國立臺灣大學理學院心理學研究所

碩士論文

Graduate Institute of Psychology
College of Science
National Taiwan University
Master Thesis

以小鼠探討紋狀體不同腦區在增強學習以及酬賞預測誤差中所扮演的角色

The Role of Striatal Subregions in Reinforcement Learning Process and Reward Prediction Error using Excitotoxic Lesion in Male Mice

劉雅文

Ya-Wen Liu

指導教授：賴文崧 博士

Advisor: Wen-Sung Lai, Ph.D.

中華民國 104 年 1 月

January, 2015

摘要

紋狀體分屬於基底核，是主要接收基底核訊息的腦區，更參與動作控制和酬賞相關的學習。近來的研究指出紋狀體與行動值以及酬賞預測誤訊號（個體預期得到的酬賞和實際得到的酬賞之差異)的更新有關。紋狀體可進一步分成三個分區，各分區分別與不同種類的學習歷程有關。背內側紋狀體主要接收來自關聯皮層的訊息、與目標導向的行為學習有關；背外側紋狀體主接收來自感覺動作皮層的訊息、與習慣學習有關；伏隔核則被認為是表徵對未來酬賞預期的重要腦區，並可根據此預期進一步影響酬賞導向的行為選擇。然而，紋狀體內各分區在增強學習以及酬賞相關的學習中所扮演的角色、及其內在機制仍未有一定論。所以，本研究的目的為檢視不同的紋狀體分區在增強學習、酬賞預測誤訊號更新所扮演的角色，使用興奮性毀壞藥物注射紋狀體不同分區搭配二選項動態酬賞作業，觀察毀壞後小鼠的學習行為是否改變。本研究使用的二選項動態酬賞作業包含兩組不同的酬賞機率學習，小鼠的每次選擇都會被記錄。我們使用增強學習模型來分析資料，酬賞預測誤的相關參數估計使用貝氏估計法，另使用配對法則分析小鼠的選擇行為傾向。本研究結果顯示，背內側紋狀體毀壞小鼠在整個學習過程裡，相較於控制組小鼠，除了達到預設標準需要更多的選擇次數外，也在學習過程中累積更多錯誤。背外側紋狀體以及伏隔核毀壞小鼠則沒有展現整體學習行為上的差異。另使用增強學習模型分析，發現背內側紋狀體以及伏隔核毀壞小鼠皆有酬賞預測誤訊號更新速度下降、行為選擇一致性些微上升的情況。配對法則分析部分，沒有發現任何毀壞組及控制組的組間差異。整體而言，本研究證實了背內側

紋狀體的功能損傷會影響酬賞相關學習和行為決策的表現。除此之外，亦證實背內側紋狀體以及伏隔核對於二選項動態酬賞作業的重要性，以及兩腦區皆在決策行為的價值評估、行為選擇兩部分扮演重要角色。

關鍵詞：增強學習、酬賞預測誤、紋狀體、興奮性毀壞、二選項動態酬賞作業、小鼠、決策行為

The Role of Striatal Subregions in Reinforcement Learning Process and Reward Prediction Error using Excitotoxic Lesion in Male Mice

Ya-Wen Liu

Abstract

The striatum is the principal input structure of the basal ganglia that influences motor control and reward-based learning. Emerging studies indicate that it also contributes to update of action value and reward prediction error (RPE), a discrepancy between the predicted and actual rewards. Previous studies imply that three different subregions of the striatum participating in different kinds of learning processes. The dorsomedial striatum (DMS, also known as “associative striatum” in primates) which receives inputs from the association cortices is implicated in goal-directed behavior in rodents. The dorsolateral striatum (DLS, a part of the sensorimotor striatum in primates) is related to habit learning in rodents. The nucleus accumbens (NA) is implicated in representing predicted future reward, and the representation can be used to guide action selection for reward. However, the precise role or mechanism of each subregion in reinforcement learning and reward-based decision making is still under debate. The aim of this study is to examine the role of different striatal subregions (including DMS, DLS, and NA) in reinforcement learning process and reward prediction error using excitotoxic lesions and 2-choice dynamic foraging task in male C57/Bl6 mice. The 2-choice dynamic foraging task is a risky-choices task which consisted of two kinds of reward ratio learning. The behavioral performance of each of the three lesioned groups and their sham controls were recorded. Their trial-by-trial choice behavior were further analyzed and fit with a standard reinforcement learning model using the Bayesian estimation approach and matching law analysis to elaborate parameters for RPE and reward sensitivity. Compared to sham controls, overall behavioral results indicated that the DMS lesioned mice had more trials to reach the preset criteria and made more cumulated errors during the learning process of this dynamic foraging task. In contrast to the DMS group, both NA and DLS lesioned groups did not exhibited more accumulated trials or more cumulated errors. Reinforcement learning model analysis further revealed that both DMS and NA lesion mice had a lower learning rate in updating the RPE signaling and a slightly higher perseveration compared to their sham controls. But no significant difference was found in the reward sensitivity among the 3 groups. Collectively, the current study confirmed the importance of DMS and NA in the 2-choice dynamic foraging task and their roles in the value component and choice component of decision making. Excitotoxic lesion of DMS can significantly impair performance of probabilistic reward-based learning and decision making.

Keywords: reinforcement learning, reward prediction error, striatum, excitotoxic lesion, 2-choice dynamic foraging task, mice, decision making

Table of Contents

Tables and Figures

Chapter 1: Introduction

1. An overview of decision making

In everyday life, there are numerous decisions waiting for us, from what food to eat, what clothes to wear, what hair style and what you are going to do in the future…etc. All of these things need us to make decisions. In short, a decision is a process that weighs priors, evidence, and values of different options to generate a choice intended to achieve particular goals. And this is the main focus of the field of decision making. Recently, a cross disciplinary approach to study decision making process has come out to the mainstream: Neuroeconomics.

Neuroeconomics is a newly established field that integrates the confluence of economics, psychology and neuroscience to the study of decision making to try and create a better model about decisions, interactions, and risks and rewards. Accordingly, neuroeconomics combines the modeling from economics with psychological studies of social and emotional influences on decision making, and utilizes tools from neuroscience that permit the observation of valuation and decision-making computations that take place in the brain. In the following section, a brief introduction of decision process and its corresponding brain areas are described.

1.1. Elements of a decision. As mentioned, a decision is a process that weighs priors, evidence, and values of different options to generate a choice intended to achieve particular goals. It also can be regarded as a form of statistical inference (Kersten, Mamassian, & Yuille, 2004; Smith, 1961). According to Doya, the process of value-based decision making can be decomposed into four steps (Doya, 2008):

a. Subject identifies the existing situation (or state).

b. Subject evaluates possible options (or actions) according to the reward or punishment every potential choice could bring.

c. Subject makes the final decision after considering own needs.

d. Based on the outcome, subject revaluates the decision.

Although decisions are not always made through these four steps, a standardizing procedure of decision making process is useful in the understanding of how these steps are executed in the brain.

1.2. Brain areas related to value functions. Subject's internal reward expectancy represents value functions in decision process. Theoretically, neural signals related to reward expectancy can be divided into two categories: action value and state value (Lee, Seo, & Jung, 2012). Action value functions are useful in choosing a particular action, especially if such signals are observed before the execution of a motor response. However, based on the dimension in which choices are made, brain areas related to the corresponding action value functions may vary substantially. In most previous studies, many brain areas are implicated in action value functions, including dorsolateral prefrontal cortex (Barraclough, Conroy, & Lee, 2004; Kim, Hwang, & Lee, 2008), posterior parietal cortex (Dorris & Glimcher, 2004; Platt & Glimcher, 1999; Sugrue, Corrado, & Newsome, 2004), medial frontal cortex (Seo & Lee, 2009; So & Stuphorn, 2010; Sul, Kim, Huh, Lee, & Jung, 2010), premotor cortex (Pastor-Bernier & Cisek, 2011), and striatum (Cai, Kim, & Lee, 2011; Kim, Sul, Huh, Lee, & Jung, 2009; Lau & Glimcher, 2008; Samejima, Ueda, Doya, & Kimura, 2005; Tai, Lee, Benavidez, Bonci, & Wilbrecht, 2012).

State value functions play a more evaluative role in the brain, and it can be further divided into two categories: pre-decision and post-decision. For the pre-decision state value functions, researchers found that some of the related brain areas overlapped with the action value functions. Neural activity in the posterior parietal cortex and dorsal striatum showed both characteristics of pre-decision state value functions and action value functions (Cai et al., 2011; Seo, Barraclough, & Lee, 2009; Yang & Shadlen, 2007). Brain areas related to pre-decision state value functions are also found in the ventral striatum (Cai et al., 2011), anterior cingulate cortex (Seo & Lee, 2007), and amygdala (Belova, Paton, & Salzman, 2008).

Post-decision state value functions are also called chosen values, and its related brain areas are also widespread, including orbitofrontal cortex (Padoa-Schioppa & Assad, 2006; Sul et al., 2010), medial frontal cortex (Sul et al., 2010), ventromedial prefrontal cortex (Hare, Camerer, & Rangel, 2009), dorsolateral prefrontal cortex (Hare et al., 2009), and striatum (Cai et al., 2011; Kim et al., 2009; Lau & Glimcher, 2008). Since the revaluation happens after subjects made their decision, the chosen value may be utilized to revaluate (i.e. compute the difference between the outcome of a choice and the chosen value) and update value functions.

1.3. Brain areas related to action selection. In decision making process, the action value must be transformed into specific action and corresponding motor structures. Hence, the brain areas involved in action value functions are likely to be related in action selection. Also, brain areas involved in motor control are likely to be related in action selection (Lee, Seo, & Jung, 2012). However, the character of a behavioral task may change the precise anatomical location involved in action selection. For instance, a well-trained motor sequence (fixed stimulus-response association) may rely more on the dorsolateral striatum (Hikosaka et al., 1999; Yin & Knowlton, 2004, 2006; Yin, 2010), whereas the dorsomedial striatum may be rely more on to perform flexible goal-directed behaviors (Yin, Knowlton, & Balleine, 2005; Yin, Ostlund, Knowlton, & Balleine, 2005). Moreover, recent study using transient optogenetic stimulation of dorsal striatal dopamine D1 and D2 receptor–expressing neurons during decision-making found that the striatal activity is involved in goal-directed action selection (Tai et al., 2012). There are cumulated evidence showing that the lateral intraparietal cortex (LIP) (Roitman & Shadlen, 2002; Rorie, Gao, McClelland, & Newsome, 2010; Seo et al., 2009), frontal eye field (Ding & Gold, 2012), and superior colliculus (Horwitz & Newsome, 2001) are involved in selecting a specific physical movement.

In addition, other brain areas may be related to more abstract action selection (Lee, Seo, & Jung, 2012). Action selections like making choices among different objects or goods may rely more on the orbitofrontal cortex (Padoa-Schioppa & Assad, 2006; Padoa-Schioppa, 2011). Compared to the orbitofrontal cortex, the medial frontal cortex may be involved more in action selection guided by endogenous cues (for example, memory) rather than external sensory stimuli. The medial frontal cortex, including the anterior cingulate cortex (Kennerley, Walton, Behrens, Buckley, & Rushworth, 2006; Lee, Rushworth, Walton, Watanabe, & Sakagami, 2007; Shidara & Richmond, 2002) and supplementary motor area (Okano & Tanji, 1987; Sohn & Lee, 2007; Soon, Brass, Heinze, & Haynes, 2008; Sul, Jo, Lee, & Jung, 2011), may integrate the information about the costs and benefits of particular behaviors and take action. Furthermore, it has been proposed that the anterior cingulate cortex might play a more important role in selecting an action voluntarily and monitoring its outcomes (Kennerley et al., 2006; Quilodran, Rothé, & Procyk, 2008; Rushworth, Walton, Kennerley, & Bannerman, 2004).

1.4. Neural mechanisms for updating value functions. Value updating functions can be divided into two parts. First, subjects need to relate an action to its corresponding outcome correctly. Deficit of this function could interfere with the process of updating value functions suitably. Previous studies showed that subjects with lesions in the orbitofrontal cortex are impaired in reversal learning tasks (Fellows & Farah, 2003; Murray, O'Doherty, & Schoenbaum, 2007; Schoenbaum, Nugent, Saddoris, & Setlow, 2002), and the deficits produced by the lesions were due to animals' choice behavior no longer reflected the history of precise conjoint relationships between particular choices and particular rewards (Walton, Behrens, Buckley, Rudebeck, & Rushworth, 2010). Thus, orbitofrontal cortex may be a critical brain area to associate an action and its corresponding outcome correctly.

Second, subjects need to realize the difference between expected reward and actual reward (i.e. the reward prediction error signal) and use this information to update the value functions. Signals related to reward prediction error were first identified in the midbrain dopamine neurons (Schultz, 1997). Recent studies found that it also exists in many brain areas, including the lateral habenula (Matsumoto & Hikosaka, 2007), globus pallidus (Hong & Hikosaka, 2008), dorsolateral prefrontal cortex (Asaad & Eskandar, 2011), anterior cingulate cortex (Seo & Lee, 2007), orbitofrontal cortex (Sul et al., 2010), and striatum (Asaad & Eskandar, 2011; Kim et al., 2009; Oyama, Hernádi, Iijima, & Tsutsui, 2010). Thus, dopamine neurons may play an important role in relaying these error signals to update the value functions represented broadly in different brain areas. Brain areas related to chosen value are also widespread, including orbitofrontal cortex (Padoa-Schioppa & Assad, 2006; Sul et al., 2010), medial frontal cortex (Sul et al., 2010), ventromedial prefrontal cortex (Hare et al., 2009), dorsolateral prefrontal cortex (Hare et al., 2009), and striatum (Cai et al., 2011; Kim et al., 2009; Lau & Glimcher, 2008). Thus, brain areas related to the chosen value and reward prediction error overlapped, such as the orbitofrontal cortex, dorsolateral prefrontal cortex and striatum. These brain areas may therefore play an important role in updating the value functions.

2. A general introduction of reinforcement learning and related models

Reinforcement learning (Sutton & Barto, 1998), a field that gets ideas from psychological theory (for example, Pavlovian and instrumental conditioning) and developed within the artificial learning community, has provided a normative framework within which such observed behavior can be understood. Reinforcement learning regards decision making as an adaptive process in which an animal utilizes its previous experience to improve the outcomes of future choices. In order to link the observed behavior and the neural functions together, decision making process is represented through complex algorithms and various mathematical models in the field of reinforcement learning. The field has developed strong mathematical foundations and various applications. The computational study of reinforcement learning is now a large field, with researchers in diverse disciplines such as psychology, control theory, artificial intelligence, and neuroscience. The field also plays a central role in the newly emerging areas of neuroeconomics and decision neuroscience. In the following section, a series of basic concepts in reinforcement learning are briefly introduced.

2.1. The basics of dopamine and reinforcement learning. The majority of dopamine secreting neurons reside in the midbrain and forms three cell groups (Bentivoglio & Morelli, 2005): the substantia nigra pars compacta (SNc; A9), the ventral tegmental area (VTA; A10), and the retrorubral nucleus which lies caudal and dorsal to the substantia nigra (RRN; cell group A8 in the rat). Studies suggested three distinct ascending dopamine projection systems from the SN–VTA complex, the mesostriatal, mesolimbic and mesocortical pathways, with widespread projections to forebrain targets (Björklund & Dunnett, 2007; Fallon & Moore, 1978; Lindvall, Bjorklund, & Divac, 1977; Lindvall & Bjorklund, 1974). The mesolimbic pathway projects dopamine axons from the SN–VTA complex to limbic areas, including amygdala, olfactory tubercle and septum. The mesocortical pathway projects to the isocortex (including prefrontal, cingulate, entorhinal, and perirhinal cortex) and allocortex (including olfactory bulb, anterior olfactory nucleus, and piriform cortex). The mesostriatal pathway projects to the striatum and nucleus accumbens.

The original link between dopamine neurons and reinforcement learning started from a series of recording studies done by Wolfram Schultz. It revealed that dopamine neurons from the SN–VTA complex responded with a phasic burst of spikes to unexpected rewards. However, if food delivery was consistently preceded by a tone or light, the response of dopamine neurons to the reward disappeared after a number of trials. The monkeys began showing conditioned responses of anticipatory licking and arm movements to the reward-predictive stimulus. Furthermore, not only the monkeys' responses to the tone, but also their dopamine neurons began responding to the tone, exhibiting phasic bursts of activity whenever the tone came on. On the other hand, when cued reward fails to arrive, dopamine neurons exhibit a momentary pause in their background firing, timed to the moment reward was expected (Hollerman & Schultz, 1998; Schultz, 1997). After years of research, converging evidence links reinforcement learning to dopamine neurons, assigning them precise computational roles. Specifically, electrophysiological recordings in behaving animals and functional imaging of human decision-making have revealed in the brain the existence of a key reinforcement learning signal, the reward prediction error (Bayer & Glimcher, 2005; Montague, Hyman, & Cohen, 2004; Schultz, 2010). Taking into consideration that many brain areas have been reported to be related to reward prediction error, dopamine neurons may play an important role in relaying these error signals to update the value functions represented broadly in different brain areas.

2.2. Rescorla-Wagner model. From the perspective of reinforcement learning, classical conditioning is considered as a typical instance of prediction learning (i.e., learning the predictive relationships between events in the environment). The Rescorla-Wagner model (Wagner & Rescorla, 1972), which was developed from the Bush and Mosteller stochastic model of learning (Bush & Mosteller, 1955), postulated that learning occurs only when events violate expectations. For instance, in a training session of classical conditioning, an unconditional stimulus (US) such as food pellets are paired with two conditional stimuli such as the sound of a tuning fork (CS1) and a light (CS2). In every trial, the sound of a tuning fork appears first, following by the light and finally the food pellets show up. According to the following equation, the associative strength of each of the conditional stimuli V (CSi) with the paired unconditional stimulus (US) will change in a trial by trial basis (Niv, 2009).

$$\mathrm{V_{new}}\ (\mathrm{CS_i}) = \mathrm{V_{old}}(\mathrm{CS_i}) + \ \eta \times \left[\lambda(\mathrm{US}) - \sum_{\mathrm{i}} \mathrm{V_{old}}(\mathrm{CS_i})\right]$$

Learning is driven by the difference between what was expected ($\Sigma\ _i$V (CS$_i$), i indexes all the CS$_s$ present in the trial) and what actually happened (λ(US), quantification of the maximal associative strength). η is a learning rate, and its value which depends on the salience properties of both the unconditional and the conditional stimuli being associated.

