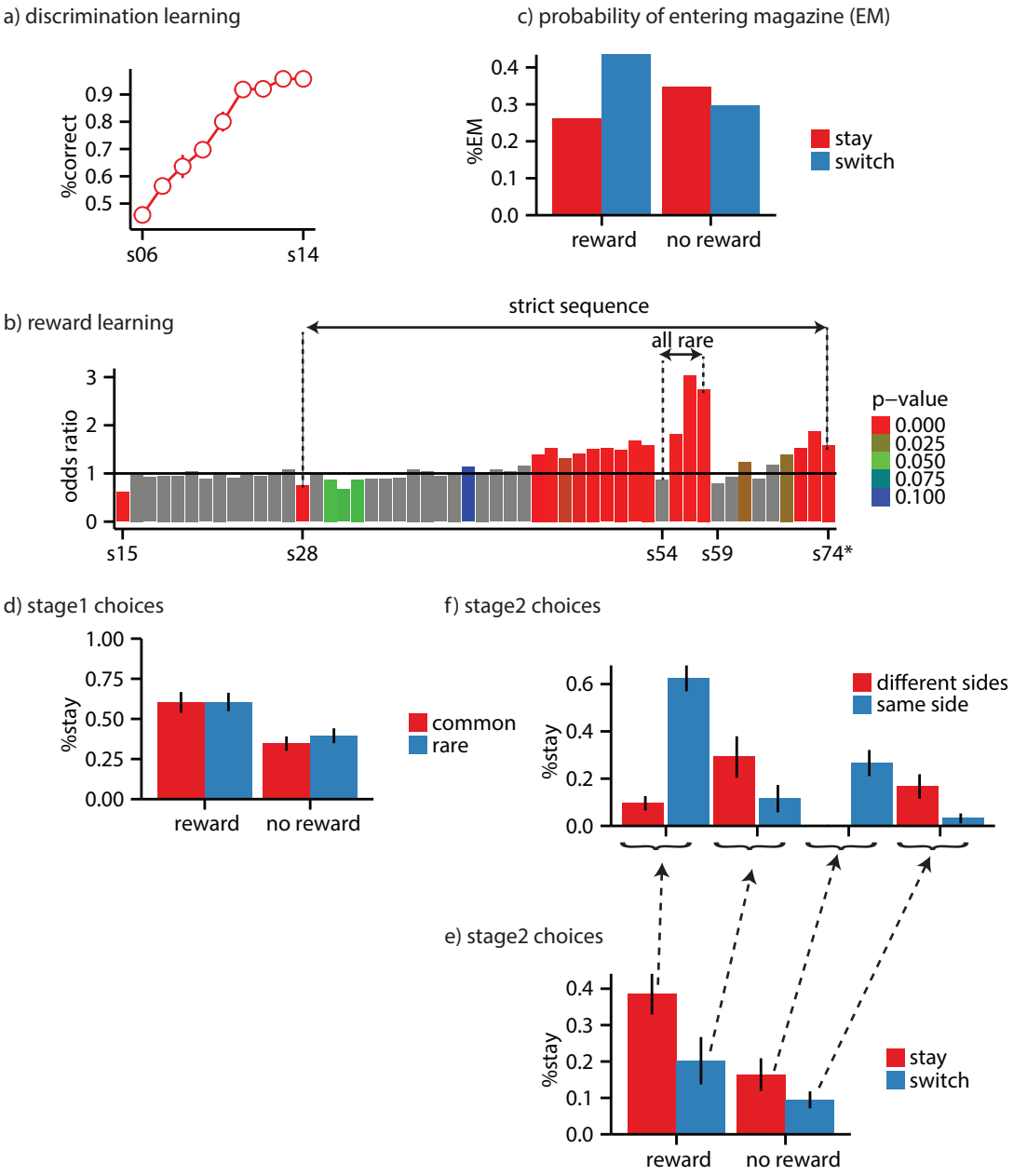


Figure 5.6 – Experiment 3. (a) Results of the discrimination training, showing the percentage of the correct responses. (b) Odds ratio of the probability of staying on the same stage 1 action after getting rewarded in the previous trial. (c) The probability of entering the magazine as a function of staying (stay) or switching (switch) to the other stage 1 action, and whether the previous trial was rewarded (first five sessions of the training; 8 * 5 data points contributed to each bar). (d) The probability of staying on the same stage 1 action (session s74) as a function of whether the previous trial was rewarded (reward/no reward), and whether the transition in the previous trial was common or rare. (e) The probability of staying on the same stage 2 action (session s74), as a function of whether the previous trial was rewarded (reward/no reward), and whether subjects stayed on the same stage 1 action (stay/switch). (f) The probability of staying on the same stage 2 action, as a function of whether the previous trial was rewarded (reward/no reward) whether the same stage 1 action was taken in the current trial, (same/switch), and whether the performance of the action sequence required switching to the other side of the box or not (same side/different sides). Only trials in which the stage 2 state was different from the previous trial are included in panels e,f. Error-bars 1SEM.

5.3.6 Experiment 3

The results of the previous experiment indicated that the lack of evidence for the engagement of hierarchal RL at stage 2 choices was likely because the performance of the action sequences that require pressing levers on both sides of the box is prone to high inhibition. Switching from one side of the box to the other side of the box in some trials was intervened by a magazine response (animals check the magazine before executing the second component of the sequence) which can potentially make such action sequences prone to interruption and inhibition. Based on this, in Experiment 3, if a subject made a magazine entry response after a stage 1 action and before the stage 2 action, then the trial aborted (strict sequences in Figure 5.6b).

Eight subjects were trained in a two-stage decision-making task that only included common transitions (Figure 5.2b, Figure 5.6b). The result of the stage 1 choices is depicted in Figure 5.6d, and the statistical analyses are presented in Table 5.2/stage1/expr3/s74, which indicates the main effect of reward, which potentially indicates that rats were using hierarchical RL. The results of the stage 2 choices are presented in Figure 5.6e, and the statistical analyses are presented in Table 5.2/stage2/expr3/s74. Similar to Experiment 2, choices at stage 2 did not indicate the operation of hierarchical decision-making; however, when animals were rewarded and stayed on the same stage 1 action, then they seemed to be able to inhibit the action sequences more if the sequence performance did not require switching to the other side of the box (Figure 5.6f), as suggested by the result of a logistic regression assessing the effect of switching sides on staying on the same stage 2 action (β=-1.4 (CI:-1.79,-.92), SE=0.26, p<1e-7).

5.3.7 Experiment 4

The result of Experiment 2 and 3 suggested increased inhibition of action sequences by the stage 2 stimulus when subjects needed to switch from one side of the box to the other side. This effect should be predicted to diminish if less salient stimuli were used to signal stage 2 states. Based on this, in this experiment instead of the clicker and tone signaling the stage 2

states, we used a constant and a blinking house light to signal the stage 2 states. We trained eight subjects in these conditions and, as expected, it took longer for the subjects to learn to discriminate between these two stimuli, compared to the tone and clicker (Figure 5.7a vs e.g., Figure 5.6a), which is indicated by the result of a logistic regression that showed a significant effect of experiment (Experiment 4 vs Experiment 3) on correct discrimination in the first 10 sessions (β=1.39 (CI: 0.60, 2.18), SE=0.40, p=5e-4).

After training (Figure 5.7b), animals were given three probe sessions (s57, s58, s70). Table 5.2/stage1/expr4/s57,s58,s70 presents the statistical analysis of the sessions. In session s57, animals did not show any sign of model-based or hierarchical decision-making at stage 1 (main effect of reward p>0.565; reward-transition interaction p>0.174), and there was no indication of discrimination between the states at stage 2 choices (Table 5.2/stage2/expr4/s57, s58:intercept, p>0.25). As such, animals were given another test (s58), which indicated a main effect of reward (Table 5.2/stage1/expr4/s58), however, at stage 2, subjects did not show any discrimination between stage 2 states (Table 5.2/stage2/expr4/s58:intercept; p>0.25). As such, subjects were given nine more training sessions (s59 to s69), and then they were given a further probe test (s70).

The results of stage 1 actions are presented in Figure 5.7d, and statistical analysis of stage 1 choices are presented in Table 5.2/stage1/expr4/s70. As the table shows, there was a significant reward-transition interaction (p<1e-4), which indicates the engagement of the model-based controller. In addition to that, the main effect of reward was also significant, which potentially indicates the engagement of HMB RL. Figure 5.7e presents choices at stage 2, which were consistent with the predictions of HMB RL about stage 2 choices, i.e., subjects tended to take the same stage 2 action when the previous trial was rewarded and they stayed on the same stage 1 actions. This is indicated by a significant interaction between staying on the same stage 1 action and reward in the previous trial (β=0.52 (CI: 0.11, 0.92), SE=0.20, p=0.011).