2.3. Temporal difference learning model. Compared to the Rescorla-Wagner model, temporal difference (TD) learning model is an elaborated model. It started from phenomena which are not explained under the Rescorla-Wagner model, such as second-order conditioning, and made predictions sensitive to the temporal relationships within a learning trial (Sutton & Barto, 1990). TD learning is a combination of two ideas from reinforcement learning theory, the Monte Carlo idea and the dynamic programming (DP) idea (Sutton & Barto, 1998; Sutton & Barto, 1990).

In TD learning, the goal of the learning system is to maximize the benefit. In order to reach the goal, the learning system needs to evaluate the estimated values of every states or situations, in terms of the possible outcomes (such as future rewards or punishments). According to that, the learning system learns at every time point within a trial, as shown in the following equation (Niv, 2009):

$$V_{\mathrm{new}}(S_i, t) = V_{\mathrm{old}}(S_i, t) + \eta \left[r(t) + \gamma \sum_{S_k, t+1} V_{\mathrm{old}}(S_k, t+1) - \sum_{S_j, t} V_{\mathrm{old}}(S_j, t) \right]$$

On the basis of the above equation, every stimulus (S_i, S_k, S_j) makes long-lasting memory traces (representations)with paired value ($V(S_i,t)$, $V(S_j,t)$, $V(S_k,t)$) which is learned for every state of this trace. η is still the learning rate as in the Rescorla-Wagner model, so as the learning is driven by the difference between actual (r(t), the reward observed at time t) and expected outcome. Nevertheless, unlike the Rescorla-Wagner model, the associative strength of the stimuli at time t is not only taken to predict the immediately forthcoming reward r(t), but also the future predictions due to those stimuli that will still be used in the next time step $\sum_{S_k \text{ at } t+1} V(S_{k,t+1})$ along with γ ($0 \leq \gamma \leq 1$) discounting these future delayed predictions.

2.4. Q-learning model. The whole purpose of prediction learning is to help selecting actions. Since the environment rewards us for our actions instead for our predictions, we need to take “action” into the Markov decision process. Q-learning model, a modified TD learning model, postulated that agent learns explicitly the predictive value (Q(S,a), the expected future reward) of taking a specific action a at a certain state S. Thus, the value learning was updated according to the following rule (Niv, 2009; Sutton & Barto, 1998; Watkins, 1989).

$$Q(S_t, a_t)_{new} = Q(S_t, a_t)_{old} + \eta \cdot \delta_t$$

$$\delta_t = r_t + \max_a \gamma Q\,(S_{t+1}, a) - Q(S_t, a_t)$$

The $\max_a$ operator represents the best available action at the subsequent state $S_{t+1.}$ Since Q-learning takes into account the best future action, it is considered an “off-policy” method, regardless of the possibility that this may not be the actual action taken at the subsequent state $S_{t+1.}$ According to that, in Q-learning, action selection is simply taking the highest Q(S,a) value. However, in a real world scenario, action selection is also stochastically dependent. For a given state s, the action value $Q(S,a_i)$ for the candidate action a_i (i = 1,…, m) are compared and the one with a higher action value is selected with a higher probability. This is the so-called softmax rule or

Boltzmann exploration (Kaelbling, Littman, & Moore, 1996), a logistic form that assigned a weight to each of the actions according to their action value estimation:

$$\mathrm{P}(\mathrm{a_i} \mid \mathrm{s}) = \frac{e^{\beta Q(\mathrm{S},\mathrm{a_i})}}{\sum_{\mathrm{i}=1}^{\mathrm{m}} e^{\beta Q(\mathrm{S},\mathrm{a_i})}}$$

The parameter β, which is called the inverse temperature, represents choice perseveration (or exploration/exploitation), a term referring to the tendency of making actions guided by reward values. A zero value of β means the agent will choose the action at random. Thus, the hypothesis of Q-learning included not only the predictive value, but also the action to explain behaviors. And it was postulated that learning is to optimize the consequences of actions in terms of some long-term measure of total obtained rewards (and/or avoided punishments). Somehow, this hypothesis seemed to be similar to the one which instrumental conditioning proposed. Thus, the study of instrumental conditioning, using TD learning model (consider both value and action), could be an approach into the fundamental form of rational decision-making.

3. An overview of striatum: anatomy and neural circuits

3.1. Anatomy of striatum. The striatum is the principal input structure of the basal ganglia that influences motor control and reward-based learning (Chang, Chen, Luo, Shi, & Woodward, 2002; Lauwereyns, Watanabe, & Coe, 2002; Tanaka et al., 2006). The principal neurons in the striatum are medium spiny neurons (MSN), which represent over 95% of total neurons. These GABAergic neurons receive two major glutamatergic inputs from the cortex and the thalamus (Kreitzer & Malenka, 2008; Lovinger, 2010; Surmeier, Ding, Day, Wang, & Shen, 2007). MSNs also receive dopaminergic inputs from the SN-VTA complex, and regulation of MSN by dopamine is important for reward learning (Lee, Seo, & Jung, 2012; Oyama et al., 2010; Schultz, 2006).

Evidence showed that the MSNs can be further divided into two categories: the striatonigral MSNs and the striatopallidal MSNs. The striatonigral MSNs express D1-like receptors, group I mGluRs (mGluR1/5), M1 and M4 muscarinic receptors, while the striatopallidal MSNs express D2-like receptors, M1 muscarinic receptors, adenosine A2A receptors and group I mGluRs (mGluR1/5) (Kreitzer & Malenka, 2008). Both subgroups of MSNs are morphologically indistinguishable and mosaically distributed (Gerfen & Young, 1988; Gerfen, 1992; Giménez-Amaya & Graybiel, 1990). However, recent studies using technique of bacterial artificial chromosome (BAC) mediated transgenesis in mice has shown differences of basal electrophysiological properties and synaptic plasticity between the striatonigral and striatopallidal MSNs (Kreitzer & Malenka, 2007; Shen, Flajolet, Greengard, &

Surmeier, 2008).

In addition, MSNs receive GABAergic synapse from local interneurons as well as other MSNs (Kawaguchi, Wilson, Augood, & Emson, 1995; Kreitzer, 2009). Striatal interneurons are grouped into four types based on the cytochemical, physiological and morphological properties. The giant cholinergic interneurons with large soma are the source of acetylcholine (ACh) in the striatum and their axonal fields are extensive compared with other interneurons. Cholinergic interneurons display tonic irregular firing pattern and are featured by a long duration after hyperpolarization, hence are also called long duration after hyperpolarization cells. The second type of interneuron is the parvalbumin-containing cell which composes 3-5% of total striatal neurons and is characterized as fast-spiking firing pattern *in vitro*. The third type of interneuron is the somatostatin (Neuropeptide Y, NOS)-containing interneuron which represents 1-2% of total striatal neurons, and the dendrites of which are relatively unbranched for longer distances. Somatostatin-containing interneuron is featured by Ca^{2+}-dependent low threshold spikes *in vitro*. The fourth type of interneuron is the calretinin-containing interneuron, the phenotype and physiology of which have not been well established (Kawaguchi et al., 1995; Kreitzer, 2009; Lovinger, 2010).

There are two pathways of projections of MSNs. One is called the direct pathway and the other is called the indirect pathway (Albin, Young, & Penney, 1989; Garrett E. Alexander & Crutcher, 1990; DeLong, 1990). The direct-pathway circuit originates from striatonigral MSNs, which project to GABAergic neurons in the internal globus pallidus (GPi in primates, GPm in rodents) and substantia nigra pars reticulata (SNr), and the GPi and SNr send axons to motor nuclei of the thalamus. The net effect of direct-pathway activity is a disinhibition of excitatory thalamocortical projections, leading to activation of cortical premotor circuits and the facilitation of movement. The indirect-pathway circuit originates from striatopallidal MSNs, which inhibit neurons in the globus pallidus (GP), which in turn project to glutamatergic neurons in the subthalamic nucleus (STN). Subthalamic neurons send axons to basal ganglia output nuclei (GPi and SNr), where they form excitatory synapses on the inhibitory output neurons. The net effect of indirect-pathway activity is an inhibition of thalamocortical projection neurons, which would reduce cortical premotor drive and inhibit movement.

3.2. Cortico-striatal circuits involved in decision making. Traditionally, the striatum has been divided into dorsal and ventral subregions. The dorsal subregion contains the dorsolateral striatum (DLS) and dorsomedial striatum (DMS). The ventral subregion contains the nucleus accumbens (NA), which itself consists of core and shell subregions (Alexander, DeLong, & Strick, 1986; Groenewegen, Berendse, Wolters, & Lohman, 1991; Zahm, 2000). The cortical inputs to striatum are topographically organized, with limbic and ventral prefrontal regions projecting to the ventral striatum, sensorimotor cortical regions projecting to the DLS and association areas of the prefrontal cortex projecting to the DMS (Alexander et al., 1986; Groenewegen et al., 1991). The connectivity between cortico-striatal regions has lead to the idea that cortico-basal-ganglia loop are corresponded to functional circuits that mediate distinct components of behavior. And researches focused on the different subregions of striatum somehow confirmed this point of view.

1. DMS: Local blockade of NMDA receptors and lesion studies all showed that DMS is crucial for the acquisition and expression of goal-directed actions (Gremel & Costa, 2013; Yin et al., 2005; Yin & Knowlton, 2004, 2006; Yin et al., 2005). However, some researchers found that the DMS may not support effort- and reward-related decision making but the flexibility of spatially guided behavior (Braun & Hauber, 2011; Ragozzino, Jih, & Tzavos, 2002; Ragozzino, Ragozzino, Mizumori, & Kesner, 2002; Ragozzino, 2007).

2. DLS: For DLS, almost all studies confirmed it crucial to habit formation (Gremel & Costa, 2013; Yin & Knowlton, 2004, 2006).

3. NA: Previous studies demonstrated that the NA plays an important role on the acquisition and reversal of instrumental contingencies (Annett, McGregor, & Robbins, 1989; Balleine & Killcross, 1994; Taghzouti, Louilot, Herman, Le Moal, & Simon, 1985), while others found that lesions of NA did not disrupt reversal performance in a go-no go odor discrimination paradigm (Schoenbaum & Setlow, 2003) and in a delayed matching task (Burk & Mair, 2001). In sum, studies investigating the contribution of the NA in reversal learning are controversial. On the other hand, there is evidence for the participation of the NA, and in particular its core sub-region, in behavioral flexibility involving changes in strategies or rules (Floresco, Ghods-Sharifi, Vexelman, & Magyar, 2006; Haluk & Floresco, 2009). Also, NA was described as having a role in the expression of conditioned emotional responses to cues and contexts associated with appetitive (or aversive) events (Belin, Jonkman, Dickinson, Robbins, & Everitt, 2009; Day & Carelli, 2007).

Despite the inconsistency, Shiflett and Balleine cnocluded the previous findings on rodents and proposed a cortico-striatal circuits involved in decision making process

(Shiflett & Balleine, 2011). According to the previous defined subregions, there are three pathways:

1. The dorsomedial striatum, also known as “associative striatum” in primates, which receives inputs from association areas of the prefrontal cortex is implicated in goal-directed behavior (i.e. reward –related actions) in rodents.
2. The dorsolateral striatum, a part of the sensorimotor striatum in primates, is related to habit learning (i.e. stimulus-response bound actions) in rodents.
3. The nucleus accumbens (NA) is implicated in representing predicted future reward, and the representations can be used to guide both goal-directed and habitual actions.

Furthermore, the basal ganglia contain intrinsic feedforward and feedback circuits that may be crucial for striatal function. In particular, bidirectional connections of striatum and midbrain through the SN-VTA complex have been found to connect neighboring striatal regions. This spiraling architecture links NA to the DMS, and the DMS to the DLS (Haber, Fudge, & McFarland, 2000). Also, as previously mentioned, the interneurons in the striatum may also contribute to connect neighboring striatal subregions. these connections may enable striatal subregions to work cooperatively to support the transition from goal-directed to habitual behavior, as well as enable information of predictied reward (from NA) to influence action control mediated by dorsal striatum (Ito & Doya, 2011; Yin, Ostlund, & Balleine, 2008).

4. The objective of this study

Through literature review, striatum has shown to participate in every step of decision making process, including value representation, action selection, and value updating functions. Furthermore, striatum is the principal input structure of the basal ganglia and cortical inputs to striatum are topographically organized, implying a functional circuits that mediate distinct components of behavior (Alexander et al., 1986; Groenewegen et al., 1991). It was reported that the DMS is implicated in goal-directed behavior in rodents, the DLS is related to habit learning in rodents, and the NA is implicated in representing predicted future reward, and the representations can be used to guide action selection for reward (Shiflett & Balleine, 2011). However, as previously mentioned, findings concerning functions of striatal subregions are somehow controversy, and the precise mechanism or role of each subregion in reinforcement learning and reward-based decision making is still under debate.

Furthermore, many previous studies on the DMS used outcome devaluation and contingency degradation as methods to detect whether action-outcome contingency changes after specific manipulation (for instance, lesion and drug manipulation) (Gremel & Costa, 2013; Yin et al., 2005; Yin et al., 2005), and results of these studies confirmed that the DMS is crucial for goal-directed behavior.

However, these studies did not directly look into the learning process, but used a post-learning assessment, examining the disappearance of an action-outcome association. These researchers used the idea that how fast a belief can be destroyed to answer the question concerning the functions of DMS. Accordingly, in the current study, we want to directly look into the learning process (i.e., to examine the process of building up an action-outcome association). Thus, the aim of this study is to examine the role of different striatal subregions (including the DMS, DLS, and NA) in reinforcement learning process and reward prediction error using excitotoxic lesions and a 2-choice dynamic foraging task in male C57/Bl6 mice. The 2-choice dynamic foraging task is a risky-choices task which consisted of 1:3 and 1:6 reward ratio as a whole learning process. Using Q-learning model and matching law analysis, the trial-by-trial choice behaviors of mice were further analyzed to elaborate parameters for RPE and reward sensitivity.

Chapter 2: Materials and Methods

1. Animals

Male C57BL/6J purchased from National Taiwan University Hospital were housed with food and water available *ad libitum* in polysulfone individually ventilated cages (Alternative Design Manufacturing & Supply, Arkansas, AR, USA) within the animal rooms in the Psychology Department, National Taiwan University. All animals were 2.5-5 month-old at the beginning of experiments. Animals were housed individually and handled at least 1 week before the behavioral experiments, and behavioral experiments were conducted during the dark phase at least half an hour after dark/light cycle began. Animals were brought to the behavioral room 30 min before experiments. All animal procedures were performed according to protocols approved by the appropriate Animal Care and Use Committees established by the National Taiwan University.

2. Experimental apparatus

Behavioral apparatus were two custom-built 5-aperture operant chambers (31.8 L × 25.8 W × 29.1 H cm^3; Coulbourn Instruments, Whitehall, PA, USA) in a behavioral testing room under a red lighting condition (11.4 lux). Each chamber had a stainless-steel grid floor, aluminum front and back modular walls, aluminum top with a hole (4 cm diameter) in the center, and clear acrylic sides. Five 1.5 cm diameter and 4 cm deep stimulus-response apertures were spaced 3 cm apart, 1 cm above the grid floor, and centered on the front, curved wall of the chamber. Each stimulus-response aperture contained three pair of white light-emitting diode (LED) lights to generate a light stimulus and a photocell sensor to signal nose poke responses. The 3 apertures in the middle were covered by a white opaque acrylic (22 L × 15 W × 0.3 H cm^3) throughout the experiment and only the 2 apertures on the side of the curved wall of the chamber were used in this study. The magazine was located in the low center of the back wall of the chamber with a yellow LED light fitted in the magazine as a cue of nose poke responses, and was spanned horizontally by a photocell sensor to signal nose poke responses. Below the magazine was a reward deliver to dispense 2 % sucrose solution. A 3 W house light was mounted above the magazine. The Graphic State 3.03 (Coulbourn Instruments, Whitehall, PA, USA) was used to perform on-line control of this apparatus and data collection.

3. Experimental procedures

3.1. Water restriction schedule. Animals were water-restricted to 80-85% of free-drinking body weight throughout the 2-choice dynamic foraging task with daily weighed. Water was given daily in their home cages at least an hour after they finished experiment. Food was available *ad libitum* in their home cages throughout the behavioral experiments.

3.2. Open field task. To measure the spontaneous locomotor activity before and after the surgery, each mouse was placed into a polyvinylchloride chamber (48 cm x 24 cm x 25 cm) for 60 minutes. Total travel distance was recorded using EthoVision video tracking system (Noldus Information Technology, Netherlands).

3.3. Surgery. Mice were anesthetised with isoflurane and placed in a stereotaxic frame fitted with an isoflurane gas anesthesia system. The scalp was incised and the skin retracted. Bregma and lambda were leveled in the horizontal plane. Bilateral burr holes were drilled through the skull according to the following coordinates, measured from bregma: dorsal medial striatum lesion (AP, + 0.5 mm; ML, ± 1.5 mm; DV, - 3 mm), dorsal lateral striatum lesion (AP, + 0.5 mm; ML, ± 2.5 mm; DV, - 3 mm), nucleus accumbens lesion (AP, + 1.8 mm; ML, ± 1.1 mm; DV, - 4.7 mm), as shown in Figure 2. 1. Injector was lowered to the target coordinates and N-methyl-D-aspartate (NMDA, 20 mg/mL; Sigma), dissolved in sterilized saline, was infused (via Hamilton syringe). Because the striatum are surrounded by fibers of passage and lesion effect may be confounded by the damage of fibers passing by (such as electrolytic lesion), we made lesions using the excitotoxin NMDA, which destroys intrinsic neurons, but not fibers of passage (Mayer & Westbrook, 1987). According to the previous studies, the effect of lesion can maintain three months (Pothuizen, Jongen-Rêlo, Feldon, & Yee, 2005), this is the other reason we made lesions using the excitotoxin NMDA. Sham animals received saline alone. The NMDA or vehicle was infused at a volume of 0.2 μL per infusion (manually across 5 min). The syringe remained in place for an additional 5 min to allow for diffusion of the drug. Following the infusion, the incision was sutured with bone wax. Mice were allowed to recover for 7 days prior to the start of behavioral testing.

3.4. Sucrose preference test. A two-bottle sucrose preference test was used to evaluate reward sensitivity after lesion surgery. Each mouse was individually tested in their home cages. Drinking water was first filled in the two bottles on day 1 and day 2 to obtain a drinking baseline and to make sure there was no side preference. Subsequently, bottles were filled with drinking water and 2% sucrose solution, respectively, on day 3 and day 4. The daily fluid intake was measured by weighing the bottles; the positions of the bottles were alternated every day. The daily sucrose preference was calculated for each mouse as follows: 100 × [weight of 2% fluid intake / (weight of water intake + weight of 2% fluid intake)].