Comparison of RL models. The above behavioral analysis suggests that HMB RL guides choices. However, the analysis only reveals the effect of feedback from one-trial-back on current

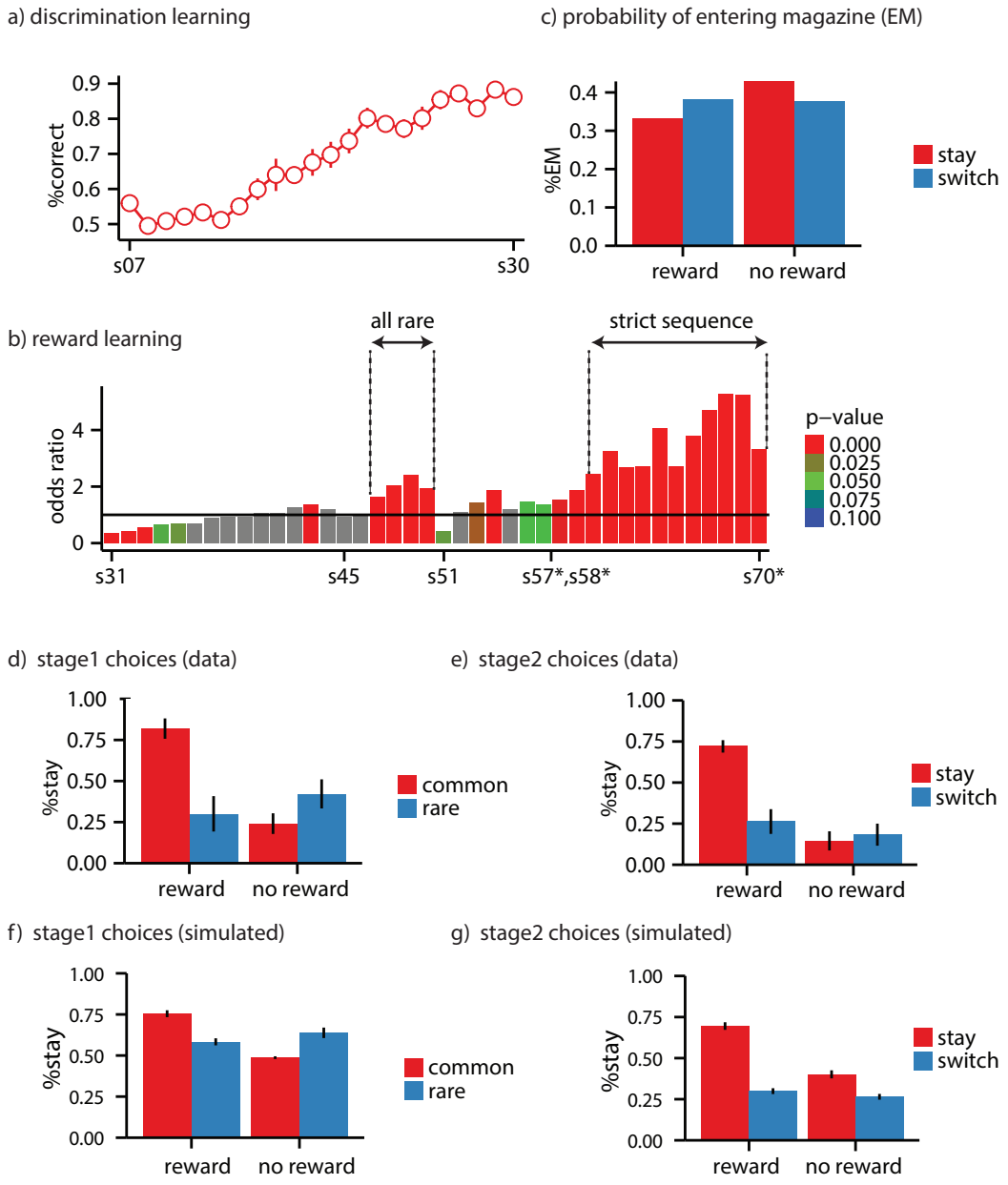


Figure 5.7 – Experiment 4. (a) Results of the discrimination training, showing the percentage of the correct responses. (b) Odds ratio of the probability of staying on the same stage 1 action after getting rewarded in the previous trial. (c) The probability of entering the magazine as a function of staying (stay) or switching (switch) to the other stage 1 action, and whether the previous trial was rewarded (first five sessions of the training; 8 * 5 data points contributed to each bar). (d) The probability of staying on the same stage 1 action (session s70) as a function of whether the previous trial was rewarded (reward/no reward) and whether the transition in the previous trial was common or rare. (e) The probability of staying on the same stage 2 action (session s70), as a function of whether the previous trial was rewarded (reward/no reward) and whether subjects stayed on the same stage 1 action (stay/switch). (f) Simulation of stage 1 choices, and (g) stage 2 choices using the best fitted parameters for each subject.

choices, whereas choices are presumably guided by the history of prior feedback, and not only the feedback received on the previous trial. As such, in order to analyze the full profile of choices, we compared different variants of the RL algorithms using a Bayesian modeling comparison procedure with the aim of confirming that subjects are using HMB RL as opposed to other non-hierarchical alternatives. Based on previous studies using similar decision-making tasks in humans (Daw et al., 2011; Dezfouli & Balleine, 2013; Gläscher, Daw, Dayan, & O'Doherty, 2010), we compared four families of RL algorithms: (1) a non-hierarchical model-based RL family, (2) a model-free RL family, (3) a hybrid model-free, model-based RL family, and (4) a hierarchical model-based RL family. In total we considered 328 different models, where each family consisted of several members with different degrees of freedom (please see section 5.5 for details of each model). We then calculated the log-model evidence for each model given the choices of subjects. The results indicated that the differences in model evidence (Bayes factor) between the best fitting model of the HMB RL family, and the non-hierarchical model-based RL (family 1)/model-free RL (family 2)/ model-free, model-based hybrid RL (family 3), were 50, 47, and 51 respectively. In Bayesian model comparison literature, Bayes factors greater than 10 are considered to be strong evidence (Jeffreys, 1961), and thus the above results of model comparison reveal that subjects were very likely utilizing HMB RL action selection.

We then simulated the HMB RL model in the task conditions with the best fitting parameters (please see section 5.5 for how parameters were obtained) and analyzed stage 1 and stage 2 choices. Analysis of stage 1 choices (Figure 5.7f) revealed a significant main effect of reward (β=0.25 (CI: 0.19, 0.32), SE=0.03, p<0.001), and a significant interaction between whether the previous trial was rewarded, and the transition type of the previous trial (β=0.29 (CI: 0.19, 0.38), SE=0.04, p<0.001). Analysis of stage 2 choices (Figure 5.7g), revealed a significant interaction between staying on the same stage 1 action and earning a reward on the previous trial (β=0.52 (CI: 0.33, 0.70), SE=0.09, p<0.001), which is consistent with predictions from HMB RL regarding stage 2 choices.

5.4 Discussion

Model-based reinforcement learning provides a computational framework for studying goal-directed actions. This notion, however, assumes that the states and available actions are given to an agent a priori, whereas the ability to (i) create new states to be able to correctly represent task space, and (ii) create complex actions out of simpler actions to make hierarchical decisions, is also an integral part of goal-directed decision-making. Using a sequential decision-making task in rats, we firstly provided behavioral evidence that, at the beginning of training, animals make decisions based on a simple representation of the environment, but with sufficient training, they learn the sequential nature of the task and build the correct state-space representation of the environment as required for the operation of model-based RL. Secondly, within this multi-stage representation, we show that animals chunk actions together and create action sequences that are then integrated into model-based RL, revealing a hierarchical organization of decision-making processes.

The results of the experiments showed that the operation of HMB RL requires certain experimental conditions, such as pre-training on action sequences and low intrusion of other factors on the performance of action sequences. Without these conditions, choices at stage 1 are consistent with the operation of HMB RL, but not choices at stage 2. This would have been due to inhibition of action sequences. Alternatively this pattern of choices could also indicate the operation of another RL system, such as model-free RL, instead of HMB RL with interrupted sequences. Within this alternative framework, choices are a mixture of goal-directed actions (model-based), and model-free actions that are guided by their 'cached' (as opposed to their current values) (Daw et al., 2005). Our results are ambivalent with respect to this interpretation, but since, in other conditions, there is positive evidence for HMB RL, it seems to be more parsimonious to interpret this result in terms of the inhibition of action sequences, rather than being the output of a model-free system.

Besides theoretical considerations regarding the hierarchical component of the task, the current decision-making task provides a behavioral protocol for investigating the acquisition

of the correct state-space in a multi-stage decision- making task, which can be used to uncover the neural substrates of this process. Furthermore, using a computational modeling procedure, the task provides a quantitative measure of the inhibition of action sequences by other monitoring processes, which again can be used to uncover the neural substrates of this process.

5.5 Appendix: computational modeling

As we mentioned in the article, we compared four different families of the RL algorithms: (1) a non-hierarchical model-based RL family, (2) a model-free RL family, (3) a hybrid model-free, model-based RL family, and (4) a hierarchical model-based RL family. Each family had several instances, that we described them below.

5.5.1 Simulation environment

We assumed that the environment has five states; the initial state denoted by S_0, stage 2 states denoted by S_1 and S_2, the reward state denoted by S_{Re} and no-reward state denoted by S_{NR}. For the case of non-hierarchical models, we assumed that actions L and R are available in states S_0, S_1 and S_2, and for the case of hierarchical models, we assumed actions L, R, LR, LL, RL, and RR are available in stats S_0, and actions L and R, are available in states S_1 and S_2.

5.5.2 Model-based RL (MB)

The model-based system works by learning the model of the environment, and then calculating the value of actions using the learned model. The model of the environment is composed of the transition function ($T(.)$), and the reward function ($R(.)$). We denote the transition function with $T(s'|a,s)$ which is the probability of reaching state s' after executing action a in state s. We assume that the transition function at the first stage is fixed ($T(S_1|R,S_0) = 0.8$ and $T(S_2|L,S_0) = 0.8$) and it will not change during learning. For other states, after executing

action a in state s and reaching state s', the transition function updates as follows:

$$\forall s'' \in \{S_{Re}, S_{NR}\} : T(s''|s,a) = \begin{cases} (1-\eta)T(s''|s,a) + \eta & : s' = s'' \\ (1-\eta)T(s''|s,a) & : s' \neq s'' \end{cases} \tag{5.1}$$

Where $\eta (0 < \eta < 1)$ is the update rate of the state-action-state transitions. For the reward functions, we assumed that the reward at state S_{Re} is one ($R(S_{Re}) = 1$), and zero in all other states.

Based on the above reward and transition functions, the goal-directed value of taking action a in state s is as follows:

$$\forall s \in \{S_0, S_1, S_2\} : V^G(s,a) = \sum_{s'} T(s'|s,a) V^G(s') \tag{5.2}$$

Where:

$$V^G(s) = \begin{cases} \max_a V^G(s,a) & : s \in \{S_0, S_1, S_2\} \\ R(s) & : s \in \{S_{Re}, S_{NR}\} \end{cases} \tag{5.3}$$

Finally, the agent uses the calculated values to choose actions. The probability of selecting action a in state s, denoted by $\pi(s,a)$, will be determined according to the soft-max rule:

$$\pi(s,a) = \frac{e^{\beta(s)V^G(s,a)+\kappa(s,a)+d(s,a)}}{\sum_{a'} e^{\beta(s)V^G(s,a')+\kappa(s,a')+d(s,a')}} \tag{5.4}$$

The above equation reflects the fact that actions with higher values are more likely to be selected. The $\beta(s)$ parameter controls the rate of exploration; the parameter $\kappa(s,a)$ is the action preservation parameter and captures the general tendency of taking the same action as the previous trial (Ito & Doya, 2009; Lau & Glimcher, 2005). Finally, the term $d(s,a)$ represents the tendency of the subjects to take the discriminative actions at the stage 2 states (taking action R in S_2 and action L in S_1). A positive value for this parameter entails that a subject has a tendency to take the discriminative action at stage 2 states. Please note that the effect of this

parameter is on top of the effect of values at stage 2 states.