3.5. Two-choice dynamic foraging task. Animals were trained and tested in a

2-choice dynamic foraging task modified from the dynamic foraging task used in human and mice previously (Chen et al., 2012; Rutledge, Lazzaro, Lau, Myers, Gluck, & Glimcher, 2009). It was a two-alternative forced-choice task, and one of the alternative apertures presented a reward at a high rate, while independently, the probability of receiving a reward in the other aperture was low. Animals conducted a 45-min daily session per day. The procedure consisted of a shaping phase and a 2-reward-ratio testing phases: the 1:3 reward ratio and the 1:6 reward ratio, as depicted in Figure 2. 2.

3.5.1. ***Shaping phase***. Before surgery, mice were first trained to operate the experimental apparatus by a series of 4 shaping stages. In each stage, each mouse was required to reach shaping criteria in 45 minutes, and then they could move to the next stage. During the first 4 shaping stages, a trial started with the illumination of the house light, and ended after animals collected their reward following a new trial started automatically. Besides, the magazine illuminated to signal the delivery of a reward. Stage 1 (MAG10): Animals were required accumulating 10 nose pokes into either the 2 stimulus-response apertures or the magazine, and each nose poke was followed by the delivery of a reward. Stage 2 (M5H5): Animals were still required to perform a nose poke into the magazine followed by the delivery of a reward. But after accumulating 5 nose pokes into the magazine, no reward was delivered from the magazine if the animal kept performing nose pokes into the magazine. Each mouse was required accumulating 5 nose pokes into one of the 2 apertures, and each nose poke into stimulus-response apertures was followed by the delivery of a reward. Stage 3 (M0H10): Each mouse was required accumulating 10 nose pokes into one of the 2 stimulus-response apertures, and nose poking into the magazine was not followed by any delivery of a reward. Additionally, each nose poke into stimulus-response apertures was followed by the delivery of a reward. Stage 4 (H11): A trial started with the illumination of the house light, and then mice had to wait an intertrial interval (ITI) of 5 sec for the illumination of stimulus-response apertures. The 2 apertures subsequently illuminated, and animals were required to accumulating 11 nose pokes into one of the illuminated apertures to show their preference for left or right stimulus-response apertures, and each nose poke into apertures was followed by the delivery of a reward.

3.5.2. ***Testing phase.*** After surgery, mice went on shaping phase (only stage 4) again to show their preference for left or right stimulus-response apertures. After mice completed stage 4, next day started the testing phase. The testing phase consisted of 2 reward ratio testing phases: the 1:3 reward ratio (including acquisition of the 1:3

reward ratio and reversal of the 1:3 reward ratio); the 1:6 reward ratio (including acquisition of the 1:6 reward ratio and reversal of the 1:6 reward ratio). The 1:3 reward ratio contained the reward rate of 20 % and 60 % in one of the 2 stimulus-response apertures. The 1:6 reward ratio had the reward rate of 11.43% and 68.57% in one of the 2 stimulus-response apertures. The location of high and low reward aperture was switched back and forth one day after each mouse completed preset criteria in each section, as shown in Figure 2. 2. On each day, each animal underwent a 45 minutes daily session or maximum 6 blocks (a block consisted of 10 trials). Daily session began with the illumination of house and magazine lights. A nose poke into the magazine initiated a trial and extinguished the magazine light. A fixed ITI of 5 sec preceded the illumination of stimulus-response apertures. The 2 stimulus-response apertures subsequently illuminated after the ITI, and animals were required nose poking into one of the illuminated apertures. Each nose poke into the illuminated aperture was followed by either the delivery of a reward or no any reward, and both of them were subsequently followed by the illumination of magazine. Each trial ended after animals collected earned reward or after animals nose poked into the illuminated magazine. Each mouse discovered these rules and chose the high reward rate aperture by trial and error. The criteria of accomplishing each section was accumulating choice of the high reward rate aperture for at least 70% accuracy in 3 consecutive blocks. Once the criterion was achieved, each mouse moved on to the next section on the next testing days and the reward rates of the 2 apertures were switched. If mice couldn’t reach the criterion after accumulating over 900 trials, mice also moved on to the next section on the next testing days. Accumulated trials, choice results, and latency both to response to the illuminated apertures and to reach the magazine were recorded trial by trial by computer software during daily training.

3.6. Histology. Mice were perfused and the brains post-fixed with 4% paraformaldehyde, with lesion placement identified through Nissl staining of 40-μm brain slices. Only mice with lesions located with DMS, DLS or OFC were included.

4. Data analysis

4.1. Q learning model. A standard reinforcement learning model was applied to estimate RPE in the 2-choice dynamic foraging task. As typically seen in other modeling work, the reinforcement learning model constitutes one value updating component (i.e. how information is updated) and one choice component (i.e. how choice is made). For the value updating rule, we used a simplified Q-learning model, which belongs to the family of temporal difference models, to characterize the dynamic process of RPE in the 2-choice dynamic foraging task (Sutton & Barto, 1998;

Watkins & Dayan, 1992). Such a rule proposes that an RPE is updated whenever the subject’s expected reward changes on each trial. Thus, the value chosen from the high-reward aperture for each trial was updated according to the following rule (Rutledge et al., 2009).

$$Q_{\mathrm{high}}(t+1) = Q_{\mathrm{high}}(t) + \boldsymbol{\alpha\delta}(t) \qquad \boldsymbol{\delta}(t) = R_{\mathrm{high}}(t) - Q_{\mathrm{high}}(t)$$

$$Q_{\mathrm{low}}(t+1) = Q_{\mathrm{low}}(t) + \alpha\delta(t) \qquad \delta(t) = R_{\mathrm{low}}(t) - Q_{\mathrm{low}}(t)$$

where $Q_{high}(t)$ is the expected value associated with choosing the high-reward rate aperture on trial t and δ(t) is the RPE representing the discrepancy between expectation and the reward just received. $R_{high}(t)$ denotes the actual outcome received from the high-reward rate aperture on trial t. The parameter α represents the learning rate, which determines how rapidly the reward prediction error signal is updated. Because the onsets of stimuli and outcomes were modeled trial-by-trial as separate δ(t) at the time of each feedback display during each trial, the magnitude of RPE was determined by the learning rate (α) from the trial-by-trial data in each testing section of the 2-choice dynamic foraging task.

Reinforcement learning also requires a balance between exploration and exploitation. For the choice rule in the reinforcement learning model, it is assumed that the probability of choosing the high-reward aperture $P_{high}(t + 1)$ was determined by the so-called softmax rule or Boltzmann exploration (Kaelbling et al., 1996), a logistic form that assigned a weight to each of the actions according to their action value estimation:

$$\mathrm{P}_{\mathrm{high}}\,(\mathrm{t}+1)=\frac{e^{\beta Q_{\mathrm{high}}(\mathrm{t})}}{e^{\beta Q_{\mathrm{high}}(\mathrm{t})}+e^{\beta Q_{\mathrm{low}}(\mathrm{t})}}$$

$$\mathrm{P}_{\mathrm{low}}\,(\mathrm{t}+1)=\frac{e^{\beta Q_{\mathrm{low}}(\mathrm{t})}}{e^{\beta Q_{\mathrm{high}}(\mathrm{t})}+e^{\beta Q_{\mathrm{low}}(\mathrm{t})}}$$

The parameter β represents choice perseveration (or exploration/exploitation), a term referring to the tendency of making actions guided by reward values. A zero value of β means the subject will choose the high-reward rate aperture at random. To estimate the learning rate (α) and the choice perseveration (β), we used a hierarchical modeling approach called Markov Chain Monte Carlo (MCMC)-based Bayesian parameter estimation to fit the reinforcement learning model to the trial-by-trial data from the 2-choice dynamic foraging task (Lee & Wagenmakers, 2014; Wetzels, Lee, & Wagenmakers, 2010). The advantage of the Bayesian approach is that it can account for inter-subject variability and other random effects in a more rigorous and satisfactory way using latent parameters. In particular, from the Bayesian perspective, parameters are described by informative probability distributions instead of point estimations. A probit transformation was used to make the construction of the Bayesian hierarchical model easier. Because the Bayesian hierarchical model requires the number of input trials to be the same, we cut the cumulated trials into the same number by use of the smallest cumulated trials as a cutting point in the lesion and sham groups. The structure of this Bayesian hierarchical modeling is depicted in Figure 2. 3. As shown in Figure 2.3, the parameters α and β for subject i (α_i and β_i) were each assumed normally distributed with respective means and standard deviations, which were from the group level of distributions (i.e. μ_a σ_a and μ_b σ_b , respectively). We used WinBUGS [the MS Windows operating system version of BUGS (Bayesian inference Using Gibbs Sampling)] and WinBUGS Development Interface (Lunn, Thomas, Best, & Spiegelhalter, 2000) to approximate the distributions of parameters by sampling values using the MCMC technique. A chain consisted of 28000 iterations, of which the first 8000 (burn-in) points were discarded to ensure that only samples from the stationary distribution were used and that the data were unaffected by the starting value. Thus, we obtained 60000 points of estimation from the three chains and collected samples at intervals of every five samples, which yielded 12000 points. All interpretations and tests were performed based on these 12000 samples. Parameters between lesion and sham groups were compared by computing the difference between the values of the two posterior distributions in each run obtained from the hierarchical Bayesian estimation. One way to evaluate the strength of evidence for differences in group-mean parameters is by checking whether the probability of the posterior distribution of differences is greater (or less) than zero (Fridberg et al., 2010). Another way is to use the Bayes factor (BF), an odd ratio of marginal likelihood of the two models (or hypotheses) of interest, to index the evidence strength of the alternative hypothesis against the null hypothesis (Kass & Raftery, 1995; Raftery, 1995). A large BF value (> 3) would (at least) “positively” favor the alternative hypothesis and a BF value between 1 and 3 would “weakly” favor the alternative hypothesis, as shown in Table 2. 1. To evaluate the differences of group-mean parameters, a method based on the Savage-Dickey density ratio was used to compute the BF values (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010).

4.2. Matching law analysis. To assess the degree to which animals in the 2-choice dynamic foraging task made their overall average choices in accord with the received rewards, a matching law analysis was also conducted (Baum, 1974; Rutledge et al., 2009), which provides a simple empirical quantification between the rate of response and the rate of reinforcement:

$$\log_2\left(\frac{\mathrm{C}_{\mathrm{left}}}{\mathrm{C}_{\mathrm{right}}}\right) = s\,\log_2\left(\frac{\mathrm{R}_{\mathrm{left}}}{\mathrm{R}_{\mathrm{right}}}\right) + \log_2 k$$

In the above formula, $\mathrm{C}_{\mathrm{left}}$ and $\mathrm{C}_{\mathrm{right}}$ denote the number of choices to the left- and right apertures, respectively. Likewise, $\mathrm{R}_{\mathrm{left}}$ and $\mathrm{R}_{\mathrm{right}}$ are the respective number of rewards received from the left and right apertures. The slope s is thought to be a measure of the sensitivity of choice allocation to reward frequency. In this study, we used least-squares regression to fit the above formula to steady-state (last 30 trials of each testing phase) choice behavior in the 2-choice dynamic foraging task. Blocks in which one aperture was never rewarded (i.e. $\mathrm{R}_{\mathrm{left}}$ or $\mathrm{R}_{\mathrm{right}} = 0$) were excluded from the analysis in order to fit the data to the above formula.

4.3. Statistical analysis and software. The behavioral data were analyzed by the Student’s t-test or the one-way analysis of variance (ANOVA) where appropriate. Adjusted t-test was applied if the Levene’s test for equality of variances reached the significant level. Statistic analyses were performed using SPSS 20.0 (SPSS Inc., Chicago, IL, USA).

Chapter 3: Results

1. Histology

Photographs of representative infusion placements in the DLS, DMS and NA were shown in Figure 3. 1. Using Nissl staining, 3 of 13 mice in the DLS lesioned group were excluded from the study; 5 of 15 mice in the DMS lesioned group were excluded from the study; 5 of 15 mice in the NA lesioned group were excluded from the study, as shown in Figure 3. 2.

2. Behavioral data

2.1. Open field task. As shown in Figure 3. 3, no significant difference was found in the three sham groups before surgery (*F*(2,27) = 0.817, *p* = .452) and after surgery (*F*(2,27) = 1.933, *p* = .164). No significant difference was found in the DLS lesioned mice after surgery (*t*(9) = -1.324, *p* = .22). A trend of hyperlocomotion was found in the DMS lesioned mice after surgery (*t*(9) = -2.184, *p* = .057). The NA lesioned mice showed hypolocomotion after surgery (*t*(9) = 2.602, *p* = .029).

2.2. Sucrose preference test. As depicted in Figure 3. 4, no significant difference was found in the three sham groups in sucrose preference (*F*(2,27) = 1.115, *p* = .343) There is no significant difference in sucrose preference between lesioned mice and sham controls within each of the 3 groups (DLS: *t*(18) = 0.169, *p* = .87; DMS: *t*(18) = -1.607, *p* = .13; NA: *t*(18) = -0.797, *p* = .44).

2.3. Assessing motivation on performing the 2-choice task after surgery. In these 3 brain lesioned groups, there is no difference in latency to response to the illuminated apertures before and after surgery (DLS: *t*(9) = 0.266, *p* = .80; DMS: *t*(9) = 0.146, *p* = .89; NA: *t*(9) = 0.034, *p* = .97), as shown in Figure 3. 5. There is no difference in latency to collect reward between lesioned mice and sham controls within each of the 3 groups (DLS: *t*(9) = -0.908, *p* = .39; DMS: *t*(9) = -0.148, *p* = .89; NA: *t*(9) = -0.412, *p* = .69), as shown in Figure 3. 6. No significant difference was found in the three sham groups in latency to response to the illuminated apertures before (*F*(2,27) = 0.096, *p* = .909) and after surgery (*F*(2,27) = 0.881, *p* = .426). No significant difference was found in the three sham groups in latency to collect reward before (*F*(2,27) = 0.250, *p* = .781) and after surgery (*F*(2,27) = 0.508, *p* = .607).

2.4. Measurement of cumulated trials and errors in the 2-choice dynamic foraging task. For overall cumulated trials, no significant difference was found in the three sham groups (*F*(2,27) = 0.144, *p* = .866). For overall cumulated trials, no significant difference was found in the DLS (*t*(18) = -0.791, *p* = .44) and NA (*t*(18) = -1.479, *p* = .16) groups. Compared to sham mice, the DMS lesioned mice required more overall trials to reach the preset criteria (*t*(10.547) = -2.576, *p* = .027), as shown in Figure 3. 7. For cumulated trials in the 1:3 reward ratio and the 1:6 reward ratio, no significant difference was found in the three sham groups (1:3 reward ratio: *F*(2,27) = 0.472, *p* = .629; 1:6 reward ratio: *F*(2,27) = 0.211, *p* = .811). For cumulated trials in the 1:3 reward ratio and the 1:6 reward ratio, no significant difference was found in the DLS group (the 1:3 reward ratio, *t*(18) = -0.406, *p* = .69; the 1:6 reward ratio, *t*(18) = -1.274, *p* = .22). Compared to sham mice, the DMS lesioned mice required more cumulated trials to reach the preset criteria in the 1:6 reward ratio (*t*(18) = -2.155, *p* = .045) and there is a marginal significant difference in the 1:3 reward ratio (*t*(11.976) = -2.089, *p* = .059). Compared to sham controls, a trend in the 1:3 reward ratio was found in the NA lesioned mice (*t*(13.487) = -2.049, *p* = .06), as shown in Figure 3. 8.

For cumulated trials in the learning of 1:3 reward ratio, the reversal of 1:3 reward ratio, learning of 1:6 reward ratio, and reversal of 1:6 reward ratio, no significant difference was found in the DLS (1: *t*(18) = -1.668, *p* = .11; 2: *t*(18) = 0.441, *p* = .67; 3: *t*(11.536) = -1.168, *p* = .27; 4: *t*(18) = -1.119, *p* = .28) and DMS (1: *t*(10.316) = -1.278, *p* = .23; 2: *t*(12.971) = -1.122, *p* = .28; 3: *t*(18) = -1.719, *p* = .10; 4: *t*(18) = -1.341, *p* = .20) groups. For cumulated trials in every section, no significant difference was found in the three sham groups (1: *F*(2,27) = 0.693, *p* = .509; 2: *F*(2,27) = 2.182, *p* = .132; 3: *F*(2,27) = 0.641, *p* = .535; 4: *F*(2,27) = 0.012, *p* = .988). But as shown in Figure 3. 9, compared to sham controls, a trend on cumulated trials was found in the NA lesioned mice (*t*(9.933) = -2.035, *p* = .069)in the reversal of the 1:3 reward ratio.

For overall cumulated errors, no significant difference was found in the three sham groups (*F*(2,27) = 0.031, *p* = .969). For overall cumulated errors, no significant difference was found in the DLS (*t*(18) = -0.975, *p* = .34) and NA (*t*(18) = -1.396, *p* = .18) group. Compared to sham mice, the DMS lesioned mice cumulated more total errors to reach the preset criteria (*t*(9.885) = -2.583, *p* = .028), as shown in Figure 3. 10. For cumulated errors in the 1:3 reward ratio and the 1:6 reward ratio, no significant difference was found in the three sham groups (1:3 reward ratio: *F*(2,27) = 0.661, *p* = .525; 1:6 reward ratio: *F*(2,27) = 0.945, *p* = .401). For cumulated errors in the 1:3 reward ratio and the 1:6 reward ratio, no significant difference was found in the DLS group (the 1:3 reward ratio, *t*(18) = -0.594, *p* = .56; the 1:6 reward ratio, *t*(18) = -1.546, *p* = .14). Compared to sham mice, the DMS lesioned mice cumulated more errors to reach the preset criteria in the 1:6 reward ratio (*t*(18) = -2.223, *p* = .039) and there was a marginal difference in the 1:3 reward ratio (*t*(10.487) = -2.110, *p* = .06). There is a trend that the NA lesioned mice cumulated more errors to reach the preset criteria in the 1:3 reward ratio compared to sham mice (*t*(11.953) = -1.869, *p* = .086), as shown in Figure 3. 11. For cumulated errors in each of the four sections, no significant difference was found in the DLS (1: $t(18) = -1.397$, $p = .18$; 2: $t(18) = -0.036$, $p = .97$; 3: $t(18) = -1.219$, $p = .24$; 4: $t(18) = -1.291$, $p = .21$) and DMS (1: $t(18) = -1.337$, $p = .20$; 2: $t(18) = -1.216$, $p = .24$; 3: $t(18) = -1.723$, $p = .10$; 4: $t(18) = -1.027$, $p = .32$) groups; whereas the NA lesioned mice seemed to cumulate more errors in the reversal of 1:3 reward ratio compared to sham mice ($t(9.983) = -1.974$, $p = .077$), as shown in Figure 3. 12. For cumulated errorss in every section, no significant difference was found in the three sham groups (1: $F(2,27) = 1.119$, $p = .341$; 2: $F(2,27) = 2.076$, $p = .145$; 3: $F(2,27) = 1.711$, $p = .200$; 4: $F(2,27) = 0.110$, $p = .897$).

3. Matching law analysis

Using least-squares regression, trial-by-trial data from the steady state (last 30 trials in each section) of the 2-choice dynamic foraging task were fitted and used to estimate reward sensitivity. The sections in which animals gained no reward from either of the two apertures (i.e., R_{low} or $R_{high} = 0$) were excluded from analysis. As depicted in Figure 3. 13, the estimated values of reward sensitivity *s* for the DLS sham and lesion groups were 0.607 and 0.616, respectively. The estimated values of reward sensitivity for the DMS sham and lesion groups were 0.60 and 0.569, respectively. And the estimated values of reward sensitivity for the DLS sham and lesion groups were 0.676 and 0.611, respectively. There is no significant difference in reward sensitivity between sham controls and lesioned mice within each of the three groups (DLS: $t(56) = 0.247$, $p = .81$; DMS: $t(47) = -0.546$, $p = .59$; NA: $t(51) = 0.536$, $p = .60$).

4. Estimation of learning rate and choice perseveration using reinforcement learning model

As depicted in Figure 3. 14, the posterior sample means and their 95% credible intervals (CI) of learning rate (α) for the DLS sham and lesion groups were 0.0072 (CI = (0.0040, 0.0124)) and 0.0069 (CI = (0.0036, 0.0120)), respectively. The posterior sample means and their 95% credible intervals (CI) of learning rate (α) for the DMS sham and lesion groups were 0.0074 (CI = (0.0038, 0.0137)) and 0.0033 (CI = (0.0014, 0.0068)), respectively. The posterior sample means and their 95% credible intervals (CI) of learning rate (α) for the NA sham and lesion groups were 0.0078 (CI = (0.0048, 0.0124)) and 0.0039 (CI = (0.0023, 0.0065)), respectively. Besides, the probability of the posterior distribution of group mean differences of the parameter α between sham and lesion groups for the DLS, DMS and NA groups were 0.556, 0.957, and 0.978, respectively. The Results from the DMS and NA groups provided marginal evidence in favor of the claim that the learning rate of sham group was higher than lesion group. The findings in DMS and NA groups are further supported by the Bayesian hypothesis test, in which we obtained BF = 3.15 and 7.10, respectively. The BF values are positively in favor of the evidence that the learning rate in the lesion (DMS and NA) groups are lower than their corresponding sham groups.

As depicted in Figure 3. 15, the posterior sample means and their 95% credible intervals (CI) of choice perseveration (β) for the DLS sham and lesion groups were 2.92 (CI = (1.89, 4.26)) and 2.77 (CI = (1.81, 4.02)), respectively. The posterior sample means and their 95% credible intervals (CI) of choice perseveration (β) for the DMS sham and lesion groups were 3.67 (CI = (1.52, 6.55)) and 5.98 (CI = (2.90, 8.86)), respectively. The posterior sample means and their 95% credible intervals (CI) of choice perseveration (β) for the NA sham and lesion groups were 3.55 (CI = (1.28, 6.51)) and 6.21 (CI = (3.19, 8.92)), respectively. Besides, the probability of the posterior distribution of group mean differences of the parameter β between lesion and sham groups for the DLS, DMS and NA groups were 0.424, 0.876, and 0.902, respectively. The findings in DMS and NA groups are further supported by the

Bayesian hypothesis test, in which we obtained BF = 1.56 and 1.69, respectively. The BF values are slightly in favor of the evidence that the choice perseveration in the lesion (DMS and NA) groups are higher than their corresponding sham groups.

Chapter 4: Discussion

1. Result summary

The present study showed that surgery did not alter motivation (i.e., no change in the latency to response to the illuminated apertures and latency to collect reward) in any group of lesioned mice. Compared to sham controls, DMS lesioned mice showed more trials to reach the preset criteria and made more errors during the whole learning process of the dynamic foraging task. In contrast to the DMS group, both NA and DLS lesioned groups did not exhibited more accumulated trials or errors during the whole learning process. In the results of model fitting and matching law analysis, both DMS and NA lesioned mice had a smaller learning rate for updating the RPE signals and a slightly higher choice perseveration compared to sham mice. But no difference was found in their reward sensitivity. Our findings suggest that both DMS and NA are involved in value updating component and decision component of reinforcement learning model in the 2-choice dynamic foraging task.

2. DMS lesion mice showed impaired learning of action-outcome association

Compared to sham controls, DMS lesioned mice showed more cumulated trials and made more errors in this task. Using reinforcement learning model, DMS lesioned mice had a smaller learning rate and higher perseveration compared to sham mice. The results from cumulated trials and trial-by-trial analysis are consistent. In the value updating component, slower rate for updating the RPE signals is indicated by more cumulated trials. Meanwhile, in the decision component, higher choice perseveration is indicated by more perseverative errors.

Furthermore, the deficit observed in the behavioral performance of the DMS lesioned mice is not specific to any reversal section or different difficulty (i.e., the 1:3 reward ratio and the 1:6 reward ratio) within the task. We further divided the overall learning process into 4 sections (the learning of 1:3 reward ratio, the reversal of 1:3 reward ratio, the learning of 1:6 reward ratio, and the reversal of 1:6 reward ratio) to see if the deficit is specific to particular section. Compared to sham controls, no significant difference of behavioral performance in the DMS lesioned mice was found in the four sections. And DMS lesioned mice required more trials and made more errors in both the 1:3 reward ratio and the 1:6 reward ratio compare to sham controls. Thus, results in the DMS lesioned mice could be explained as an impaired learning of action-outcome association.

Goal-directed and habitual actions differ in two ways. Firstly, they differ in the sensitivity to changes in the value of the consequences previously associated with the action. Secondly, they differ in the sensitivity to changes in the causal relationship between the action and those consequences. Therefore, two kinds of experimental test have been used to establish these differences, referred to as outcome devaluation and contingency degradation (B. W. Balleine & O'Doherty, 2010; Yin, Ostlund, et al., 2005). Previous studies used post-learning methods, such as outcome devaluation and extinction test to assess the role of DMS in reinforcement learning and decision making process (Gremel & Costa, 2013; Yin, Knowlton, et al., 2005; Yin & Knowlton, 2004, 2006; Yin, Ostlund, et al., 2005). In the current study, we directly looked into the learning process (i.e. to examine the process of building up an action-outcome association). As a result, bi-directional assessment confirmed that DMS is crucial for the reinforcement learning and decision making process. Based on our current result, it suggests that the DMS is important in both value and choice components.

Recently, instead of a functional segregation, more and more researches showed that DMS- and DLS-mediated learning strategies develop in parallel and compete for the control of the behavioral response early in learning (Ito & Doya, 2011; Moussa, Poucet, Amalric, & Sargolini, 2011; Thorn, Atallah, Howe, & Graybiel, 2010). The DMS is necessary for goal-directed actions, and lesions or inactivation of DMS render actions habitual instead of goal-directed (Yin, Knowlton, & Balleine, 2004).

Conversely, the DLS is necessary for habitual actions, and lesions or temporary inactivation of DLS bias behavior towards goal-directed actions (Yin et al., 2004; Yin, Knowlton, & Balleine, 2006). Moreover, researchers observed region-specific changes in neural activity during the different phases of learning, with the DMS being preferentially engaged early in training and the DLS being engaged later in training (Yin et al., 2009). These previous studies indicate that if function of the DMS is impaired, it could be compensated by function of the DLS. And the DLS may express the functional compensation with the same behavioral outcome but different mechanism inside. In the current study, the DLS may involve more during learning of the 2-choice dynamic foraging task after lesion of the DMS. And it may be the reason why the DMS lesioned mice only showed a tendency of more cumulated trials and errors in learning of the 1:3 reward ratio compared to sham controls.

Through Q-learning model, the DMS lesioned mice showed lower learning rate compared to sham mice. The learning rate is a characteristic of value updating. It implies that the DMS lesioned mice showed slower rate in updating the RPE signals. This could be resulted from changes of reward sensitivity, dysfunction in RPE signal, or the mice simply responded slower to RPE. After surgery, sucrose preference test was done to ensure that reward preference was not altered in the mice with brain lesion. The matching law analysis with data from the 2-choice dynamic task was also conducted. The DMS lesioned mice did not show any significant difference in either case. Accordingly, the possibility of reward sensitivity can be excluded. Since our study did not directly measure the RPE signal in the DMS using electrophysiological recording, it is possible that the DMS lesioned mice may have deficits in the representation of RPE signal, or slower response to it.

3. NA lesion mice only learned slower in more difficult task

Compared to sham controls, NA lesioned mice only showed a tendency of more cumulated trials and errors in the 1:3 reward ratio which is more difficult compared to the 1:6 reward ratio. And specifically, the NA lesioned mice made more errors in the reversal of the 1:3 reward ratio. Using reinforcement learning model, the NA lesioned mice had a smaller learning rate and a slightly higher perseveration compared to sham mice. Compared to DMS lesioned mice, NA lesioned mice only showed behavioral changes in more difficult learning (i.e., the 1:3 reward ratio learning), especially in reversal section of it. The observed deficit in NA lesioned mice appeared to be a failure in suppressing perseverative responding to the original action-outcome contingency. Because the behavioral changes were revealed only in more difficult part of the task, it may indicate that NA participates in harder action-outcome association.

Previous studies demonstrated that the NA plays an important role on the acquisition and reversal of instrumental contingencies (Annett et al., 1989; B. Balleine & Killcross, 1994; Taghzouti et al., 1985), while others found that lesions of NA did not disrupt reversal performance in a go-no go odor discrimination paradigm (Schoenbaum & Setlow, 2003) and in a delayed matching task (Burk & Mair, 2001). Thus, the role of the NA in reversal learning appears to be controversial. On the other hand, it is evident that the NA, especially its core, participates in behavioral flexibility which is related to changes in strategies or rules (Floresco et al., 2006; Haluk & Floresco, 2009). Besides, the NA is considered as having a role in the expression of conditioned emotional responses to cues and contexts associated with appetitive (or aversive) events (Belin et al., 2009; Day & Carelli, 2007). These findings suggest that the NA plays a role in behavioral flexibility. Its functions were also revealed by more perseverative errors in reversal of the 1:3 reward ratio in the dynamic foraging task.

In addition, compared to sham controls, the NA lesioned mice showed lower learning rate in updating RPE signals. As described above, this could be also resulted from change of reward sensitivity, dysfunction in RPE signal, or the mice simply responded slower to RPE signal. As mentioned previously, sucrose preference test and matching law analysis were conducted to ensure that reward sensitivity was not changed in these mice after surgery. Compared to sham controls, the NA lesioned mice did not show any significant difference in either case. Accordingly, alteration of reward sensitivity might be ruled out, and it is possible that the NA lesioned mice may have deficits in the representation of RPE signal, or slower response to it. Nevertheless, the result that no difference in reward sensitivity was found in the NA lesioned mice compared to sham controls seems to contradict with the literature review. Since the NA is implicated in representing predicted future reward (Shiflett & Balleine, 2011), it's somehow inconsistent that we did not find change of reward sensitivity in the NA lesioned mice after excitotoxic lesion.

According to review of Balleine and Shiflett, the NA is implicated in representing predicted future reward, and the representations can be used to guide both goal-directed and habitual actions (Shiflett & Balleine, 2011). Additionally, nucleus accumbens core (NA core) appears to promote a flexible approach toward reward-related locations (Ambroggi, Ishikawa, Fields, & Nicola, 2008; Dalton, Phillips, & Floresco, 2014; Nicola, 2010), whereas nucleus accumbens shell (NA shell) has been implicated in suppression of non-rewarded actions and in learning to ignore irrelevant stimuli (Ambroggi, Ghazizadeh, Nicola, & Fields, 2011; Blaiss & Janak, 2009; Dalton et al., 2014; Floresco, McLaughlin, & Haluk, 2008; Weiner, 2003). Taken together, these results suggest the NA shell and NA core facilitate reward seeking in a distinct yet complementary manner when the relationship between specific actions and reward is uncertain. The NA core promotes approach toward reward-associated stimuli, whereas the NA shell refines response selection to those specific actions more likely to yield reward.

Because the mouse brain is small and it is very challenging to bilaterally inject neurotoxin specific into NA core or NA shell. The coordinates we used here were intended to cover the whole NA, including NA core and NA shell. So the results of the NA lesioned mice could be included both subregions (i.e., NA core and NA shell). In the current study, the NA lesioned mice showed lower learning rate but without change in reward sensitivity. It could be resulted from the complementary effect of NA core and shell. In the current study, mice with lesion of the NA showed no change in reward sensitivity, which might indicate intact functions of normal approach toward reward-associated stimuli. In contrast to that, the functions of response selection to those specific actions with higher reward were affected. As a result, the NA lesioned mice were required to have more trials to reach the preset criteria and made more perseverative errors in reversal of the 1:3 reward ratio.

4. The constraint on Bayesian hierarchical model

In the present study, compared to sham controls, both DMS and NA lesioned mice had a smaller learning rate for updating the RPE signals and a slightly higher choice perseveration compared to sham mice. However, the results of cumulated trials and errors in the two groups are different. The inconsistency between the behavioral data and the parameters from the reinforcement learning model could be explained through the constraint on Bayesian hierarchical model. Because the Bayesian hierarchical model requires the number of input trials to be the same, we cut the cumulated trials into the same number by use of the smallest cumulated trials of mouse as a cutting point in the lesion and sham groups. Maintain the same number of trials in lesion and sham groups ensured the parameters coming from the same criterion. However, the deletion of trials after the cutting point could result in incomplete representation of the parameters. For example, the results of model fitting in the NA lesion and sham groups might be a consequence of cumulated trials in reversal of the 1:6 reward ratio were deleted to fit the requirement of Bayesian hierarchical model in minority of the mice. Thus, it might be possible that the inconsistency between the behavioral data and the model fitting resulted from the constraint on Bayesian hierarchical model.

5. Motivation control of 2-choice dynamic foraging task

In the current study, the NA lesioned mice showed hypolocomotion, whereas a trend of hyperlocomotion was found in the DMS lesioned mice after surgery. These results could confound with the behavioral data of 2-choice dynamic foraging task. It is possible that animals' motivation on performing the task may alter the responses in locomotion. To rule out this possibility, we also recorded and compared the animals' response latency to reach the illuminated apertures and to collect reward before and after surgery. The results showed that they were intact after surgery. Thus, even though mice slightly displayed alterations in locomotion, their motivation on performing the 2-choice dynamic foraging task was not changed during the learning process.

6. Hierarchical reinforcement learning in the cortico-striatal circuits

Using reinforcement learning model, we found similar characteristics in both DMS and NA lesioned mice during the reward learning process, but altered in cumulated trials and errors. The DMS lesioned mice showed impairment in their overall learning despite of task difficulty, whereas the NA lesioned mice only learned slower in more difficult task. These results suggest that there might be a collaborative and hierarchical cortico-striatal circuits as shown in Figure 3. 16 (Ito & Doya, 2011). And the nature of a behavior task may decide the detailed collaboration within striatum as well.

One possible implementation of hierarchical reinforcement learning in the cortico-striatal circuits is the topographically organization within the striatum, in which limbic and ventral prefrontal regions project to the ventral striatum (i.e., NA), sensorimotor cortical regions project to the DLS, and association areas of the prefrontal cortex project to the DMS (Alexander et al., 1986; Groenewegen et al., 1991). The ventral striatum is connected with the limbic system, which represents primary reward information and regulates the affects and motivation of the animal. The DLS, on the contrary, is connected with the sensory-motor cortices that control detailed body movements in response to get reward or avoid punishment. The DMS is connected with the prefrontal cortex that controls more abstract action selection in response to get reward or avoid punishment. Moreover, the connections between the striatum and the dopamine neurons might be used for passing reward signal from area to area (Haruno & Kawato, 2006).

According to the nature of task used in the current study, the DMS is important to the learning of new action-association contingency in the 2-choice dynamic foraging task, and it might use the cue (i.e., the illumination of stimulus-response apertures) in the environment to direct behavior. The NA might involve more when the task gets harder, and it might integrate the context information to affect behavior. Since the current study did not find significant effect in the NA lesioned mice compare to sham controls, increasing task difficulty might be a way to further confirm this hypothesis.

7. Future directions

Based on previous findings and our current results, two potential studies are listed below as future directions.

1. Because our results support the idea of the hierarchical reinforcement learning in the cortico-striatal circuits, it is of great interest to use electrophysiological recording in the DMS and NA (including NA core and NA shell) to see if there are specific change related to certain action or certain step during animals' choice process in the 2-choice dynamic foraging task.
2. As described previously, NA core appears to promote a flexible approach toward reward-related locations, whereas NA shell has been implicated in suppression of non-rewarded actions and in learning to ignore irrelevant stimuli. As a result, specific modulation of striatonigral MSNs and striatopallidal MSNs using optogenetic technique in the NA core and NA shell is worth further exploring. It is of interest to see its effect on decision making. For example, activation of NA core of striatonigral MSNs or striatopallidal MSNs when animal approaches to reward to see if the manipulation disrupts the animals' value representation of reward; activcation of NA shell of striatonigral MSNs or striatopallidal MSNs when animal is going to make perseverative errors to see if the manipulation disrupts the

animals' suppression of non-rewarded actions and in learning to ignore irrelevant stimuli.