For the exploration parameter, we assume that $\beta(s) = \beta_1$ if $s = S_0$ and $\beta(s) = \beta_2$ if $s \in \{S_1, S_2\}$. For the perseveration parameter we assumed that if $s = S_0$ and a being the action taken in the previous trial in the S_0 state, then $\kappa(s,a) = k$, otherwise it will be zero. Finally, for the discrimination parameter, we assume $d(s,a) = \phi$ if $(s,a) \in \{(S_1,L),(S_2,R)\}$, and $d(s,a) = 0$ otherwise.

In the most general form, all the parameters $(\beta_1, \beta_2, \eta, k, \phi)$ were treated as free parameters. We also generated eight variants by (1) setting $\beta_1 = \beta_2$ (i.e., rate of exploration at stage 1 and stage 2 states are the same), (2) setting $k = 0$ (there is no tendency to perseverate on the previously taken actions), and (3) setting $\phi = 0$ (there is not tendency to take the discriminative action at stage 2).

5.5.3 Model-free RL (MF)

Here we used Q-learning (Watkins, 1989) for model-free learning. After taking action a in state s, and reaching state s', model-free values update as follows:

$$Q^H(s,a) \leftarrow Q^H(s,a) + \alpha(V^H(s') - Q^H(s,a)) \tag{5.5}$$

Where $\alpha (0 < \alpha < 1)$ is the learning rate, which can be different in stage 1 and stage 2 actions. For the first stage actions (actions executed in S_0), $\alpha = \alpha_1$, and for the second stage actions $\alpha = \alpha_2$. In equation 5.5, $V^H(s)$ is the value of the best action in state s:

$$V^H(s) = \begin{cases} \max_a Q^H(s,a) & : s \in \{S_0, S_1, S_2\} \\ R(s) & : s \in \{S_{Re}, S_{NR}\} \end{cases} \tag{5.6}$$

In the trials in which the best action is executed in $s \in \{S_1, S_2\}$ the value of the action executed in state S_0 also updates according to the outcome. If a was to be the action which was taken in

S_0, a' the action taken in s, and s' the state visited after executing a', values update as follows:

$$Q^H(S_0,a) \leftarrow Q^H(S_0,a) + \alpha_1\lambda(V^H(s') - Q^H(s,a')) \tag{5.7}$$

Where $\lambda(0 < \lambda < 1)$ is the reinforcement eligibility parameter, and determines how the first stage action values are affected by receiving the outcome after executing the second stage actions. The action selection method, and variants of this form of learning are described in the next section.

5.5.4 Model-free, model-based hybrid RL (MF-MB)

This model is a combination of model-free RL, and model-based RL, in which final values are computed by combining the values provided by model-free and model-based processes:

$$V(s,a) = wV^G(s,a) + (1-w)Q^H(s,a) \tag{5.8}$$

Were $w(0 < w < 1)$ determines the relative contribution of model-free and model-based values into the final values.

The probability of selecting action a in state s will be determined according to the soft-max rule:

$$\pi(s,a) = \frac{e^{\beta(s)V(s,a)+\kappa(s,a)+d(s,a)}}{\sum_{a'} e^{\beta(s)V(s,a')+\kappa(s,a')+d(s,a')}} \tag{5.9}$$

Were parameters are same as the ones we described in the 'Model-based (MB)' section.

In the most general form, all the free parameters are included in the model: β_1, β_2, η, α_1, λ, w, k, ϕ (we assumed that $\alpha_2 = \eta$). We generated 32 simpler models by setting (1) $\lambda = 0$, (2) $\alpha_1 = \alpha_2$ (learning rate of model-free system is the same at stage 1 and stage 2 states), (3) $\beta_1 = \beta_2$ (rate of exploration is the same at stage 1 and stage 2 states), (4) $k = 0$ (there is no tendency to perseverate on the previously taken actions), and (5) $\phi = 0$ (there is not tendency

to take the discriminative action at stage 2).

By setting $w = 0$ the above hybrid model degenerated to a model-free system described in the previous section, and therefore, we generated 32 variants of model-free RL (similar to the hybrid model), system by setting $w = 0$.

5.5.5 Hierarchical model-based RL (HMB)

Implementation of the hierarchical structure is similar to hierarchical RL, with action sequences (*LL*, *LR*, etc) as options (section 2.4.5). We assumed actions *L*, *R*, *LL*, *LR*, *RL*, and *RR* are available in stats S_0, and actions *L* and *R*, are available in states S_1 and S_2. After reaching a terminal state (S_{Re} or S_{NR}), transition functions of both the action sequence, and the single action that led to that state update according to equation 5.1. In the case of single actions, the transition function will be updated by the $\eta = \eta_1$ update rate, and in the case of action sequences, the transition function will be updated by the $\eta = \eta_2$ update rate. Based on the learned transition function, value of action a in state s is calculated by the goal-directed system using equation 5.2.

After calculating values of actions ($V^G(s,a)$) using equation 5.2, the probability of selecting each action will be as follows:

$$\pi(s,a) = \frac{e^{\beta(s)V^G(s,a)+\kappa(s,a)+d(s,a)}}{\sum_{a'} e^{\beta(s)V^G(s,a')+\kappa(s,a')+d(s,a')}} \tag{5.10}$$

$\beta(s,a)$ is the rate of exploration. The rate of exploration for stage 2 states ($s \in \{S_1, S_2\}$) is $\beta(s,a) = \beta_2$. For stage 1 actions ($s = S_0$), if a is a single action, we assume $\beta(s,a) = \beta_1$, and if a is an action sequence $\beta(s,a) = \beta_3$. As before, $\kappa(s,a)$ captures action perseveration. We assumed that $\kappa(s,a) = k_1$ if action a is a single action, and $\kappa(s,a) = k_2$ if action a is an action sequence. Parameter $d(s,a)$ is similar to the previous sections.

For calculating the probability of selecting actions in the second stage, given the first choice of the subject, we need to know whether that action is a part of an action sequence selected

earlier, or is it under goal-directed control. Assume we know action L has been executed in state S_0 by the subject; then, the probability of this action being due to performing the LR action sequence is:

$$p(LR|S_0, L) = \frac{\pi(LR|S_0)}{\pi(L|S_0) + \pi(LR|S_0) + \pi(LL|S_0)} \tag{5.11}$$

Similarly, the probability of observing L due to selecting the single action L is:

$$p(L|S_0, L) = \frac{\pi(L|S_0)}{\pi(L|S_0) + \pi(LR|S_0) + \pi(LL|S_0)} \tag{5.12}$$

Based on this, the probability that the model assigns to action a in state $s \in \{S_1, S_2\}$, given that action a' is being observed in S_0 is:

$$p(a|s) = p(a'|S_0, a')\pi(a|s) + p(a'a|S_0, a') \tag{5.13}$$

Where $p(aa'|S_0, a')$ and $p(a'|S_0, a')$ are calculated using equations 5.11 and 5.12 respectively. Next, we assumed that even under the conditions in which an action sequence is being executed, there is a chance that the performance of the action sequence will be interrupted. Let's assume that the probability of interrupting an action sequence is I, then equation 5.13, will become as follows:

$$p(a|s) = \pi(a|s)\big(p(a'|S_0, a')(1 - I) + I\big) + (1 - I)p(a'a|S_0, a') \tag{5.14}$$

The above equation, in the case of $I = 0$, i.e., action sequences never become interrupted, will degenerate to equation 5.13. In the case of $I = 1$, i.e., all the action sequences are interrupted and they have no effect on stage 2 choices, and we have:

$$p(a|s) = \pi(a|s) \tag{5.15}$$

Which indicates probability of taking each action at stage 2 is guided only by the rewards

earned on that stage 2, and not by the action sequences in the first stage.

In the most general form, all the free parameters are included in the model: β_1, β_2, β_3, η_1, η_2, k_1, k_2, ϕ,I. We generated 256 simpler models by setting (1) $\beta_1 = \beta_2$ (exploration rates at stage 1 and stage 2 choices are the same), (2) $\beta_1 = \beta_3$ (exploration rates for action sequences and single actions are the same), (3) $\eta_1 = \eta_3$ (learning rates for action sequences and single actions are the same), (4) $k_1 = 0$ (no perseveration for single actions), (5) $k_2 = 0$ (no perseveration for action sequences), (6) $\phi = 0$ (no tendency to take discriminative actions), (7) $I = 1$ (action sequences are always interrupted), (8) $I = 0$ (action sequences are never interrupted).

5.5.6 Model comparison

We took a hierarchical Bayesian approach to compare different models. This approach provides a framework to compare models based on their complexity, and their fit to data. Bayesian model comparison is based on the model evidence quantity, which is the probability of the data given a model. The approach that we took to calculate this quantity is similar to the approach taken in (Piray et al., 2014), which is based on (Huys et al., 2011).

For each model, there are two sets of free parameters: group-level parameters denoted by Θ (we call these parameters hyper-parameters), and subject-level parameters denoted with θ_i for subject i. The hyper-parameters define the prior distribution over subject-level parameters. The aim is to calculate the probability of data (denoted by D) given model M:

$$P(D|M) = \int_{\Theta} P(D|M,\Theta)P(\Theta)d\Theta \tag{5.16}$$

Since the above integral is intractable, we approximate it using Bayesian Information Criterion (BIC) (Schwarz, 1978):

$$\log P(D|M) \approx \sum_i \log P(D_i|M,\Theta^{ML}) - \frac{1}{2}|\Theta|\log|D| \tag{5.17}$$

Where Θ^{ML} is the maximum-likelihood estimate of Θ, and D_i is the data of subject i. $|\Theta|$ is

the number of hyper-parameters, and $|D|$ is the sum of number of choices made by all the subjects. In the above formula, the term inside the sum is:

$$P(D_i|M,\Theta^{ML}) = \int_{\theta_i} P(D_i|M,\theta_i)P(\theta_i|\Theta^{ML})d\theta_i \tag{5.18}$$

Which is again intractable to compute, and we use Laplace method (MacKay, 2003) to approximate it:

$$\log P(D_i|M,\Theta^{ML}) \approx \log P(D_i|M,\theta_i^{MAP}) + \log P(\theta_i^{MAP}|\Theta^{ML}) + \frac{1}{2}|\theta_i|\log 2\pi - \frac{1}{2}\log|H_i| \tag{5.19}$$

Where θ_i^{MAP} is the maximum a posterior (MAP) estimate of θ_i. $|\theta_i|$ is the number of free parameters for model M, and $|H_i|$ is determinant of the Hessian matrix at θ_i^{MAP}.