References

Albin, R. L., Young, A. B., & Penney, J. B. (1989). The functional anatomy of basal ganglia disorders. *Trends in Neurosciences*, *12*(10), 366–375. doi: 10.1016/0166-2236(89)90074-X

Alexander, G. E., & Crutcher, M. D. (1990). Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends in Neurosciences*, *13*(7), 266–271. doi:10.1016/0166-2236(90)90107-L

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*, 357–381. doi:10.1146/annurev.ne.09.030186.002041

Ambroggi, F., Ghazizadeh, A., Nicola, S. M., & Fields, H. L. (2011). Roles of nucleus accumbens core and shell in incentive-cue responding and behavioral inhibition. *The Journal of Neuroscience*, *31*(18), 6820–6830. doi: 10.1523/JNEUROSCI.6491-10.2011

Ambroggi, F., Ishikawa, A., Fields, H. L., & Nicola, S. M. (2008). Basolateral amygdala neurons facilitate reward-seeking behavior by exciting nucleus accumbens neurons. *Neuron*, *59*(4), 648–661. doi:10.1016/j.neuron.2008.07.004

Annett, L. E., McGregor, A., & Robbins, T. W. (1989). The effects of ibotenic acid lesions of the nucleus accumbens on spatial learning and extinction in the rat. *Behavioural Brain Research*, *31*(3), 231–242. doi: 10.1016/0166-4328(89)90005-3

Asaad, W. F., & Eskandar, E. N. (2011). Encoding of both positive and negative reward prediction errors by neurons of the primate lateral prefrontal cortex and caudate nucleus. *The Journal of Neuroscience*, *31*(49), 17772–17787. doi: 10.1523/JNEUROSCI.3793-11.2011

Balleine, B., & Killcross, S. (1994). Effects of ibotenic acid lesions of the Nucleus Accumbens on instrumental action. *Behavioural Brain Research*, *65*(2), 181–193. doi:10.1016/0166-4328(94)90104-X

Balleine, B. W., & O’Doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*(1), 48–69. doi:10.1038/npp.2009.131

Barraclough, D. J., Conroy, M. L., & Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. *Nature Neuroscience*, *7*(4), 404–410. doi:10.1038/nn1209

Baum, W. M. (1974). On two types of deviation from the matching law: bias and undermatching. *Journal of the Experimental Analysis of Behavior*, *22*(1), 231–242. doi:10.1901/jeab.1974.22-231

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*(1), 129–141. doi:10.1016/j.neuron.2005.05.020

Belin, D., Jonkman, S., Dickinson, A., Robbins, T. W., & Everitt, B. J. (2009). Parallel and interactive learning processes within the basal ganglia: relevance for the understanding of addiction. *Behavioural Brain Research*, *199*(1), 89–102. doi:10.1016/j.bbr.2008.09.027

Belova, M. A, Paton, J. J., & Salzman, C. D. (2008). Moment-to-moment tracking of state value in the amygdala. *The Journal of Neuroscience*, *28*(40), 10023–10030. doi:10.1523/JNEUROSCI.1400-08.2008

Bentivoglio, M., & Morelli, M. (2005). Chapter I The organization and circuits of mesencephalic dopaminergic neurons and the distribution of dopamine receptors in the brain. In S. B. Dunnett, M. Bentivoglio, A. Björklund, & T. Hökfelt (Eds.), *Dopamine* (Vol. 21, pp. 1–107). Elsevier. doi: http://dx.doi.org/10.1016/S0924-8196(05)80005-3

Björklund, A., & Dunnett, S. B. (2007). Dopamine neuron systems in the brain: an update. *Trends in Neurosciences*, *30*(5), 194–202. doi:10.1016/j.tins.2007.03.006

Blaiss, C. A, & Janak, P. H. (2009). The nucleus accumbens core and shell are critical for the expression, but not the consolidation, of Pavlovian conditioned approach. *Behavioural Brain Research*, *200*(1), 22–32. doi:10.1016/j.bbr.2008.12.024

Braun, S., & Hauber, W. (2011). The dorsomedial striatum mediates flexible choice behavior in spatial tasks. *Behavioural Brain Research*, *220*(2), 288–293. doi:10.1016/j.bbr.2011.02.008

Burk, J. A, & Mair, R. G. (2001). Effects of dorsal and ventral striatal lesions on delayed matching trained with retractable levers. *Behavioural Brain Research*, *122*(1), 67–78. doi:10.1016/S0166-4328(01)00169-3

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning*. John Wiley & Sons, Inc.

Cai, X., Kim, S., & Lee, D. (2011). Heterogeneous coding of temporally discounted values in the dorsal and ventral striatum during intertemporal choice. *Neuron*, *69*(1), 170–182. doi:10.1016/j.neuron.2010.11.041

Chang, J.-Y., Chen, L., Luo, F., Shi, L.-H., & Woodward, D. J. (2002). Neuronal responses in the frontal cortico-basal ganglia system during delayed matching-to-sample task: ensemble recording in freely moving rats. *Experimental Brain Research. Experimentelle Hirnforschung. Expérimentation Cérébrale*, *142*(1), 67–80. doi:10.1007/s00221-001-0918-3

Dalton, G. L., Phillips, A. G., & Floresco, S. B. (2014). Preferential involvement by nucleus accumbens shell in mediating probabilistic learning and reversal shifts. *The Journal of Neuroscience*, *34*(13), 4618–4626. doi: 10.1523/JNEUROSCI.5058-13.2014

Day, J. J., & Carelli, R. M. (2007). The nucleus accumbens and Pavlovian reward learning. *The Neuroscientist : A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, *13*(2), 148–159. doi:10.1177/1073858406295854

DeLong, M. R. (1990). Primate models of movement disorders of basal ganglia origin. *Trends in Neurosciences*, *13*(7), 281–285. doi:10.1016/0166-2236(90)90110-V

Ding, L., & Gold, J. I. (2012). Neural correlates of perceptual decision making before, during, and after decision commitment in monkey frontal eye field. *Cerebral Cortex*, *22*(5), 1052–1067. doi:10.1093/cercor/bhr178

Dorris, M. C., & Glimcher, P. W. (2004). Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. *Neuron*, *44*(2), 365–378. doi:10.1016/j.neuron.2004.09.009

Doya, K. (2008). Modulators of decision making. *Nature Neuroscience*, *11*(4), 410–416. doi:10.1038/nn2077

Fallon, J. H., & Moore, R. Y. (1978). Catecholamine innervation of the basal forebrain. IV. Topography of the dopamine projection to the basal forebrain and neostriatum. *The Journal of Comparative Neurology*, *180*(3), 545–580. doi:10.1002/cne.901800310

Fellows, L. K., & Farah, M. J. (2003). Ventromedial frontal cortex mediates affective shifting in humans: evidence from a reversal learning paradigm. *Brain*, *126*(8), 1830–1837. doi:10.1093/brain/awg180

Floresco, S. B., Ghods-Sharifi, S., Vexelman, C., & Magyar, O. (2006). Dissociable roles for the nucleus accumbens core and shell in regulating set shifting. *The Journal of Neuroscience*, *26*(9), 2449–2457. doi: 10.1523/JNEUROSCI.4431-05.2006

Floresco, S. B., McLaughlin, R. J., & Haluk, D. M. (2008). Opposing roles for the nucleus accumbens core and shell in cue-induced reinstatement of food-seeking behavior. *Neuroscience*, *154*(3), 877–884. doi: 10.1016/j.neuroscience.2008.04.004

Fridberg, D. J., Queller, S., Ahn, W.-Y., Kim, W., Bishara, A. J., Busemeyer, J. R., … Stout, J. C. (2010). Cognitive Mechanisms Underlying Risky Decision-Making in Chronic Cannabis Users. *Journal of Mathematical Psychology*, *54*(1), 28–38. doi:10.1016/j.jmp.2009.10.002

Gerfen, C. R. (1992). The neostriatal mosaic: multiple levels of compartmental organization in the basal ganglia. *Annual Review of Neuroscience*, *15*, 285–320. doi:10.1146/annurev.ne.15.030192.001441

Gerfen, C. R., & Scott Young, W. (1988). Distribution of striatonigral and striatopallidal peptidergic neurons in both patch and matrix compartments: an in situ hybridization histochemistry and fluorescent retrograde tracing study. *Brain Research*, *460*(1), 161–167. doi:10.1016/0006-8993(88)91217-6

Giménez-Amaya, J. M., & Graybiel, A. M. (1990). Compartmental origins of the striatopallidal projection in the primate. *Neuroscience*, *34*(1), 111–126. doi:10.1016/0306-4522(90)90306-O

Gremel, C. M., & Costa, R. M. (2013). Orbitofrontal and striatal circuits dynamically encode the shift between goal-directed and habitual actions. *Nature Communications*, *4*. doi:10.1038/ncomms3264

Groenewegen, H. J., Berendse, H. W., Wolters, J. G., & Lohman, A. H. M. (1991). *The Prefrontal Its Structure, Function and Cortex Pathology*. *Progress in Brain Research* (Vol. 85, pp. 95–118). Elsevier. doi:10.1016/S0079-6123(08)62677-1

Haber, S. N., Fudge, J. L., & McFarland, N. R. (2000). Striatonigrostriatal pathways in primates form an ascending spiral from the shell to the dorsolateral striatum. *The Journal of Neuroscience*, *20*(6), 2369–2382.

Haluk, D. M., & Floresco, S. B. (2009). Ventral striatal dopamine modulation of different forms of behavioral flexibility. *Neuropsychopharmacology*, *34*(8), 2041–2052. doi:10.1038/npp.2009.21

Hare, T. A., Camerer, C. F., & Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science*, *324*(5927), 646–648. doi:10.1126/science.1168450

Haruno, M., & Kawato, M. (2006). Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. *Neural Networks*, *19*(8), 1242–1254. doi:10.1016/j.neunet.2006.06.007

Hikosaka, O., Nakahara, H., Rand, M. K., Sakai, K., Lu, X., Nakamura, K., … Doya, K. (1999). Parallel neural networks for learning sequential procedures. *Trends in Neurosciences*, *22*(10), 464–471. doi:10.1016/S0166-2236(99)01439-3

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience, 1*(4), 304–309. doi:10.1038/1124

Hong, S., & Hikosaka, O. (2008). The globus pallidus sends reward-related signals to the lateral habenula. *Neuron, 60*(4), 720–729. doi:10.1016/j.neuron.2008.09.035

Horwitz, G. D., & Newsome, W. T. (2001). Target selection for saccadic eye movements: prelude activity in the superior colliculus during a direction-discrimination task. *Journal of Neurophysiology*, *86*(5), 2543–2558.

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. doi:10.1016/j.conb.2011.04.001

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement Learning : A Survey. *Journal of Artificial Intelligence Research, 4*, 237-285.

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. doi: 10.1080/01621459.1995.10476572

Kawaguchi, Y., Wilson, C. J., Augood, S. J., & Emson, P. C. (1995). Striatal interneurones: chemical, physiological and morphological characterization. *Trends in Neurosciences*, *18*(12), 527–535. doi:10.1016/0166-2236(95)98374-8

Kennerley, S. W., Walton, M. E., Behrens, T. E. J., Buckley, M. J., & Rushworth, M. F. S. (2006). Optimal decision making and the anterior cingulate cortex. *Nature Neuroscience*, *9*(7), 940–947. doi:10.1038/nn1724

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55,* 271–304. doi: 10.1146/annurev.psych.55.090902.142005

Kim, H., Sul, J. H., Huh, N., Lee, D., & Jung, M. W. (2009). Role of striatum in updating values of chosen actions. *The Journal of Neuroscience*, *29*(47), 14701–14712. doi:10.1523/JNEUROSCI.2728-09.2009

Kim, S., Hwang, J., & Lee, D. (2008). Prefrontal coding of temporally discounted values during intertemporal choice. *Neuron*, *59*(1), 161–172. doi: 10.1016/j.neuron.2008.05.010

Kirkby, R. J. (1969). Caudate nucleus lesions and perseverative behavior. *Physiology & Behavior*, *4*(4), 451–454. doi:10.1016/0031-9384(69)90135-8

Kolb, B. (1977). Studies on the caudate-putamen and the dorsomedial thalamic nucleus of the rat: Implications for mammalian frontal-lobe functions. *Physiology & Behavior*, *18*(2), 237–244. doi:10.1016/0031-9384(77)90128-7

Kreitzer, A. C. (2009). Physiology and pharmacology of striatal neurons. *Annual Review of Neuroscience, 32*, 127–147. doi: 10.1146/annurev.neuro.051508.135422

Kreitzer, A. C., & Malenka, R. C. (2007). Endocannabinoid-mediated rescue of striatal LTD and motor deficits in Parkinson’s disease models. *Nature*, *445*(7128), 643–647.

Kreitzer, A. C., & Malenka, R. C. (2008). Striatal plasticity and basal ganglia circuit function. *Neuron*, *60*(4), 543–554. doi:10.1016/j.neuron.2008.11.005

Lau, B., & Glimcher, P. W. (2008). Value representations in the primate striatum during matching behavior. *Neuron*, *58*(3), 451–463. doi:10.1016/j.neuron.2008.02.021

Lauwereyns, J., Watanabe, K., & Coe, B. (2002). A neural correlate of response bias in monkey caudate nucleus, *Nature*, *418*(6896), 413–7. doi: 10.1038/nature00844.1.

Lee, D., Rushworth, M. F. S., Walton, M. E., Watanabe, M., & Sakagami, M. (2007). Functional specialization of the primate frontal cortex during decision making. *The Journal of Neuroscience*, *27*(31), 8170–8173. doi: 10.1523/JNEUROSCI.1561-07.2007

Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience*, *35*, 287–308. doi:10.1146/annurev-neuro-062111-150512

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lindvall, O., & Bjorklund, A. (1974). The organization of the ascending catecholamine neuron systems in the rat brain as revealed by the glyoxylic acid fluorescence method. *Acta Physiologica Scandinavica. Supplementum*, *412*, 1–48.

Lindvall, O., Bjorklund, A., & Divac, I. (1977). Organization of mesencephalic dopamine neurons projecting to neocortex and septum. *Advances in Biochemical Psychopharmacology*, *16*, 39–46.

Lovinger, D. M. (2010). Neurotransmitter roles in synaptic modulation, plasticity and learning in the dorsal striatum. *Neuropharmacology*, *58*(7), 951–961. doi:10.1016/j.neuropharm.2010.01.008

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing, 10*, 325–337. doi:10.1023/A:1008929526011

Matsumoto, M., & Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. *Nature, 447*(7148), 1111–1115. doi:10.1038/nature05860

Mayer, M. L., & Westbrook, G. L. (1987). Cellular mechanisms underlying excitotoxicity. *Trends in Neurosciences*. doi:10.1016/0166-2236(87)90023-3

Montague, P. R., Hyman, S. E., & Cohen, J. D. (2004). Computational roles for dopamine in behavioural control. *Nature, 431*(7010), 760–767. doi: 10.1038/nature03015

Moussa, R., Poucet, B., Amalric, M., & Sargolini, F. (2011). Contributions of dorsal striatal subregions to spatial alternation behavior. *Learning & Memory*, *18*(7), 444–451. doi:10.1101/lm.2123811

Murray, E. a, O'Doherty, J. P., & Schoenbaum, G. (2007). What we know and do not know about the functions of the orbitofrontal cortex after 20 years of cross-species studies. *The Journal of Neuroscience*, *27*(31), 8166–8169. doi:10.1523/JNEUROSCI.1556-07.2007

Nicola, S. M. (2010). The flexible approach hypothesis: unification of effort and cue-responding hypotheses for the role of nucleus accumbens dopamine in the activation of reward-seeking behavior. *The Journal of Neuroscience, 30*(49), 16585–16600. doi:10.1523/JNEUROSCI.3958-10.2010

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology, 53*(3), 139–154. doi:10.1016/j.jmp.2008.12.005

Okano, K., & Tanji, J. (1987). Neuronal activities in the primate motor fields of the agranular frontal cortex preceding visually triggered and self-paced movement. *Experimental Brain Research, 66*(1), 155–166. doi:10.1007/BF00236211

Oyama, K., Hernádi, I., Iijima, T., & Tsutsui, K.-I. (2010). Reward prediction error coding in dorsal striatal neurons. *The Journal of Neuroscience, 30*(34), 11447–11457. doi:10.1523/JNEUROSCI.1719-10.2010

Padoa-Schioppa, C. (2011). Neurobiology of economic choice: a good-based model. *Annual Review of Neuroscience, 34*, 333–359. doi: 10.1146/annurev-neuro-061010-113648

Padoa-Schioppa, C., & Assad, J. a. (2006). Neurons in the orbitofrontal cortex encode economic value. *Nature, 441*(7090), 223–226. doi:10.1038/nature04676

Pastor-Bernier, A., & Cisek, P. (2011). Neural correlates of biased competition in premotor cortex. *The Journal of Neuroscience, 31*(19), 7083–7088. doi: 10.1523/JNEUROSCI.5681-10.2011

Platt, M. L., & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature, 400*(6741), 233–238. doi:10.1038/22268

Pothuizen, H. H. J., Jongen-Rêlo, A. L., Feldon, J., & Yee, B. K. (2005). Double dissociation of the effects of selective nucleus accumbens core and shell lesions on impulsive-choice behaviour and salience learning in rats. *The European*

Journal of Neuroscience, *22*(10), 2605–16. doi: 10.1111/j.1460-9568.2005.04388.x

Quilodran, R., Rothé, M., & Procyk, E. (2008). Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron, 57*(2), 314–325. doi: 10.1016/j.neuron.2007.11.031

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–164.

Ragozzino, M. E. (2007). The contribution of the medial prefrontal cortex, orbitofrontal cortex, and dorsomedial striatum to behavioral flexibility. *Annals of the New York Academy of Sciences*, *1121*, 355–375. doi: 10.1196/annals.1401.013

Ragozzino, M. E., Jih, J., & Tzavos, A. (2002). Involvement of the dorsomedial striatum in behavioral flexibility: role of muscarinic cholinergic receptors. *Brain Research*, *953*(1-2), 205–214. doi:10.1016/S0006-8993(02)03287-0

Ragozzino, M. E., Ragozzino, K. E., Mizumori, S. J. Y., & Kesner, R. P. (2002). Role of the dorsomedial striatum in behavioral flexibility for response and visual cue discrimination learning. *Behavioral Neuroscience*, *116*(1), 105–115. doi: 10.1037//0735-7044.116.1.105

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *The Journal of Neuroscience*, *22*(21), 9475–9489.

Rorie, A. E., Gao, J., McClelland, J. L., & Newsome, W. T. (2010). Integration of sensory and reward information during perceptual decision-making in lateral intraparietal cortex (LIP) of the macaque monkey. *PloS One*, *5*(2), e9308. doi:10.1371/journal.pone.0009308

Rushworth, M. F. S., Walton, M. E., Kennerley, S. W., & Bannerman, D. M. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences*, *8*(9), 410–417. doi:10.1016/j.tics.2004.07.009

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. *The Journal of Neuroscience*, *29*(48), 15104–15114. doi:10.1523/JNEUROSCI.3524-09.2009

Samejima, K., Ueda, Y., Doya, K., & Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, *310*(5752), 1337–1340. doi:10.1126/science.1115270

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

Schoenbaum, G., & Setlow, B. (2003). Lesions of nucleus accumbens disrupt learning about aversive outcomes. *The Journal of Neuroscience*, *23*(30), 9833–9841.

Schultz, W. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. doi:10.1126/science.275.5306.1593

Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology*, *57*, 87–115. doi: 10.1146/annurev.psych.56.091103.070229

Schultz, W. (2010). Dopamine signals for reward value and risk: basic and recent data. *Behavioral and Brain Functions, 6,* 24. doi:10.1186/1744-9081-6-24

Seo, H., Barraclough, D. J., & Lee, D. (2009). Lateral intraparietal cortex and reinforcement learning during a mixed-strategy game. *The Journal of Neuroscience*, *29*(22), 7278–7289. doi:10.1523/JNEUROSCI.1479-09.2009

Seo, H., & Lee, D. (2007). Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. *The Journal of Neuroscience, 27*(31), 8366–8377. doi:10.1523/JNEUROSCI.2369-07.2007

Seo, H., & Lee, D. (2009). Behavioral and neural changes after gains and losses of conditioned reinforcers. *The Journal of Neuroscience*, *29*(11), 3627–3641. doi:10.1523/JNEUROSCI.4726-08.2009

Shen, W., Flajolet, M., Greengard, P., & Surmeier, D. J. (2008). Dichotomous dopaminergic control of striatal synaptic plasticity. *Science*, *321*(5890), 848–851. doi:10.1126/science.1160575

Shidara, M., & Richmond, B. J. (2002). Anterior cingulate: single neuronal signals related to degree of reward expectancy. *Science*, *296*(5573), 1709–1711. doi:10.1126/science.1069504

Shiflett, M. W., & Balleine, B. W. (2011). Molecular substrates of action control in cortico-striatal circuits. *Progress in Neurobiology*, *95*(1), 1–13. doi: 10.1016/j.pneurobio.2011.05.007

Smith, C. A. B. (1961). Consistency in Statistical Inference and Decision. *Journal of the Royal Statistical Society. Series B (Methodological)*, *23*(1), 1–37. doi: 10.2307/2983842

So, N.-Y., & Stuphorn, V. (2010). Supplementary eye field encodes option and action value for saccades with variable reward. *Journal of Neurophysiology*, *104*(5), 2634–2653. doi:10.1152/jn.00430.2010

Sohn, J.-W., & Lee, D. (2007). Order-dependent modulation of directional signals in the supplementary and presupplementary motor areas. *The Journal of Neuroscience*, *27*(50), 13655–13666. doi:10.1523/JNEUROSCI.2982-07.2007

Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, *11*(5), 543–545. doi:10.1038/nn.2112

Sugrue, L. P., Corrado, G. S., & Newsome, W. T. (2004). Matching behavior and the representation of value in the parietal cortex. *Science*, *304*(5678), 1782–1787. doi:10.1126/science.1094765

Sul, J. H., Jo, S., Lee, D., & Jung, M. W. (2011). Role of rodent secondary motor cortex in value-based action selection. *Nature Neuroscience*, *14*(9), 1202–1208. doi:10.1038/nn.2881

Sul, J. H., Kim, H., Huh, N., Lee, D., & Jung, M. W. (2010). Distinct roles of rodent orbitofrontal and medial prefrontal cortex in decision making. *Neuron*, *66*(3), 449–460. doi:10.1016/j.neuron.2010.03.033

Surmeier, D. J., Ding, J., Day, M., Wang, Z., & Shen, W. (2007). D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. *Trends in Neurosciences*, *30*(5), 228–235. doi:10.1016/j.tins.2007.03.008

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of pavlovian reinforcement. In Gabriel, M & Moore, J (Eds.), *Learning and Computational Neuroscience: Foundations of Adaptive Networks* (pp. 497-537). Cambridge, MA: MIT Press.

Sutton, R. S., & Barto, A. G. (1998). Reinforcement Learning: An Introduction. Cambridge, MA: MIT Press.

Taghzouti, K., Louilot, A., Herman, J. P., Le Moal, M., & Simon, H. (1985). Alternation behavior, spatial discrimination, and reversal disturbances following 6-hydroxydopamine lesions in the nucleus accumbens of the rat. *Behavioral and Neural Biology*, *44*(3), 354–363. doi:10.1016/S0163-1047(85)90640-5

Tai, L.-H., Lee, a M., Benavidez, N., Bonci, A., & Wilbrecht, L. (2012). Transient stimulation of distinct subpopulations of striatal neurons mimics changes in action value. *Nature Neuroscience*, *15*(9), 1281–1289. doi:10.1038/nn.3188

Tanaka, S. C., Samejima, K., Okada, G., Ueda, K., Okamoto, Y., Yamawaki, S., & Doya, K. (2006). Brain mechanism of reward prediction under predictable and unpredictable environmental dynamics. *Neural Networks*, *19*(8), 1233–1241. doi:10.1016/j.neunet.2006.05.039

Thorn, C. a, Atallah, H., Howe, M., & Graybiel, A. M. (2010). Differential dynamics of activity changes in dorsolateral and dorsomedial striatal loops during learning. *Neuron*, *66*(5), 781–795. doi:10.1016/j.neuron.2010.04.036

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: a tutorial on the Savage-Dickey method. *Cognitive Psychology*, *60*(3), 158–189. doi:10.1016/j.cogpsych.2009.12.001

Wagner, A. R., & Rescorla, R. A. (1972). Inhibition in Pavlovian conditioning: Application of a theory. In Boakes, R. A. & Halliday, M. S. (Eds.), *Inhibition and Learning* (pp. 301-336). London: Academic press.

Walton, M. E., Behrens, T. E. J., Buckley, M. J., Rudebeck, P. H., & Rushworth, M. F. S. (2010). Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. *Neuron*, *65*(6), 927–939. doi:10.1016/j.neuron.2010.02.027

Watkins, C. J. C. H. (1989). *Learning from delayed rewards.* University of Cambridge.

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3-4), 279–292. doi:10.1007/BF00992698

Weiner, I. (2003). The “two-headed” latent inhibition model of schizophrenia: modeling positive and negative symptoms and their treatment. *Psychopharmacology, 169*(3-4), 257–297. doi:10.1007/s00213-002-1313-x

Wetzels, R., Lee, M. D., & Wagenmakers, E. J. (2010). Bayesian inference using WBDev: A tutorial for social scientists. *Behavior Research Methods,42*(3), 884-897.

Yang, T., & Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature, 447*(7148), 1075–1080. doi:10.1038/nature05852

Yin, H. H. (2010). The sensorimotor striatum is necessary for serial order learning. *The Journal of Neuroscience, 30*(44), 14719–14723. doi: 10.1523/JNEUROSCI.3989-10.2010

Yin, H. H., & Knowlton, B. J. (2004). Contributions of striatal subregions to place and response learning. *Learning & Memory*, *11*(4), 459–463. doi: 10.1101/lm.81004

Yin, H. H., & Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience, 7*(6), 464–476. doi:10.1038/nrn1919

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2004). Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *The European Journal of Neuroscience*, *19*(1), 181–189.

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2005). Blockade of NMDA receptors in the dorsomedial striatum prevents action-outcome learning in instrumental conditioning. *The European Journal of Neuroscience*, *22*(2), 505–512. doi:10.1111/j.1460-9568.2005.04219.x

Yin, H. H., Knowlton, B. J., & Balleine, B. W. (2006). Inactivation of dorsolateral striatum enhances sensitivity to changes in the action-outcome contingency in instrumental conditioning. *Behavioural Brain Research*, *166*(2), 189–196. doi:10.1016/j.bbr.2005.07.012

Yin, H. H., Mulcare, S. P., Hilário, M. R. F., Clouse, E., Holloway, T., Davis, M. I., … Costa, R. M. (2009). Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill. *Nature Neuroscience*, *12*(3), 333–341. doi:10.1038/nn.2261

Yin, H. H., Ostlund, S. B., & Balleine, B. W. (2008). Reward-guided learning beyond dopamine in the nucleus accumbens: the integrative functions of cortico-basal ganglia networks. *The European Journal of Neuroscience*, *28*(8), 1437–1448. doi:10.1111/j.1460-9568.2008.06422.x

Yin, H. H., Ostlund, S. B., Knowlton, B. J., & Balleine, B. W. (2005). The role of the dorsomedial striatum in instrumental conditioning. *The European Journal of Neuroscience*, *22*(2), 513–523. doi:10.1111/j.1460-9568.2005.04218.x

Zahm, D. S. (2000). An integrative neuroanatomical perspective on some subcortical substrates of adaptive responding with emphasis on the nucleus accumbens. *Neuroscience & Biobehavioral Reviews*, *24*(1), 85–105. doi:10.1016/S0149-7634(99)00065-2

Table 2. 1

The grades of evidence corresponding to values of the Bayes factor.

Bayes factor	Evidence
<1	Negative (supports H_0)
1-3	Weak
3-20	Positive
20-150	Strong
>150	Very strong

Note. The table illustrates value of Bayes factor and its corresponding grade of evidence. Adapted from “Bayesian model selection in social research,” by A. E. Raftery, 1995, *Sociological Methodology*, *25*, 111-164.

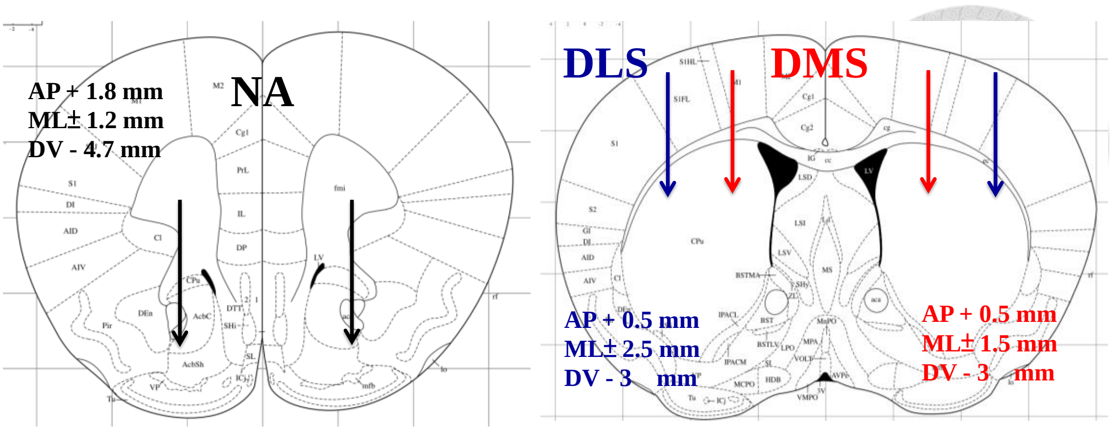
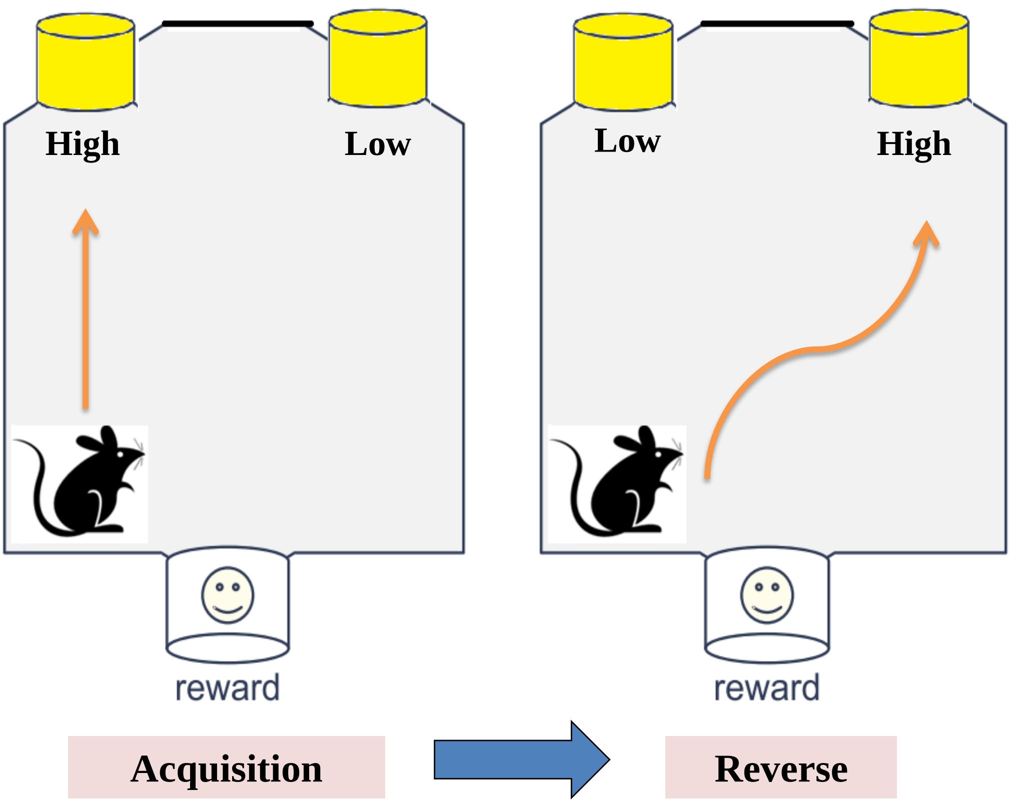


Figure 2. 1. Schematic diagram of drug injection site.

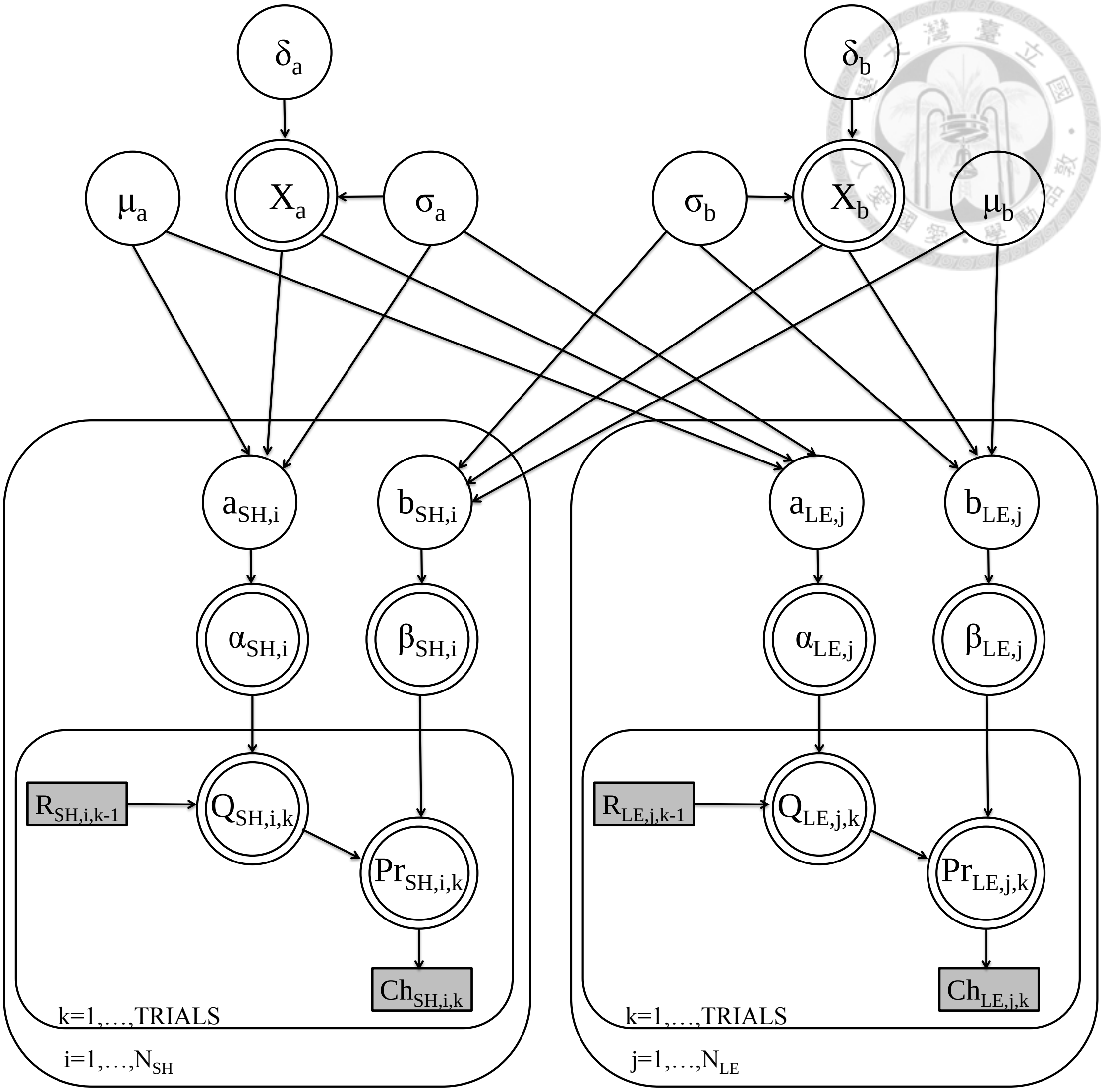
Note. Black bar: NA group; blue bar: DLS group; red bar: DMS group.



Reward ratio	1:3		1:6	
Left Hole	60 %	20 %	68.57 %	11.43 %
Right Hole	20%	60 %	11.43 %	68.57 %

Figure 2. 2. The procedure of the 2-choice dynamic foraging task.

Note. Mice had to nose poke into the food magazine to initiate a trial. A 5 sec intertribal interval (ITI) then preceded the illumination of stimulus-response apertures, and light stimulus was illuminated in the two apertures. Mice were required nose poking into one of the illuminated apertures. Each nose poke into the illuminated aperture was followed by either the delivery of a reward or no any reward, and both of them were subsequently followed by the illumination of magazine. Each trial ended after animals collected earned reward or after animals nose poked into the illuminated magazine. Each mouse discovered these rules and chose the high reward rate aperture by trial and error. The criteria of accomplishing each section was accumulating choice of the high reward rate aperture for at least 70% accuracy in 3 consecutive blocks.