Thus in summary, we calculated the model evidence for each subject using equation 5.19, and then we summed over all the model evidence for all the subjects to calculate equation 5.17, which is the model evidence over the whole group.

The only remaining question is how to calculate Θ^{ML}, which is:

$$\Theta^{ML} = \arg\max_{\Theta} \sum_i \log \int P(D_i|M,\theta_i)P(\theta_i|\Theta)d\theta_i \tag{5.20}$$

Similar to (Huys et al., 2011), we solved the above optimization problem using the expectation-maximization (EM) procedure (Dempster, Laird, & Rubin, 1977). This procedure starts with an initial value for the hyper-parameters Θ, using which the posterior distribution of each individual's parameter will be estimated using the Laplace approximation. These individual posterior distributions then shape a new value for the hyper-parameters (Θ), which will be used again to get new posterior distribution for each individual. This process continues until the hyper-parameters do not change anymore across iterations. Please refer to (Huys et al., 2011) for the details of the method.

The prior over all the individual level parameters (θ_i) where assumed to be a Gaussian distribu-

tion, and the mean and variance of the Gaussian were included in the hyper-parameters (thus the number of hyper-parameters for each model were twice as the number of free parameters of the model). Parameters that had a limited range (e.g., learning rates), were transformed to stratify the constrains.

We used the NLOPT software package (S. G. Johnson, n.d.) for nonlinear optimization using 'BOBYQA' algorithm. Finally, we used 'DerApproximator' package (Kroshko, n.d.) in order to estimate the Hessian at the MAP point.

5.5.7 Model comparison results

In total we tested 328 models (MB:n=8, MF:n=32, MB-MF:n=32, HMB:n=256). For the best four models in each family, Table 5.3 below represents the negative log-model evidence ($-\log P(D|M)$) (obtained from equation 5.17) for each model, the number of free parameters of each model (df), the free parameters of each model, and the family of each model. The table also represents a pseudo-r statistic (p–r^2), which is a normalized measure of the degree of variance accounted for in comparison to a model with random choices (averaged over subjects).

Out of 328 models that we tested, one of the models, which had eight free parameters ($\beta_1, \beta_3, \eta_1, \eta_2, k_1, k_2, \phi, I$), was not identifiable (the estimated hessian matrix was not a positive-definite matrix), and therefore it was excluded from the analysis.

Table 5.4 below represents estimated parameters for each individual in the best model (indicated by * in the table). The term '$-\log P(D_i|M, \Theta^{ML})$' represents the negative log-model evidence for each subject, obtained from equation 5.19.

Table 5.3 – For the best four models in each family, the table represents the negative log-model evidence ($-\log P(D|M)$) for each model, the number of free parameters of each model (df), the free parameters of each model, and the family of each model. The table also represents a pseudo-r statistic (p–r^2), which is a normalized measure of the degree of variance accounted for in comparison to a model with random choices (averaged over subjects).

model family	free-parameters	p-r^2	df	$-\log P(D\|M)$
HMB*	$\beta_1, \eta_1, k_2, \phi, I=0$	0.21	4	1172.09
HMB	$\beta_1, \eta_1, k_2, k_3, \phi, I=0$	0.21	5	1176.48
HMB	$\beta_1, \eta_1, k_1, k_2, \phi, I=0$	0.21	5	1178.90
HMB	$\beta_1, \eta_1, k_2, \phi, I$	0.21	5	1178.93
MF	$\beta_1, \beta_2, \alpha_1, \lambda$	0.17	4	1219.57
MF	$\beta_1, \beta_2, \alpha_1, \lambda, k$	0.18	5	1220.41
MB	$\beta_1, \beta_2, \eta, k$	0.18	4	1222.35
MF-MB	$\beta_1, \beta_2, \alpha_1, \lambda, w$	0.18	5	1223.02
MF-MB	$\beta_1, \beta_2, \alpha_1, \lambda, k, w$	0.19	6	1224.45
MF	$\beta_1, \alpha_1, \lambda$	0.16	3	1224.62
MF	$\beta_1, \alpha_1, \lambda, \phi$	0.17	4	1224.74
MF-MB	$\beta_1, \alpha_1, \lambda, w$	0.17	4	1224.82
MB	β_1, β_2, η	0.17	3	1225.02
MB	β_1, η, k, ϕ	0.17	4	1225.07
MF-MB	$\beta_1, \alpha_1, \lambda, w, \phi$	0.18	5	1225.89
MB	$\beta_1, \beta_2, \eta, \phi, k$	0.18	5	1226.82

Table 5.4 – Value of the estimated parameters for each subject.

subject	df	p-r^2	no. choices	$-\log P(D_i\|M, \Theta^{ML})$	β_1	η_1	ϕ	k_2
0	4	0.25	244	128.11	1.91	0.83	1.33	0.67
1	4	0.30	214	105.23	1.94	0.77	2.53	0.82
2	4	0.18	264	150.42	1.60	0.78	2.20	0.68
3	4	0.32	206	98.92	1.84	0.78	2.68	0.66
4	4	0.06	286	188.33	0.99	0.78	1.96	0.47
5	4	0.20	272	152.94	1.53	0.73	2.38	0.66
6	4	0.19	296	168.19	1.38	0.82	2.36	0.94
7	4	0.15	252	149.48	1.26	0.79	2.28	0.62

6 Conclusion

6.1 Summary of contributions

The current thesis investigated the theoretical and behavioral properties of automatic actions, and their relations to goal-directed actions. In chapter 2, we introduced different forms of decision-making processes in the brain, and motivated the fact that the hierarchical decision-making process can be regarded as the unifying principle underlying various forms of actions. Building on this assumption, in chapter 3, we provided a new normative computational model for learning action sequences as the building blocks of hierarchical decision-making, and we explored the power of the model in explaining different forms of automatic actions. In particular, using the proposed computational model, we elaborated the role that dopamine plays in learning action sequences, and in hierarchical decision-making. Then, we tested the predictions of the proposed model in chapter 4, and we provided experimental data in humans, that suggested the computational model can explain some behavioral phenomena that the previous accounts cannot explain. Finally, in chapter 5, we translated the task that was developed in chapter 4, to an experimental protocol in rats, and showed, that similar to the results of humans, rats also utilize hierarchical decision-making.

6.2 Future directions

In chapter 3, we proposed a normative model for the formation of action sequences, and their integration with model-based reinforcement learning. This normative model includes certain underlying computational signals, which can be investigated using model-based fMRI experiments. In chapter 4, we provided an experimental method for investigating hierarchical decision-making in humans, which can be used for studying the operation of this method in humans from different populations with psychiatric disorders. Similarly, the task developed in chapter 5 for rats, can be followed by investigating the lesion or inactivation of different brain regions, to uncover the role of each region of the brain in implementing the model.

On the theoretical side, it is likely that other reinforcement learning systems, such as model-free reinforcement learning, as suggested by previous works, operate concurrently with hierarchical model-based reinforcement learning. Investigating this possibility using outcome devaluation experiments, in conjunction with the task developed in chapter 5, can be an important step toward obtaining a more complete picture of the decision-making processes in the brain.

References

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

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

Adams, J. A. (1976). Issues for a Closed-Loop Theory of Motor Learning. In G. E. Stelmach (Ed.), *Motor control: Issues and trends* (pp. 87–107). San Diego, CA: Academic Press.

Astrom, K. J., & Murray, R. M. (2008). *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton, NJ: Princeton University Press.

Averbeck, B. B., Chafee, M. V., Crowe, D. a., & Georgopoulos, A. P. (2002, October). Parallel processing of serial movements in prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(20), 13172–7.

Averbeck, B. B., Chafee, M. V., Crowe, D. A., & Georgopoulos, A. P. (2003, May). Neural activity in prefrontal cortex during copying geometrical shapes. I. Single cells encode shape, sequence, and metric parameters. *Experimental brain research*, *150*(2), 127–41.

Averbeck, B. B., Sohn, J.-W., & Lee, D. (2006, February). Activity in prefrontal cortex during dynamic selection of action sequences. *Nature neuroscience*, *9*(2), 276–82.

Baayen, R. H. (2011). languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics". [Computer software manual]. Retrieved from `http://cran.r-project.org/package=languageR`

Badre, D., & Frank, M. J. (2012). Mechanisms of hierarchical reinforcement learning in

cortico-striatal circuits 2: evidence from fMRI. *Cerebral cortex*, *22*(3), 527–36.

Baird, L. C. (1993). *Advantage updating* (Tech. Rep.). Wright-Patterson Air Force Base Ohio: Wright Laboratory: Wright-Patterson Air Force Base Ohio: Wright Laboratory.

Balleine, B. W., Delgado, M. R., & Hikosaka, O. (2007). The role of the dorsal striatum in reward and decision-making. *Journal of Neuroscience*, *27*(31), 8161–5.

Balleine, B. W., Garner, C., Gonzalez, F., & Dickinson, A. (1995). Motivational control of heterogeneous instrumental chains. *Journal of Experimental Psychology: Animal Behavior Processes*, *21*(3), 203–217.

Balleine, B. W., & O'Doherty, J. P. (2010, January). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*(1), 48–69.

Bapi, R. S., & Doya, K. (2001). Multiple forward model architecture for sequence processing. In R. Sun & C. L. Giles (Eds.), *Sequence learning: paradigms, algorithms, and applications* (pp. 309–320). New York: Springer Verlag.

Barnes, T. D., Kubota, Y., Hu, D., Jin, D. Z., & Graybiel, A. M. (2005). Activity of striatal neurons reflects dynamic encoding and recoding of procedural memories. *Nature*, *437*(7062), 1158–61.