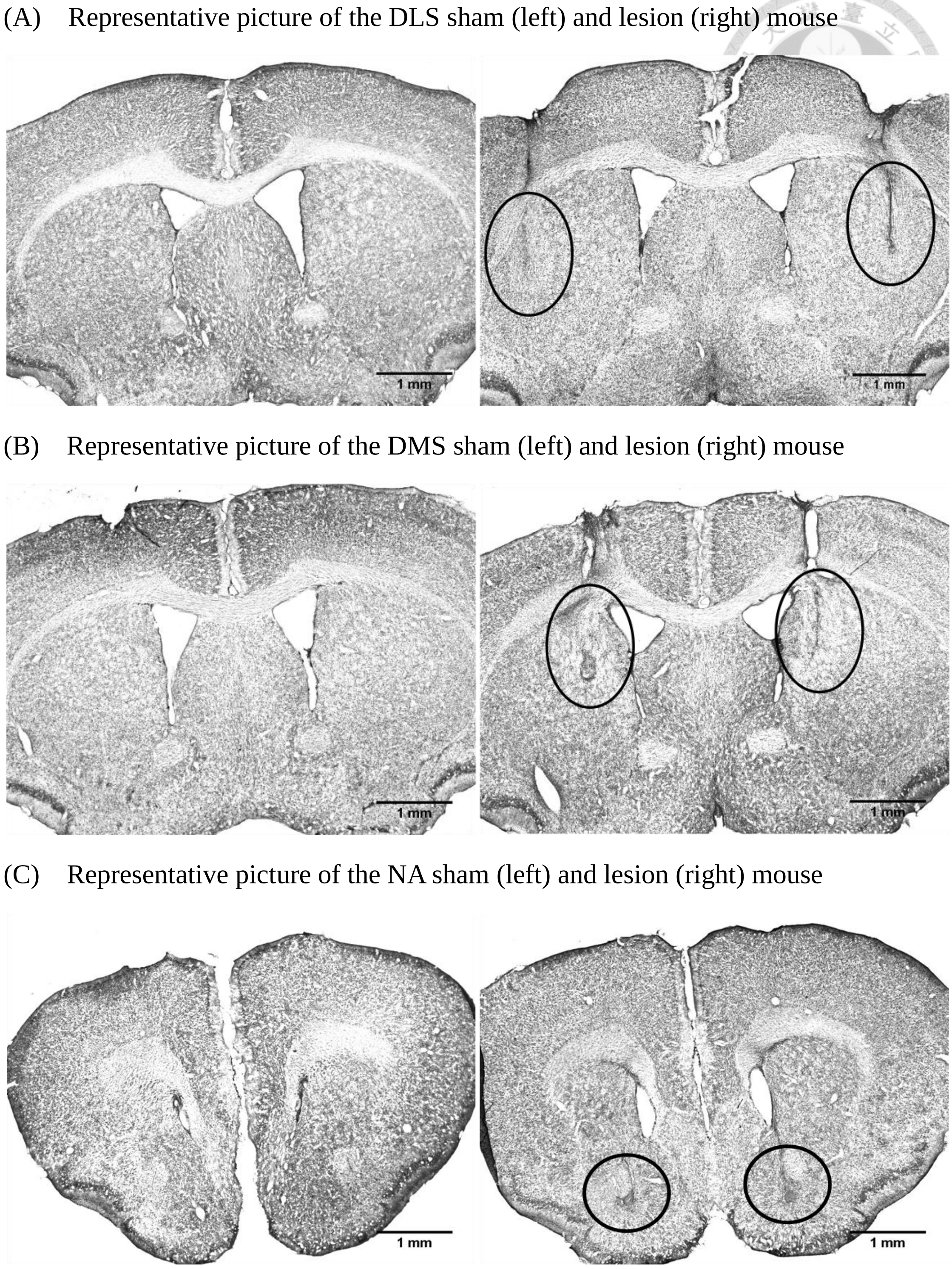


$\alpha_{SH,i} = \Phi(a_{SH,i})$
$\alpha_{LE,i} = \Phi(a_{LE,i})$
$a_{SH,i} \sim \text{Normal } (\mu_a - x_a/2, \sigma^2{}_a)$
$a_{LE,i} \sim \text{Normal } (\mu_a + x_a/2, \sigma^2{}_a)$
$x_a = \delta_a \times \sigma_a$
$\mu_a \sim \text{Normal } (0, 1)$
$\sigma_a \sim \text{Uniform } (0, 10)$
$\delta_a \sim \text{Normal } (0, 1)$
$H0 : \delta_a = 0$
$H1 : \delta_a \neq 0$
$H2 : \delta_a < 0$

$\beta_{SH,i} = \Phi(b_{SH,i})$
$\beta_{LE,i} = \Phi(b_{LE,i})$
$b_{SH,i} \sim \text{Normal } (\mu_b - x_b/2, \sigma^2{}_b)$
$b_{LE,i} \sim \text{Normal } (\mu_b + x_b/2, \sigma^2{}_b)$
$x_b = \delta_b \times \sigma_b$
$\mu_b \sim \text{Normal } (0, 1)$
$\sigma_b \sim \text{Uniform } (0, 10)$
$\delta_b \sim \text{Normal } (0, 1)$
$H0 : \delta_b = 0$
$H1 : \delta_b \neq 0$
$H2 : \delta_b < 0$

Figure 2. 3. Reinforcement learning model fitting using Bayesian Hierachical estimation.

Note. This figure showed the model fitting process and the structure of the model. In this graphical model, nodes are the variables of interest, and the arrows indicate dependencies between the variables. For nodes having double borders mean that the variables are deterministic rather than stochastic. Whereas circular nodes represent continuous variables, square nodes represent discrete variables. Shaded nodes are the observed variables, nodes that are not shaded indicating variables unobserved. $R_{SH,i,k-1}$ indicates the reward sham mouse i received in trial k-1. $R_{LE,j,k-1}$ indicates the reward lesion mouse j received in trial k-1. $CH_{SH,i,k}$ represents the observed choice of sham mouse i in trial k. $CH_{LE,i,k}$ represents the observed choice of lesion mouse j in trial k. $i = 1,\ldots,N_{SH}$ represents the number of sham mice. $j = 1,\ldots,N_{LE}$ represents the number of leion mice. $k = 1,\ldots,$ TRIALS corresponds to the number choice in the 2-choice dynamic foraging task. H0 represents the hypothesis that there is no difference in α or β between the lesion and sham groups. H1 represents the hypothesis that there is significant difference in α or β between the lesion and sham groups. H2 in left part of figure represents the hypothesis that α in sham group is higher than lesion group. H2 in right part of figure represents the hypothesis that β in sham group is lower than lesion group.



(A) Representative picture of the DLS sham (left) and lesion (right) mouse

(B) Representative picture of the DMS sham (left) and lesion (right) mouse

(C) Representative picture of the NA sham (left) and lesion (right) mouse

Figure 3. 1. The pictures of representative infusion placements in the DLS, DMS and NA.

Note. Photographs of representative infusion placements in the DLS, DMS and NA were shown. **Circle areas indicate lesion site.**

(A) Bilateral injection sites of the DLS lesioned mice

(B) Bilateral injection sites of the DMS lesioned mice

(C) Bilateral injection sites of the NA lesioned mice

Figure 3. 2. Schematics of coronal section showing the range of acceptable location of infusions within the striatal subregions.

Note. Lesion sites of striatal subregions were shown. (A) DLS, (B) DMS and (C) NA lesioned group. Circles represented the acceptable lesion sites, whereas triangles represented the excluded animals in every lesion group.

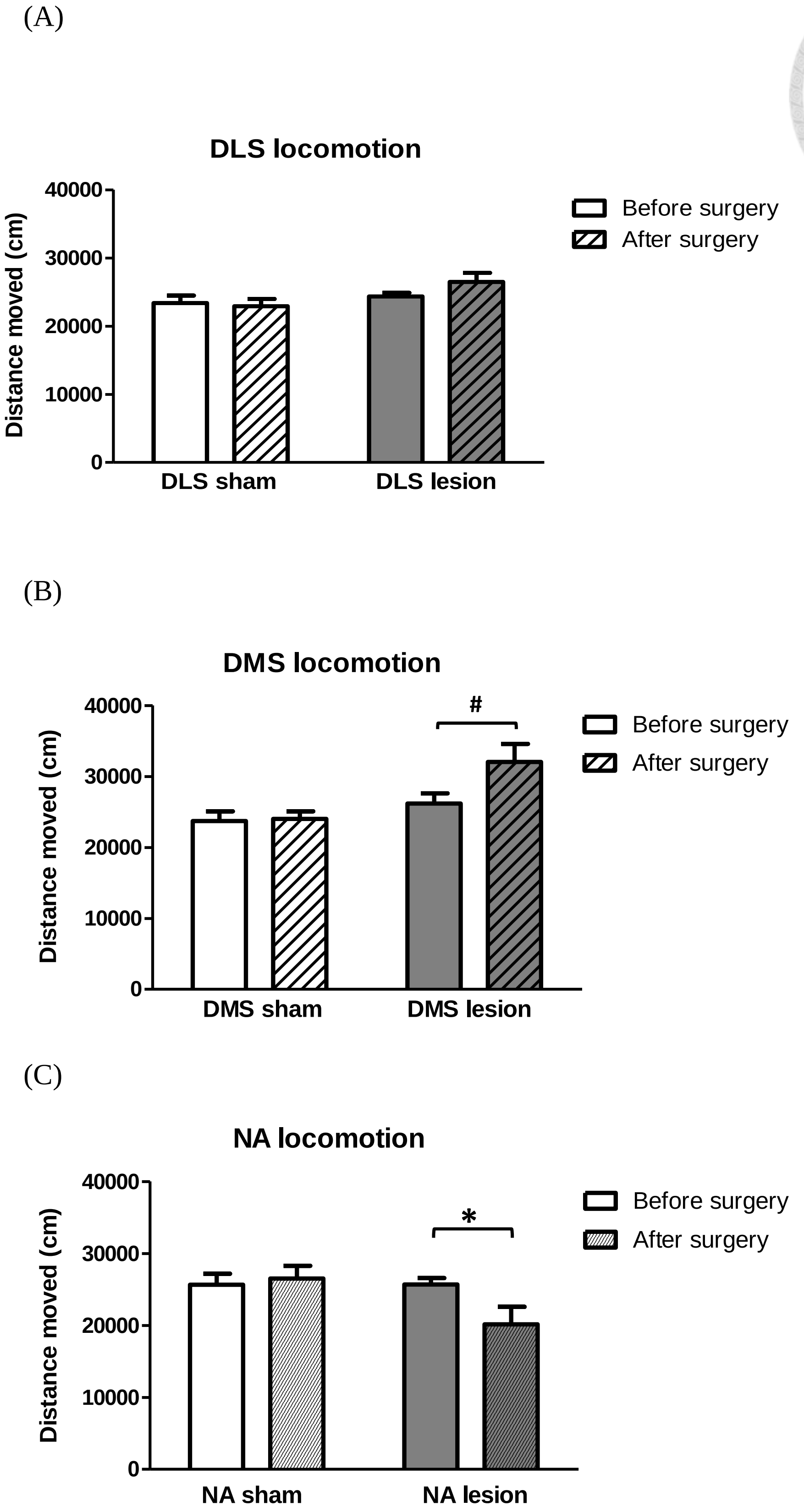


Figure 3. 3. Total moving distance in open field task.

Note. Animals' free moving distance in an open field was recorded. (A) DLS, (B) DMS and (C) NA group. These figures were depicted as mean + SEM. *: p < .05; #: p = .057; white bar: sham group before surgery; white bar with oblique line: shame group after surgery; gray bar: lesion group before surgery; gray bar with oblique line: lesion group after surgery. DMS lesion mice showed a trend of hyperlocomotion after surgery, whereas NA lesion mice showed hypolocomotion after surgery.

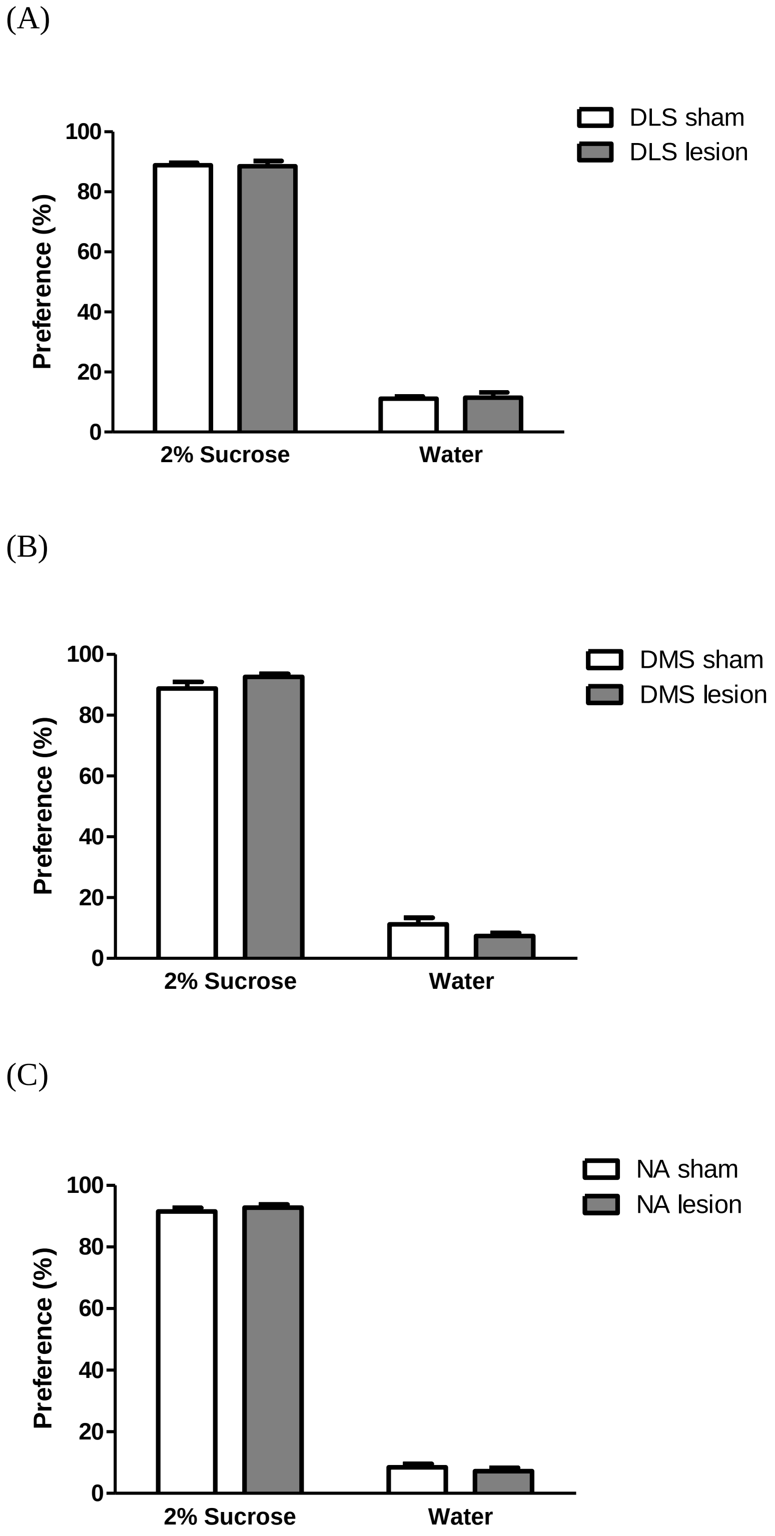


Figure 3. 4. Sucrose preference test.

Note. Animals' sucrose preference was recorded. (A) DLS, (B) DMS and (C) NA group. These figures were depicted as mean + SEM. White bar: sham group; gray bar:

lesion group. There is no difference in sucrose preference between lesion and sham groups in these 3 brain regions.

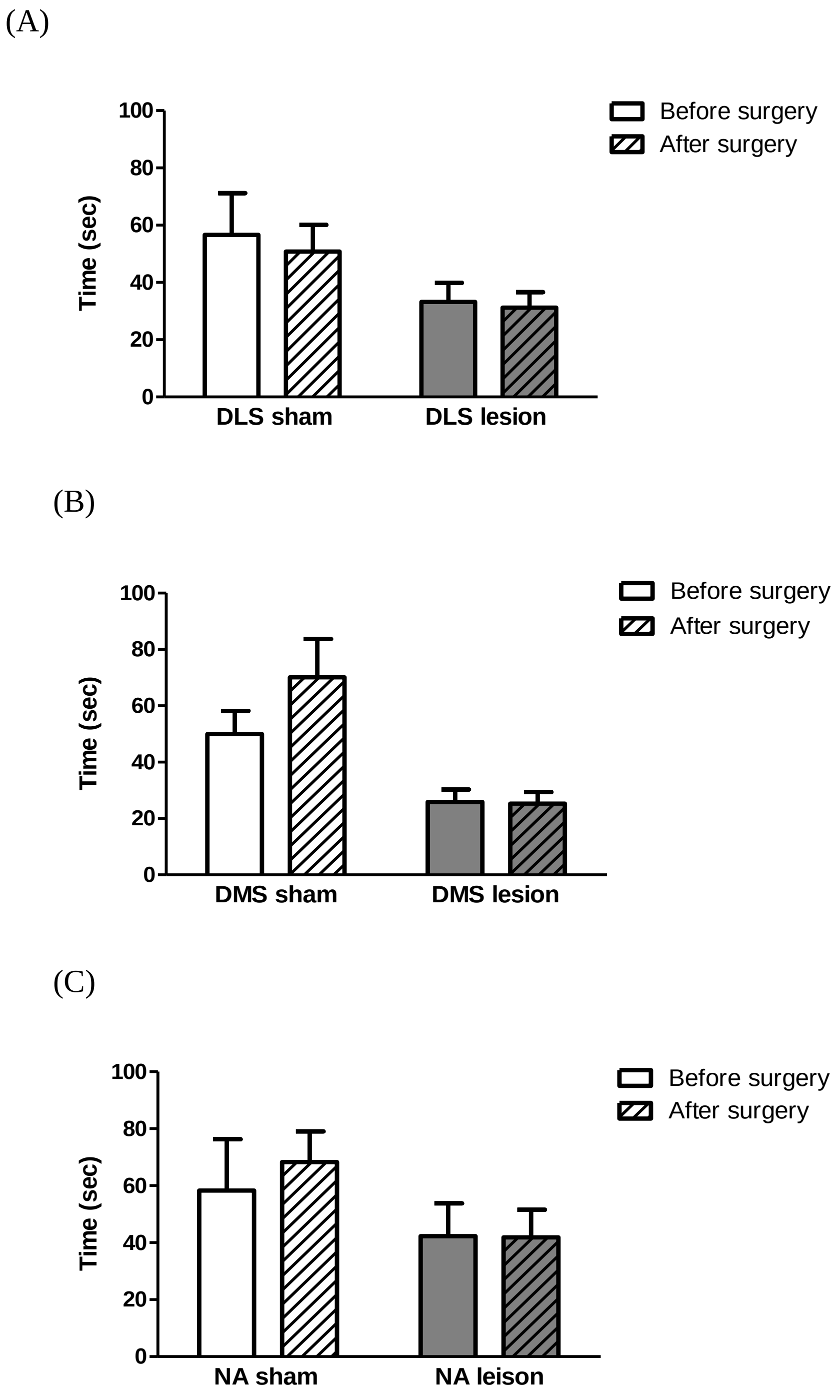


Figure 3. 5. Latency to response to the illuminated apertures in the 2-choice dynamic foraging task.