Barto, A. G. (1995). Adaptive Critics and the Basal Ganglia. In J. Houk, J. Davis, & B. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 215–232). Cambridge, MA: MIT Press.

Barto, A. G., & Mahadevan, S. (2003). Recent Advances in Hierarchical Reinforcement Learning. *Discrete Event Dynamic Systems*, *13*(4).

Bates, D., & Maechler, M. (2009). *lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-32*. Retrieved from `http://cran.r-project.org/web/packages/lme4/index.html`

Bates, D., & Maechler, M. (2014). *lme4: Linear mixed-effects models using S4 classes. R package version 1.1-7*. URL http://CRAN.R-project.org/package=lme4. Retrieved from `http://cran.r-project.org/web/packages/lme4/index.html`

Baxter, D., & Zamble, E. (1982). Reinforcer and response specificity in appetitive transfer of

control. *Animal Learning & Behavior*, *10*(2), 201–210.

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton, NJ: Princeton University Press.

Benecke, R., Rothwell, J. C., Dick, J. P., Day, B. L., & Marsden, C. D. (1987). Disturbance of sequential movements in patients with Parkinson's disease. *Brain*, *110 (Pt 2*, 361–379.

Berns, G. S., & Sejnowski, T. J. (1998, January). A computational model of how the basal ganglia produce sequences. *Journal of cognitive neuroscience*, *10*(1), 108–21.

Bissmarck, F., Nakahara, H., Doya, K., & Hikosaka, O. (2008). Combining modalities with different latencies for optimal motor control. *Journal of cognitive neuroscience*, *20*(11), 1966–79.

Blaisdell, A. P., Denniston, J. C., & Miller, R. R. (1997, May). Unblocking with Qualitative Change of Unconditioned Stimulus. *Learning and Motivation*, *28*(2), 268–279.

Bódi, N., Kéri, S., Nagy, H., Moustafa, A., Myers, C. E., Daw, N., … Gluck, M. a. (2009). Reward-learning and the novelty-seeking personality: A between-and within-subjects study of the effects of dopamine agonists on young parkinsons patients. *Brain*, *132*, 2385–2395.

Book, W. (1908). *The Psychologyof Skill*. Missoula, MT: Montana Press.

Botvinick, M. M. (2008). Hierarchical models of behavior and prefrontal function. *Trends in cognitive sciences*, *12*(5), 201–8.

Botvinick, M. M., Niv, Y., & Barto, A. G. (2009, December). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, *113*(3), 262–80.

Botvinick, M. M., & Plaut, D. C. (2004). Doing without schema hierarchies: a recurrent connectionist approach to normal and impaired routine sequential action. *Psychological review*, *111*(2), 395–429.

Botvinick, M. M., & Weinstein, A. (2014). Model-based hierarchical reinforcement learning and human action control. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *34*, 536–544.

Brafman, R. I., & Tennenholtz, M. (2003, March). R-max - a General Polynomial Time Algorithm for Near-optimal Reinforcement Learning. *Journal of Machine Learning Research*, *3*,

213–231.

Brown, T. L., & Carr, T. H. (1989). Automaticity in skill acquisition: Mechanisms for reducing interference in concurrent performance. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(4).

Bruce, D. (1994). Lashley and the Problem of Serial Order. *American Psychologist*, *49*(2), 93–103.

Brunskill, E., & Li, L. (2014). PAC-inspired Option Discovery in Lifelong Reinforcement Learning. In *International conference on machine learning (icml)*.

Buitrago, M. M., Ringer, T., Schulz, J. B., Dichgans, J., & Luft, A. R. (2004). Characterization of motor skill and instrumental learning time scales in a skilled reaching task in rat. *Behavioural Brain Research*, *155*(2), 249–256.

Buitrago, M. M., Schulz, J. B., Dichgans, J., & Luft, A. R. (2004). Short and long-term motor skill learning in an accelerated rotarod training paradigm. *Neurobiology of learning and memory*, *81*(3), 211–6.

Canic, M. J., & Franks, I. M. (1989, April). Response preparation and latency in patterns of tapping movements. *Human Movement Science*, *8*(2), 123–139.

Carr, H., & Watson, J. B. (1908). Orientation in the white rat. *Journal of Comparative Neurology and Psychology*, *18*(1).

Chang, Q., & Gold, P. E. (2004). Inactivation of dorsolateral striatum impairs acquisition of response learning in cue-deficient, but not cue-available, conditions. *Behavioral neuroscience*, *118*(2), 383–8.

Clark, J. J., Hollon, N. G., & Phillips, P. E. M. (2012, December). Pavlovian valuation systems in learning and decision making. *Current opinion in neurobiology*, *22*(6), 1054–61.

Cleland, G. G., & Davey, G. C. L. (1982, August). The effects of satiation and reinforcer develuation on signal-centered behavior in the rat. *Learning and Motivation*, *13*(3), 343–360.

Colwill, R. M., & Motzkin, D. (1994). Encoding of the unconditioned stimulus in Pavlovian conditioning. *Animal Learning & Behavior*, *22*(4), 384–394.

Colwill, R. M., & Rescorla, R. A. (1985). *Postconditioning devaluation of a reinforcer affects*

instrumental responding. (Vol. 11) (No. 1).

Cools, R., Altamirano, L., & D'Esposito, M. (2006). Reversal learning in Parkinson's disease depends on medication status and outcome valence. *Neuropsychologia*, *44*, 1663–1673.

Cooper, R. P., & Shallice, T. (2000, June). Contention scheduling and the control of routine activities. *Cognitive neuropsychology*, *17*(4), 297–338.

Cooper, R. P., & Shallice, T. (2006). Hierarchical schemas and goals in the control of sequential behavior. *Psychological Review*, *113*(4), 887–916; discussion 917–931.

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. *Journal of Neuroscience*, *21*(9), 3251–60.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006, July). Bayesian theories of conditioning in a changing world. *Trends in cognitive sciences*, *10*(7), 294–300.

Coutureau, E., & Killcross, S. (2003). Inactivation of the infralimbic prefrontal cortex reinstates goal-directed responding in overtrained rats. *Behavioural brain research*, *146*(1-2), 167–74.

Davey, G. C. L., & McKenna, I. (1983). The effect of postconditioning revaluation of CSI and UCS following Pavlovian second-order electrodermal conditioning in humans. *The Quarterly Journal of Experimental Psychology*, *35B*, 125–133.

Daw, N. D. (2002). Opponent interactions between serotonin and dopamine. *Neural networks*, *15*(4-6), 603–616.

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In M. R. Delgado, E. A. Phelps, & T. W. Robbins (Eds.), *Decision making, affect, and learning*. Oxford University Press.

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), 1204–15.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, *8*(12), 1704–11.

Daw, N. D., & Tourtezky, D. S. (2000, June). Behavioral considerations suggest an average reward TD model of the dopamine system. *Neurocomputing*, *32-33*(1-4), 679–684.

Dayan, P., & Balleine, B. W. (2002). Reward, motivation, and reinforcement learning. *Neuron*, *36*(2), 285–98.

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

Denny, M., Wells, R. H., & Maatsch, J. L. (1957). Resistance to extinction as a function of the discrimination habit established during fixed-ratio reinforcement. *Journal of Experimental Psychology*, *54*(6), 451–456.

Derusso, A. L., Fan, D., Gupta, J., Shelest, O., Costa, R. M., & Yin, H. H. (2010, January). Instrumental uncertainty as a determinant of behavior under interval schedules of reinforcement. *Frontiers in integrative neuroscience*, *4*(May), 1–8.

Desmurget, M., & Turner, R. S. (2010). Motor sequences and the basal ganglia: kinematics, not habits. *Journal of Neuroscience*, *30*(22), 7685–90.

Devan, B. D., & White, N. M. (1999). Parallel information processing in the dorsal striatum: relation to hippocampal function. *Journal of Neuroscience*, *19*(7), 2789–98.

Dezfouli, A., & Balleine, B. W. (2012). Habits, action sequences and reinforcement learning. *European Journal of Neuroscience*, *35*(7), 1036–1051.

Dezfouli, A., & Balleine, B. W. (2013). Actions, Action Sequences and Habits: Evidence that Goal-Directed and Habitual Action Control are Hierarchically Organized. *PLoS Computational Biology*, *9*(12).

Dezfouli, A., Lingawi, N. W., & Balleine, B. W. (2014). Habits as action sequences: hierarchical action control and changes in outcome value. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *369*(1655), 20130482–.

Dickinson, A. (1977). Appetitive-aversive interactions: Superconditioning of fear by an appetitive CS. *Quarterly Journal of Experimental Psychology*, *29*, 71–83.

Dickinson, A. (1994). Instrumental conditioning. In N. J. Mackintosh (Ed.), *Animal cognition and learning* (pp. 4–79). London: Academic Press.

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

Dickinson, A., & Balleine, B. W. (2002). The Role of Learning in the Operation of Motivational

Systems. In *Stevens' handbook of experimental psychology*. John Wiley & Sons, Inc.

Dickinson, A., Balleine, B. W., Watt, A., Gonzalez, F., & Boakes, R. A. (1995). Motivational control after extended instrumental training. *Animal Learning & Behavior*, *23*(2), 197–206.

Dickinson, A., Nicholas, D., & Adams, C. D. (1983). The effect of the instrumental training contingency on susceptibility to reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B Comparative and Physiological Psychology*, *35*(1), 35–51.

Dickinson, A., Squire, S., Varga, Z., & Smith, J. (1998). Omission learning after instrumental pretraining. *The Quarterly Journal of Experimental Psychology B: Comparative and Physiological Psychology*, *51(B)*(3), 271–286.

Dietterich, T. G. (2000, August). Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research*, *13*(1), 227–303.

Diuk, C., Tsai, K., Wallis, J., Botvinick, M. M., & Niv, Y. (2013, March). Hierarchical Learning Induces Two Simultaneous, But Separable, Prediction Errors in Human Basal Ganglia. *Journal of Neuroscience*, *33*(13), 5797–5805.

Dolan, R. J., & Dayan, P. (2013, October). Goals and habits in the brain. *Neuron*, *80*(2), 312–25.

Domingos, A. I., Vaynshteyn, J., Voss, H. U., Ren, X., Gradinaru, V., Zang, F., ... Friedman, J. (2011). Leptin regulates the reward value of nutrient. *Nature Neuroscience*, *14*(12), 1562–1568.

Doya, K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*, *12*(7-8), 961–974.

Endress, A. D., & Wood, J. (2011). From movements to actions: Two mechanisms for learning action sequences. *Cognitive psychology*, *63*(3), 141–171.

Estes, W. (1972). An associative basis for coding and organization in memory. In A. Melton & E. Martin (Eds.), *Coding processes in human memory* (pp. 161–190). Washington, DC: V.H. Winston & Sons.

Faure, A., Haberland, U., Condé, F., & El Massioui, N. (2005, March). Lesion to the nigrostriatal dopamine system disrupts stimulus-response habit formation. *Journal of Neuroscience*, *25*(11), 2771–80.

Frank, M. J., & Badre, D. (2012). Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: computational analysis. *Cerebral cortex (New York, N.Y. : 1991)*, *22*(3), 509–26.

Frank, M. J., Seeberger, L. C., & O'reilly, R. C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science*, *306*(December), 1940–1943.

Gallagher, M., McMahan, R. W., & Schoenbaum, G. (1999). Orbitofrontal cortex and representation of incentive value in associative learning. *Journal of Neuroscience*, *19*(15), 6610–6614.

Ganesan, R., & Pearce, J. M. (1988). Effect of changing the unconditioned stimulus on appetitive blocking. *Journal of experimental psychology. Animal behavior processes*, *14*, 280–291.

Garcia, J., Kimeldorf, D. J., & Koelling, R. A. (1955). Conditioned aversion to saccharin resulting from exposure to gamma radiation. *Science*, *122*(3160), 157–158.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological review*, *117*(1), 197–209.

Gershman, S. J., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & Behavior*, *40*(3), 255–268.

Gläscher, J., Daw, N. D., Dayan, P., & O'Doherty, J. P. (2010, May). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-Based and Model-Free Reinforcement Learning. *Neuron*, *66*(4), 585–595.

Glencross, D. J. (1977). Control of Skilled Movements. *Psychological Bulletin*, *84*(1), 14–29.

Graybiel, A. M. (1998). The basal ganglia and chunking of action repertoires. *Neurobiology of learning and memory*, *70*(1-2), 119–36.

Graybiel, A. M. (2008, January). Habits, rituals, and the evaluative brain. *Annual review of neuroscience*, *31*, 359–87.

Gremel, C. M., & Costa, R. M. (2013, January). Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. *Nature communications*, *4*, 2264.

Grossberg, S., & Pearson, L. R. (2008). Laminar cortical dynamics of cognitive and motor

working memory, sequence learning and performance: toward a unified theory of how the cerebral cortex works. *Psychological review*, *115*(3), 677–732.

Guthrie, E. R. (1935). *The Psychology of Learning*. New York: Harpers.

Guthrie, E. R. (1952). *The psychology of learning*. New York: Harper.

Hansen, E. A., Barto, A. G., & Zilberstein, S. (1996). Reinforcement Learning for Mixed Open-Loop and Closed Loop Control. In M. Mozer, M. I. Jordan, & T. Petsche (Eds.), *Advances in neural information processing systems (nips)* (Vol. 9, pp. 1026–1032). Denver, CO, USA: MIT Press.

Haruno, M., & Kawato, M. (2006, October). Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. *Neural networks*, *19*(8), 1242–54.

He, R., & Brunskill, E. (2011). Efficient Planning under Uncertainty with Macro-actions. *Journal of Artificial Intelligence Research*, *40*, 523–570.

Helie, S., Roeder, J. L., Vucovich, L., Rünger, D., & Ashby, F. G. (n.d.). A neurocomputational model of automatic sequence production. *Journal of Cognitive Neuroscience*.

Henry, F. M., & Rogers, D. E. (1960). Increased Response Latency for Complicated Movements and A "Memory Drum" Theory of Neuromotor Reaction. *Research Quarterly. American Association for Health, Physical Education and Recreation*, *31*(3), 448–458.

Hikosaka, O., Rand, M. K., Miyachi, S., & Miyashita, K. (1995). Learning of sequential movements in the monkey: process of learning and retention of memory. *Journal of neurophysiology*, *74*(4), 1652–61.

Holland, P., & Rescorla, R. A. (1975). *The effect of two ways of devaluing the unconditioned stimulus after first- and second-order appetitive conditioning*. (Vol. 1) (No. 4). (C) 1975 by the American Psychological Association: Yale U.

Holland, P. C. (1984, October). Unblocking in Pavlovian appetitive conditioning. *Journal of experimental psychology. Animal behavior processes*, *10*(4), 476–97.

Holland, P. C. (1988). Excitation and inhibition in unblocking. *Journal of experimental psychology. Animal behavior processes*, *14*, 261–279.

Holland, P. C. (2004, April). Relations between Pavlovian-instrumental transfer and reinforcer

devaluation. *Journal of Experimental Psychology: Animal Behavior Processes*, *30*(2), 104–17.

Holroyd, C. B., & Yeung, N. (2012). Motivation of extended behaviors by anterior cingulate cortex. *Trends in Cognitive Sciences*, *16*(2), 121–127.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal Of Computational And Graphical Statistics*, *15*(3), 651–674.

Houk, J., Adams, J., & Barto, A. (1994, November). A model of how the basal ganglia generate and use neural signals that predict reinforcement. In *Models of information processing in the basal ganglia (computational neuroscience)*. The MIT Press.

Hull, C. L. (1943). *Principles of Behavior*. New York: Appleton.

Hull, C. L. (1952). *A behavior system*. New Haven, CT: Yale University Press.

Hulstijn, W., & van Galen, G. P. (1983, October). Programming in handwriting: Reaction time and movement time as a function of sequence length. *Acta Psychologica*, *54*(1–3), 23–49.

Huys, Q. J. M., Cools, R., Gölzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011, April). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLoS computational biology*, *7*(4), e1002028.

Huys, Q. J. M., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012, January). Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, *8*(3), e1002410.

Iba, G. a. (1989, March). A heuristic approach to the discovery of macro-operators. *Machine Learning*, *3*(4), 285–317.

Ito, M., & Doya, K. (2009). Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *Journal of Neuroscience*, *29*(31), 9861–74.

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.

Izquierdo, A., Suda, R. K., & Murray, E. A. (2004, August). Bilateral orbital prefrontal cortex lesions in rhesus monkeys disrupt choices guided by both reward value and reward contingency. *Journal of Neuroscience*, *24*(34), 7540–8.

James, W. (1890). *The Principles of Psychology, Vol 1*. New York : Holt.

Jin, X., & Costa, R. M. (2010, July). Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature*, *466*(7305), 457–62.

Jin, X., Tecuapetla, F., & Costa, R. M. (2014, January). Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences. *Nature neuroscience*.

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

Jog, M. S., Kubota, Y., Connolly, C. I., Hillegaart, V., & Graybiel, A. M. (1999, November). Building neural representations of habits. *Science*, *286*(5445), 1745–9.

Johnson, A., van der Meer, M. A. A., & Redish, A. D. (2007, December). Integrating hippocampus and striatum in decision-making. *Current opinion in neurobiology*, *17*(6), 692–7.

Johnson, S. G. (n.d.). *The NLopt nonlinear-optimization package*.

Jong, N. K., Hester, T., & Stone, P. (2008). The Utility of Temporal Abstraction in Reinforcement Learning. In *Proceedings of the 7th international joint conference on autonomous agents and multiagent systems - volume 1* (pp. 299–306). Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Jueptner, M., Frith, C. D., Brooks, D. J., Frackowiak, R. S., & Passingham, R. E. (1997). Anatomy of motor learning. II. Subcortical structures and learning by trial and error. *Journal of neurophysiology*, *77*, 1325–1337.

Kakade, S. M. (2003). *On the Sample Complexity of Reinforcement Learning* (Unpublished doctoral dissertation). UCL.

Kamin, L. J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior* (pp. 279–296). New York: Appleton-Century-Crofts.

Keele, S. W. (1968). Movement control in skilled motor performance. *Psychological Bulletin*, *70*(6, Pt.1), 387–403.

Keele, S. W., Ivry, R., Mayr, U., Hazeltine, E., & Heuer, H. (2003, April). The cognitive and neural architecture of sequence representation. *Psychological review*, *110*(2), 316–39.

Keramati, M. M., Dezfouli, A., & Piray, P. (2011, May). Speed/Accuracy Trade-Off between the

Habitual and the Goal-Directed Processes. *PLoS computational biology*, *7*(5), e1002055.

Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral cortex (New York, N.Y. : 1991)*, *13*(4), 400–8.

Kim, H., Sul, J. H., Huh, N., Lee, D., & Jung, M. W. (2009). Role of striatum in updating values of chosen actions. *Journal of Neuroscience*, *29*(47), 14701–14712.

Kim, K. M., Baratta, M. V., Yang, A., Lee, D., Boyden, E. S., & Fiorillo, C. D. (2012). Optogenetic mimicry of the transient activation of dopamine neurons by natural reward is sufficient for operant reinforcement. *PLoS ONE*, *7*(4), 1–8.

Klapp, S. T. (1977, January). Reaction time analysis of programmed control. *Exercise and sport sciences reviews*, *5*, 231–53.

Klapp, S. T. (1995). Motor response programming during simple choice reaction time: The role of practice. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(5), 1015–1027.

Kolter, J. Z., Plagemann, C., Jackson, D. T., Ng, A. Y., & Thrun, S. (2010). A Probabilistic Approach to Mixed Open-loop and Closed-loop Control , with Application to Extreme Autonomous Driving. In *International conference on robotics and automation (icra)* (pp. 839–845). Anchorage, AK: IEEE.

Korf, R. E. (1985). Macro-Operators: A Weak Method for Learning. *Artificial Intelligence*, *26*(1), 35–77.

Kosaki, Y., & Dickinson, A. (2010, July). Choice and contingency in the development of behavioral autonomy during instrumental conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(3), 334–42.

Kroshko, D. (n.d.). *OpenOpt: Free scientific-engineering software for mathematical modeling and optimization.* Retrieved from `http://www.openopt.org/`

Kruse, J. M., Overmier, J., Konz, W. A., & Rokke, E. (1983). Pavlovian conditioned stimulus effects upon instrumental choice behavior are reinforcer specific. *Learning and Motivation*, *14*(2), 165–181.

Kubota, Y., Liu, J., Hu, D., DeCoteau, W. E., Eden, U. T., Smith, A. C., & Graybiel, A. M. (2009, October). Stable encoding of task structure coexists with flexible coding of task events

in sensorimotor striatum. *Journal of neurophysiology*, *102*(4), 2142–60.

Lashley, K. S. (1917). The accuracy of movement in the absence of excitation from the moving organ. *American Journal of Physiology*, *43*, 169–194.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–136). New York: Wiley.

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the experimental analysis of behavior*, *84*(3), 555–79.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, *81*(3), 687–99.

Lehéricy, S., Benali, H., Van de Moortele, P.-F., Pélégrini-Issac, M., Waechter, T., Ugurbil, K., & Doyon, J. (2005, August). 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), 12566–71.

Levesque, M., Bedard, M. A., Courtemanche, R., Tremblay, P. L., Scherzer, P., & Blanchet, P. J. (2007). Raclopride-induced motor consolidation impairment in primates: role of the dopamine type-2 receptor in movement chunking into integrated sequences. *Experimental brain research. Experimentelle Hirnforschung. Experimentation cerebrale*, *182*(4), 499–508.

Li, B. Y. L. (2009). *A Unifying Framework for Computational Reinforcement Learning Theory* (Unpublished doctoral dissertation). Rutgers University.

Lingawi, N. W., & Balleine, B. W. (2012, January). Amygdala central nucleus interacts with dorsolateral striatum to regulate the acquisition of habits. *Journal of Neuroscience*, *32*(3), 1073–81.

Littman, M. L. (1996). *Algorithms for Sequential Decision Making* (Unpublished doctoral dissertation). Brown University.

MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

Mahadevan, S. (1996). Average Reward Reinforcement Learning: Foundations , Algorithms , and Empirical Results. *Machine Learning*, *22*(1), 159–195.

Mann, T. a., & Mannor, S. (2014). Scaling Up Approximate Value Iteration with Options : Better Policies with Fewer Iterations. In *Proceedings of the 31ˆst international conference on machine learning*.

Matsumoto, N., Hanakawa, T., Maki, S., Graybiel, A. M., & Kimura, M. (1999). Nigrostriatal dopamine system in learning to perform sequential motor tasks in a predictive manner. *Journal of neurophysiology*, *82*(2), 978–98.

Matsuzaka, Y., Picard, N., & Strick, P. L. (2007). Skill representation in the primary motor cortex after long-term practice. *Journal of neurophysiology*, *97*(2), 1819–32.

McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y., & Schoenbaum, G. (2011, February). Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *Journal of Neuroscience*, *31*(7), 2700–5.

Mcgovern, E. A. (2002). *Autonomous Discovery Of Temporal Abstractions From Interaction With An Environment* (PhD thesis). University of Massachusetts - Amherst.

Miller, G., Galanter, E., & Pribram, K. (1960). *Plans and the structure of behavior*. New York: Holt, Rinehart & Winston.

Miyachi, S., Hikosaka, O., & Lu, X. (2002, September). Differential activation of monkey striatal neurons in the early and late stages of procedural learning. *Experimental brain research*, *146*(1), 122–6.

Miyachi, S., Hikosaka, O., Miyashita, K., Kárádi, Z., & Rand, M. K. (1997, June). Differential roles of monkey striatum in learning of sequential hand movement. *Experimental brain research*, *115*(1), 1–5.

Miyashita, K., Rand, M. K., Miyachi, S., & Hikosaka, O. (1996). Anticipatory saccades in sequential procedural learning in monkeys. *Journal of neurophysiology*, *76*(2), 1361–6.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996, March). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*(5), 1936–47.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature neuroscience*, *9*, 1057–1063.

Moussa, R., Poucet, B., Amalric, M., & Sargolini, F. (2011). Contributions of dorsal striatal

subregions to spatial alternation behavior. *Learning & memory*, *18*(7), 444–51.

Mowrer, O. H., & Jones, H. (1945). Habit strength as a function of the pattern of reinforcement. *ournal of Experimental Psychology*, *35*(4), 293–311.

Mushiake, H., Saito, N., Sakamoto, K., Itoyama, Y., & Tanji, J. (2006, May). Activity in the lateral prefrontal cortex reflects multiple steps of future events in action plans. *Neuron*, *50*(4), 631–41.

Nakahara, H., Doya, K., & Hikosaka, O. (2001). Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences - a computational approach. *Journal of cognitive neuroscience*, *13*(5), 626–47.

Neuringer, A. (2004). Reinforced variability in animals and people: implications for adaptive action. *The American psychologist*, *59*(9), 891–906.

Neuringer, A., & Jensen, G. (2010). Operant variability and voluntary action. *Psychological review*, *117*(3), 972–93.

Newell, A., & Simon, H. (1963). GPS, a program that simulates human thought. In E. Feigenbaum & J. Feldman (Eds.), *Computers and thought* (pp. 279–293). New York: McGraw-Hill.

Nissen, M. J., & Bullemer, P. (1987). Attentional Requirements of Learning : Performance Measures Evidence from. *Cognitive Psychology*, *19*(1), 1–32.

Niv, Y. (2007). *The effects of motivation on habitual instrumental behavior* (Unpublished doctoral dissertation). The Hebrew University of Jerusalem.

Niv, Y., & Schoenbaum, G. (2008, July). Dialogues on prediction errors. *Trends in cognitive sciences*, *12*(7), 265–72.

Ostlund, S. B., Winterbauer, N. E., & Balleine, B. W. (2009). Evidence of action sequence chunking in goal-directed instrumental conditioning and its dependence on the dorsomedial prefrontal cortex. *Journal of Neuroscience*, *29*(25), 8280–7.

Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological science*, *24*(5), 751–61.

Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A., & Daw, N. D. (2013). Working-memory capacity

protects model-based learning from stress. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(52), 20941–6. doi: 10.1073/pnas.1312011110

Overmann, S. R., & Denny, M. (1974). The free-operant partial reinforcement effect: A discrimination analysis. *Learning and Motivation*, *5*(2), 248–257.

Packard, M. G., & McGaugh, J. L. (1996, January). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of learning and memory*, *65*(1), 65–72.

Parr, R., & Russell, S. (1998). *Reinforcement learning with hierarchies of machines* (Unpublished doctoral dissertation). University of California at Berkeley.

Pavlov, I. P. (1927). *Conditioned Reflexes* (Vol. 17). Oxford University Press.

Penny, W. D., Stephan, K. E., Daunizeau, J., Rosa, M. J., Friston, K. J., Schofield, T. M., & Leff, A. P. (2010). Comparing families of dynamic causal models. *PLoS computational biology*, *6*(3), e1000709.

Pew, R. W. (1966). Acquisition of hierarchical control over the temporal organization of a skill. *Journal of Experimental Psychology*, *71*(5), 764–771.

Pezzulo, G., Rigoli, F., & Chersi, F. (2013, January). The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. *Frontiers in psychology*, *4*, 92.

Piray, P., Zeighami, Y., Bahrami, F., Eissa, A. M., Hewedi, D. H., & Moustafa, A. A. (2014). Impulse control disorders in Parkinson's disease are associated with dysfunction in stimulus valuation but not action valuation. *The Journal of neuroscience*, *34*(23), 7814–24.

Platt, J. R., & Day, R. B. (1979). A hierarchical response-unit analysis of resistance to extinction following fixed-number and fixed-consecutive-number reinforcement. *Journal of Experimental Psychology: Animal Behavior Processes*, *5*(4), 307–320.

Puterman, M. L. (1994). *Markov Decision Processes*. New York: Wiley Interscience.

Quinn, J. J., Pittenger, C., Lee, A. S., Pierson, J. L., & Taylor, J. R. (2013). Striatum-dependent habits are insensitive to both increases and decreases in reinforcer value in mice. *The European journal of neuroscience*, *37*(6), 1012–21.

R Core Team. (2012). R: A Language and Environment for Statistical Computing [Computer

software manual]. Vienna, Austria. Retrieved from `http://www.r-project.org/`

R Core Team. (2014). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.r-project.org/`

Randlø v, J. (1998). Learning Macro-Actions in Reinforcement Learning. In M. J. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems (nips)* (pp. 1045–1051). Denver, Colorado, USA: MIT Press.

Rescorla, R. A. (1973). Second-order conditioning: Implications for theories of learning. In F. J. McGuigan & D. B. Lumsden (Eds.), *Contemporary approaches to conditioning and learning.* Oxford, England: V. H. Winston & Sons.

Rescorla, R. A. (1999). Learning about qualitatively different outcomes during a blocking procedure. *Animal Learning & Behavior*, *27*(2), 140–151.

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

Restle, F. (1957). Discrimination of cues in mazes: a resolution of the place-vs.-response question. *Psychological review*, *64*(4), 217–28.

Reynolds, J. R., & O'Reilly, R. C. (2009). Developing PFC representations using reinforcement learning. *Cognition*, *113*(3), 281–292.

Rhodes, B. J., Bullock, D., Verwey, W. B., Averbeck, B. B., & Page, M. P. a. (2004, November). Learning and production of movement sequences: behavioral, neurophysiological, and modeling perspectives. *Human movement science*, *23*(5), 699–746.

Ribas-Fernandes, J. J. F., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., & Botvinick, M. M. (2011). A neural signature of hierarchical reinforcement learning. *Neuron*, *71*(2), 370–9.

Ritchie, B. F., Aeschliman, B., & Pierce, P. (1950). Studies in spatial learning: VIII. Place performance and the acquisition of place dispositions. *Journal of Comparative and Physiological Psychology*, *43*, 73–85.

Roesch, M. R., Calu, D. J., & Schoenbaum, G. (2007). Dopamine neurons encode the better op-

tion in rats deciding between differently delayed or sized rewards. *Nature neuroscience*, *10*, 1615–1624.

Rondi-Reig, L., Petit, G. H., Tobin, C., Tonegawa, S., Mariani, J., & Berthoz, A. (2006). Impaired sequential egocentric and allocentric memories in forebrain-specific-NMDA receptor knock-out mice during a new task dissociating strategies of navigation. *Journal of Neuroscience*, *26*(15), 4071–81.

Rosenbaum, D. A. (2009). *Human Motor Control*. Elsevier Science.

Rosenbaum, D. A., Cohen, R. G., Jax, S. a., Weiss, D. J., & van der Wel, R. (2007, August). The problem of serial order in behavior: Lashley's legacy. *Human movement science*, *26*(4), 525–54.

Rumelhart, D. E., & Norman, D. A. (1982). Simulating a Skilled Typist: A Study of Skilled Cognitive-Motor Performance. *Cognitive Science*, *6*(1), 1–36.

Rutledge, R. B., Lazzaro, S. C., Lau, B., Myers, C. E., Gluck, M. a., & Glimcher, P. W. (2009). Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. *Journal of Neuroscience*, *29*(48), 15104–15114.

Saga, Y., Iba, M., Tanji, J., & Hoshi, E. (2011, July). Development of multidimensional representations of task phases in the lateral prefrontal cortex. *Journal of Neuroscience*, *31*(29), 10648–65.

Schmitzer-Torbert, N., & Redish, A. D. (2004, May). Neuronal activity in the rodent dorsal striatum in sequential navigation: separation of spatial and reward responses on the multiple T task. *Journal of neurophysiology*, *91*(5), 2259–72.

Schneider, D. W., & Logan, G. D. (2006). Hierarchical control of cognitive processes: switching tasks in sequences. *Journal of experimental psychology General*, *135*(4), 623–640.

Schultz, W. (2000). Multiple reward signals in the brain. *Nature reviews. Neuroscience*, *1*(3), 199–207.

Schultz, W., Dayan, P., & Montague, P. R. (1997, March). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593–9.

Schwarting, R. K. W. (2009). Rodent models of serial reaction time tasks and their implementation in neurobiological research. *Behavioural brain research*, *199*(1), 76–88.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 461–464.

Shah, A., & Barto, A. G. (2009). Effect on movement selection of an evolving sensory representation: a multiple controller model of skill acquisition. *Brain research*, *1299*, 55–73.

Shima, K., Isoda, M., Mushiake, H., & Tanji, J. (2007, January). Categorization of behavioural sequences in the prefrontal cortex. *Nature*, *445*(7125), 315–8.

Shull, R. L., Gaynor, S. T., & Grimes, J. a. (2002, May). Response rate viewed as engagement bouts: resistance to extinction. *Journal of the experimental analysis of behavior*, *77*(3), 211–31.

Shull, R. L., & Grimes, J. a. (2003, September). Bouts of responding from variable-interval reinforcement of lever pressing by rats. *Journal of the experimental analysis of behavior*, *80*(2), 159–71.

Skinner, B. F. (1938). *The behavior of organisms*. New York: Appleton-Century-Crofts.

Smith, K. S., & Graybiel, A. (2013). A dual operator view of habitual behavior reflecting cortical and striatal dynamics. *Neuron*, *79*(2), 361–374.

Smith, K. S., & Graybiel, A. M. (2014). Investigating habits: strategies, technologies and models. *Frontiers in Behavioral Neuroscience*, *8*(February), 1–17.

Solway, A., & Botvinick, M. M. (2012). Goal-directed decision making as probabilistic inference: A computational framework and potential neural correlates. *Psychological Review*, *119*, 120–154.

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y., & Botvinick, M. M. (2014). Optimal Behavioral Hierarchy. *PLoS computational biology*, *10*(8), e1003779.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–17.

Sternberg, S., Monsell, S., Knoll, R. L., & Wright, C. E. (1978). The Latency and Duration of Rapid Movement Sequences: Comparisons of Speech and Typewriting. In G. E. B. T. I. P. i. M. C. Stelmach & Learning (Eds.), *Information processing in motor control and learning* (pp. 117–152). Academic Press.

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004, June). Matching behavior and the representation of value in the parietal cortex. *Science (New York, N.Y.)*, *304*(5678), 1782–

7.

Sutton, R. S. (1988). Learning to Predict by the Methods of Temporal Differences. *Machine Learning*, *vol*, 3pp9–44.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. Cambridge, MA: MIT Press.

Sutton, R. S., Precup, D., & Singh, S. (1999). Between MDPs and Semi-MDPs: A Framework for Temporal Abstraction in Reinforcement Learning. *Artificial Intelligence*, *112*(1-2), 181 – 211.

Tan, K. R., Yvon, C., Turiault, M., Mirzabekov, J. J., Doehner, J., Labouèbe, G., ... Lüscher, C. (2012). GABA Neurons of the VTA Drive Conditioned Place Aversion. *Neuron*, *73*, 1173–1183.

Tang, C., Pawlak, A. P., Prokopenko, V., & West, M. O. (2007, February). Changes in activity of the striatum during formation of a motor habit. *The European journal of neuroscience*, *25*(4), 1212–27.

Tanji, J. (2001, January). Sequential organization of multiple movements: involvement of cortical motor areas. *Annual review of neuroscience*, *24*, 631–51.

Terrace, H. S. (1991, January). Chunking during serial learning by a pigeon: I. Basic evidence. *Journal of Experimental Psychology: Animal Behavior Processes*, *17*(1), 81–93.

Thorn, C. a., Atallah, H., Howe, M., & Graybiel, A. M. (2010, June). Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning. *Neuron*, *66*(5), 781–95.

Thorndike, E. L. (1911). *Animal Intelligence*. New York: Macmillan.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*(4), 189–208.

Tolman, E. C., Ritchie, B. F., & Kalish, D. (1946). Studies in spatial learning: II. Place learning versus response learning. *Journal of Experimental Psychology*, *36*, 221–229.

Trapold, M., & Overmier, J. (1972). The second learning process in instrumental learning. In A. Black & W. Prokasy (Eds.), *Classical conditioning ii: Current research and theory* (pp. 427–452). New York: Appleton-Century-Crofts.

Tremblay, P.-L., Bedard, M.-A., Langlois, D., Blanchet, P. J., Lemay, M., & Parent, M. (2010).

Movement chunking during sequence learning is a dopamine-dependant process: a study conducted in Parkinson's disease. *Experimental brain research. Experimentelle Hirnforschung. Experimentation cerebrale*, *205*(3), 375–385.

Tremblay, P.-L., Bedard, M.-A., Levesque, M., Chebli, M., Parent, M., Courtemanche, R., & Blanchet, P. J. (2009, March). Motor sequence learning in primate: role of the D2 receptor in movement chunking during consolidation. *Behavioural brain research*, *198*(1), 231–9.

Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009, June). A specific role for posterior dorsolateral striatum in human habit learning. *The European journal of neuroscience*, *29*(11), 2225–32.

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*(May), 1080–1084.

Tsitsiklis, J. N., & Roy, B. V. (1999). Average Cost Temporal-Difference Learning. *Automatica*, *35*, 1799–1808.

Turner, R. S., & Desmurget, M. (2010). Basal ganglia contributions to motor control: a vigorous tutor. *Current opinion in neurobiology*, *20*(6), 704–16.

Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007, April). Determining the neural substrates of goal-directed learning in the human brain. *Journal of Neuroscience*, *27*(15), 4019–26.

van Mier, H., & Hulstijn, W. (1993, December). The effects of motor complexity and practice on initiation time in writing and drawing. *Acta psychologica*, *84*(3), 231–51.

Verbruggen, F., & Logan, G. D. (2008). Response inhibition in the stop-signal paradigm. *Trends in Cognitive Sciences*, *12*(11), 418–424.

Verwey, W. B. (1994). Evidence for the development of concurrent processing in a sequential keypressing task. *Acta Psychologica*, *85*(3), 245–262.

Verwey, W. B. (1999). Evidence for a multistage model of practice in a sequential movement task. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(6), 1693–1708.

Wächter, A., & Biegler, L. T. (2005). On the implementation of an interior-point filter line-

search algorithm for large-scale nonlinear programming. *Mathematical Programming*, *106*(1), 25–57.

Wang, L. P., Li, F., Wang, D., Xie, K., Wang, D., Shen, X., & Tsien, J. Z. (2011, December). NMDA receptors in dopaminergic neurons are crucial for habit learning. *Neuron*, *72*(6), 1055–66.

Wassum, K. M., Cely, I. C., Maidment, N. T., & Balleine, B. W. (2009, October). Disruption of endogenous opioid activity during instrumental learning enhances habit acquisition. *Neuroscience*, *163*(3), 770–80.

Watkins, C. (1989). *Learning from Delayed Rewards* (Ph.D. thesis). Cambridge University.

White, K., & Davey, G. C. (1989, January). Sensory preconditioning and UCS inflation in human 'fear' conditioning. *Behaviour research and therapy*, *27*(2), 161–6.

Wiener, J. M., & Mallot, H. A. (2003, December). 'Fine-to-Coarse' Route Planning and Navigation in Regionalized Environments. *Spatial Cognition & Computation*, *3*(4), 331–358.

Willingham, D. B. (1998, July). A neuropsychological theory of motor skill learning. *Psychological review*, *105*(3), 558–84.

Willingham, D. B., Nissen, M. J., & Bullemer, P. (1989). On the development of procedural knowledge. *Journal of experimental psychology. Learning, memory, and cognition*, *15*(6), 1047–60.

Wunderlich, K., Smittenaar, P., & Dolan, R. J. (2012, August). Dopamine enhances model-based over model-free choice behavior. *Neuron*, *75*(3), 418–24.

Wymbs, N. F., Bassett, D. S., Mucha, P. J., Porter, M. A., & Grafton, S. T. (2012). Differential Recruitment of the Sensorimotor Putamen and Frontoparietal Cortex during Motor Chunking in Humans. *Neuron*, *74*(5), 936–946.

Yin, H. H. (2010, November). The sensorimotor striatum is necessary for serial order learning. *Journal of Neuroscience*, *30*(44), 14719–23.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004, January). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *The European journal of neuroscience*, *19*(1), 181–9.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2005, July). Blockade of NMDA receptors in the

dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *The European journal of neuroscience*, *22*(2), 505–12.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2006, January). Inactivation of dorsolateral striatum enhances sensitivity to changes in the action-outcome contingency in instrumental conditioning. *Behavioural brain research*, *166*(2), 189–96.