Note. Animals' latency of nose poke to one of the apertures before and after surgery was recorded. (A) DLS, (B) DMS and (C) NA group. These figures were depicted as mean + SEM. White bar: sham group before surgery; white bar with oblique line: shame group after surgery; gray bar: lesion group before surgery; gray bar with oblique line: lesion group after surgery. In these 3 brain lesion groups, there is no difference in latency to response to the illuminated apertures before and after surgery.

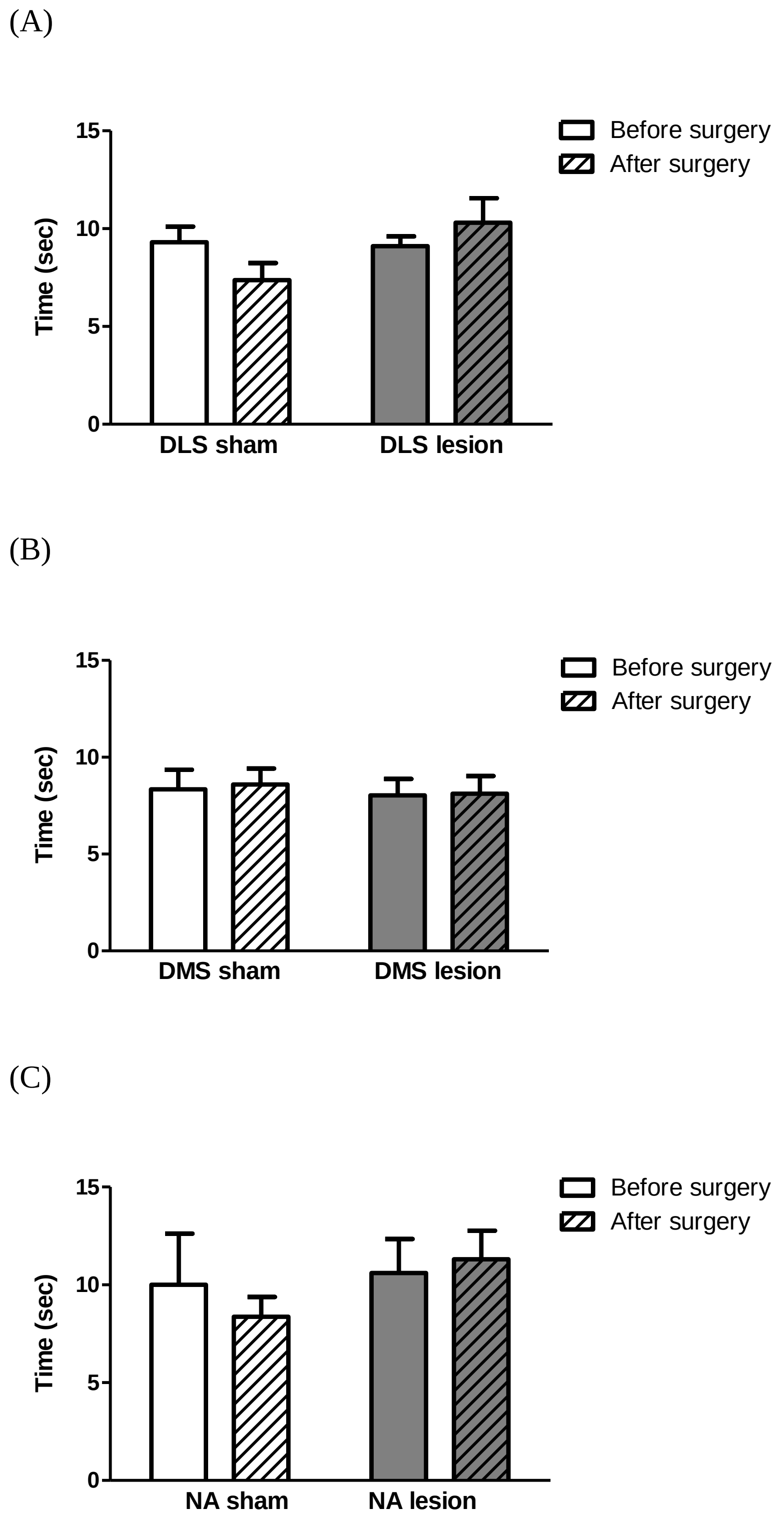


Figure 3. 6. Latency to collect reward in the 2-choice dynamic foraging task.

Note. Animals' latency to magazine to get 2% sucrose solution reward before and after surgery was recorded. (A) DLS, (B) DMS and (C) NA group. These figures were depicted as mean + SEM. White bar: sham group before surgery; white bar with oblique line: shame group after surgery; gray bar: lesion group before surgery; gray bar with oblique line: lesion group after surgery. In these 3 brain lesion groups, there is no difference in latency to collect reward before and after surgery.

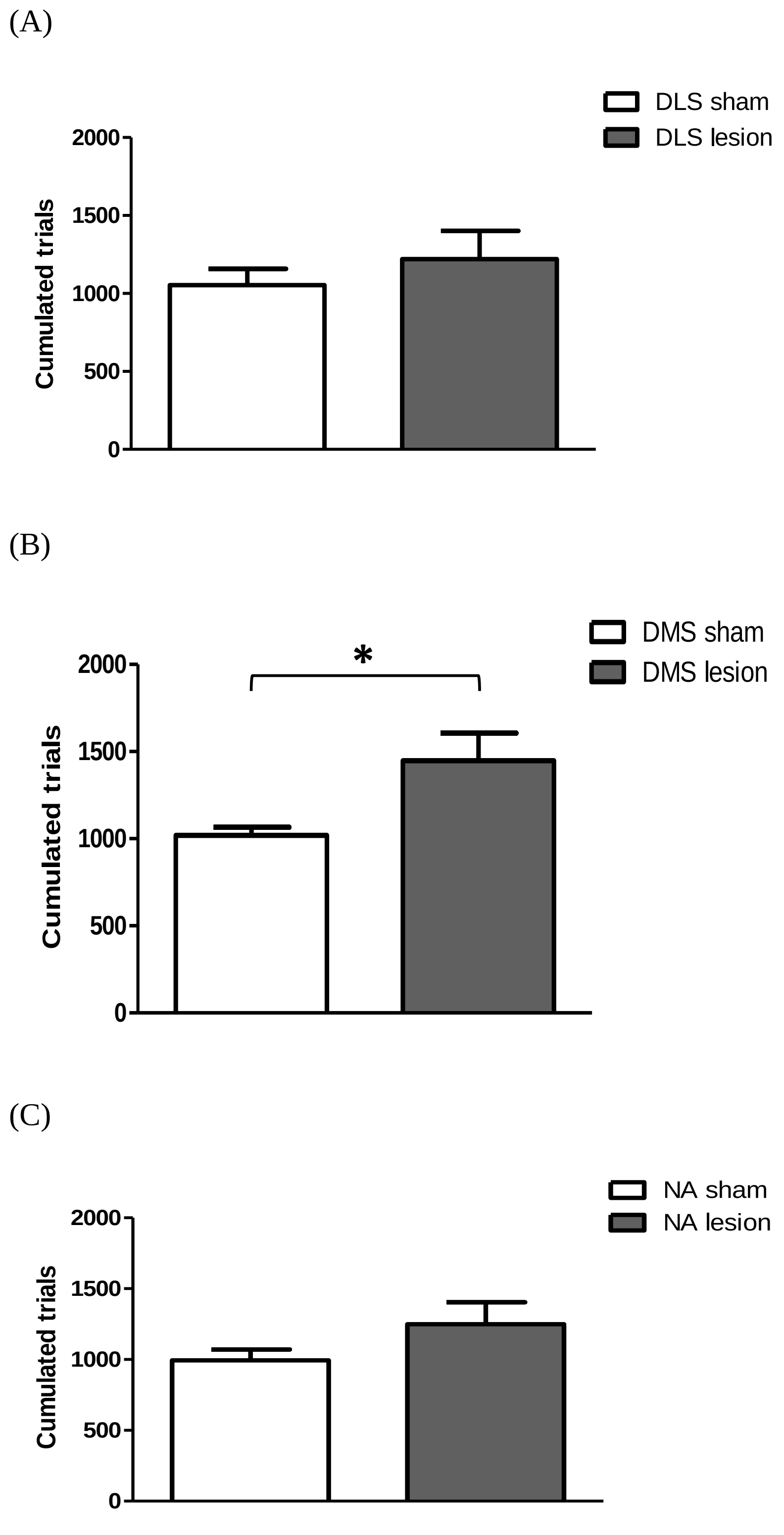


Figure 3. 7. Cumulated trials in overall testing.

Note. Animals' cumulated trials to reach the set criteria in overall testing were recorded. (A) DLS, (B) DMS and (C) NA group. These figures were depicted as mean

+ SEM. White bar: sham group; gray bar: lesion group. For cumulated trials in overall testing, there is no significant difference in DLS and NA group. Compare to sham mice, DMS lesion mice required more trials to reach the criteria in overall testing. * represented $p < .05$.

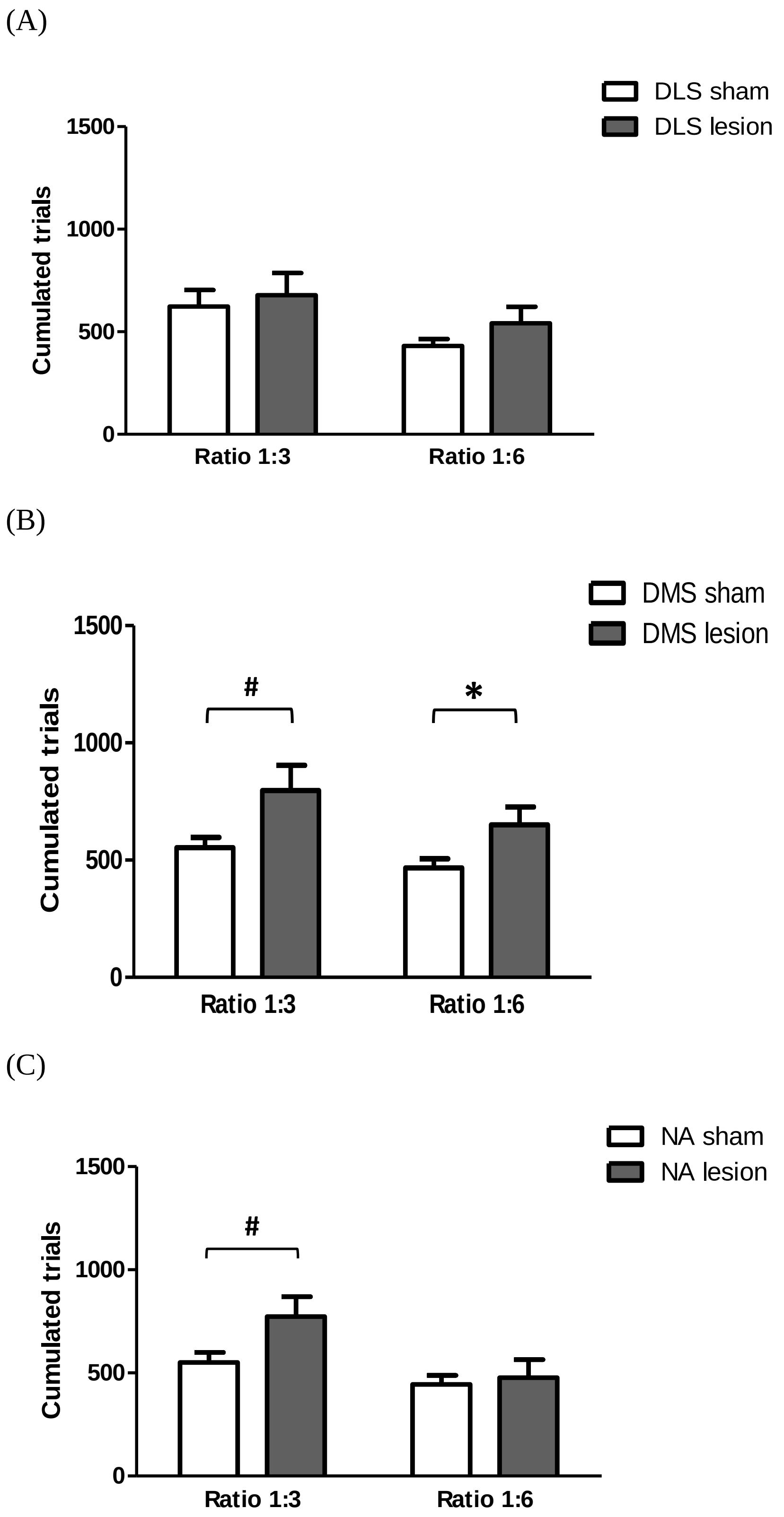


Figure 3. 8. Cumulated trials in 1:3 reward ratio and 1:6 reward ratio learning.

Note. Animals' cumulated trials to reach the set criteria were recorded. (A) DLS, (B) DMS and (C) NA group. These figures were depicted as mean + SEM. White bar: sham group; gray bar: lesion group. "Ratio 1:3" represents 1:3 reward ratio learning and reversal; "Ratio 1:6" represents 1:6 reward ratio learning and reversal. For cumulated trials in 1:3 reward ratio and 1:6 reward ratio, no significant difference was found in DLS group. Compare to sham mice, DMS lesion mice required more trials to reach the criteria in 1:6 reward ratio, with a marginal effect in 1:3 reward ratio. There is a trend that NA lesion mice needed more trials to reach the criteria in 1:3 reward ratio compare to sham mice. * represented $p < .05$, # in DMS group represented $p = .059$, and # in NA group represented p = .06.

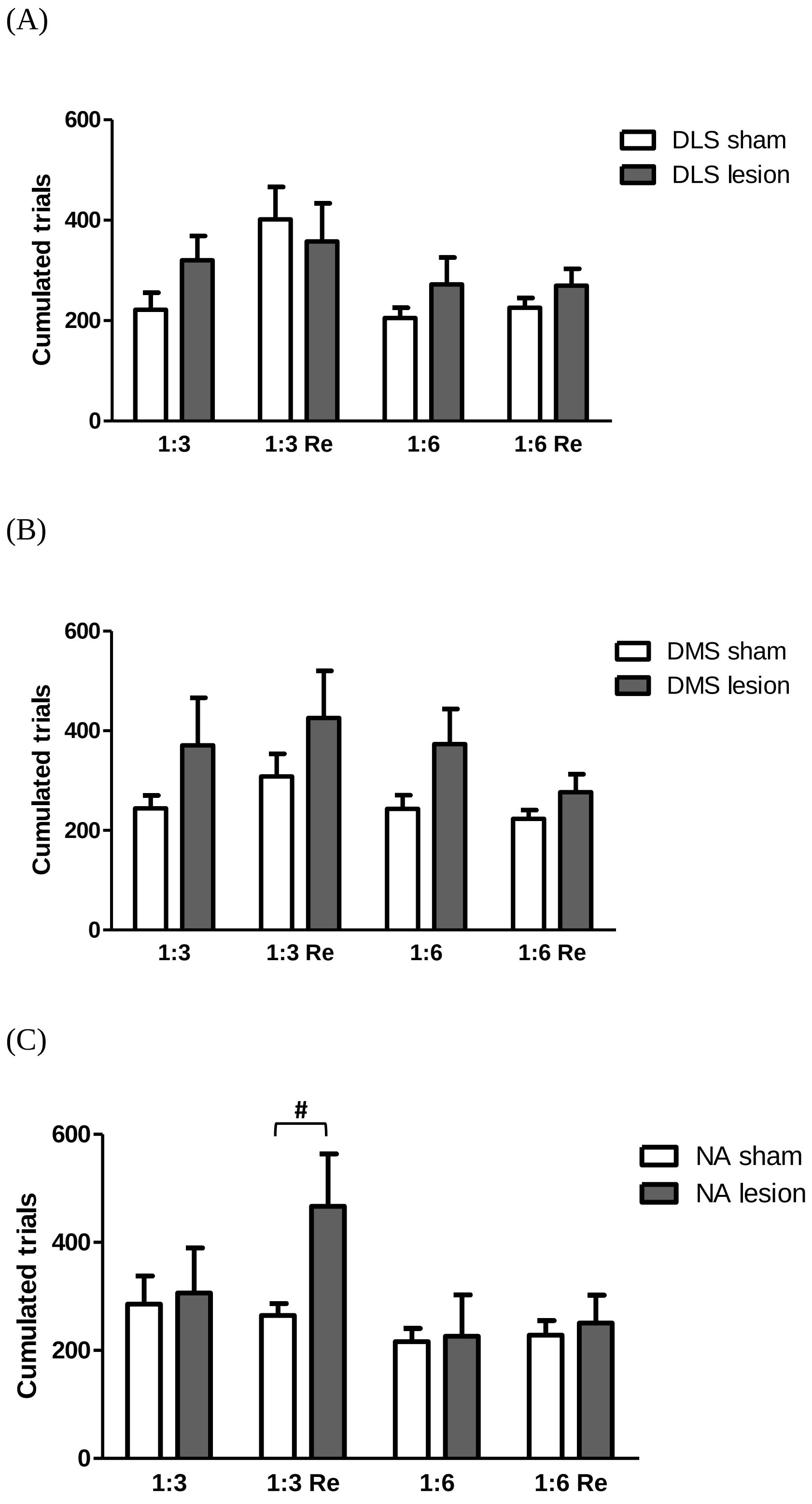


Figure 3. 9. Cumulated trials in each section of the 2-choice dynamic foraging task.

Note. Animals' cumulated trials to reach the set criteria were recorded. (A) DLS, (B) DMS and (C) NA group. These figures were depicted as mean + SEM. White bar: sham group; gray bar: lesion group. "1:3" represents 1:3 reward ratio learning; "1:3 Re" represents reversal of 1:3 reward ratio; "1:6" represents 1:6 reward ratio learning; "1:6 Re" represents reversal of 1:6 reward ratio. The NA lesion mice seemed to need more trials to reach the criteria in the reversal of 1:3 reward ratio compare to sham mice. # represented *p* = .069.

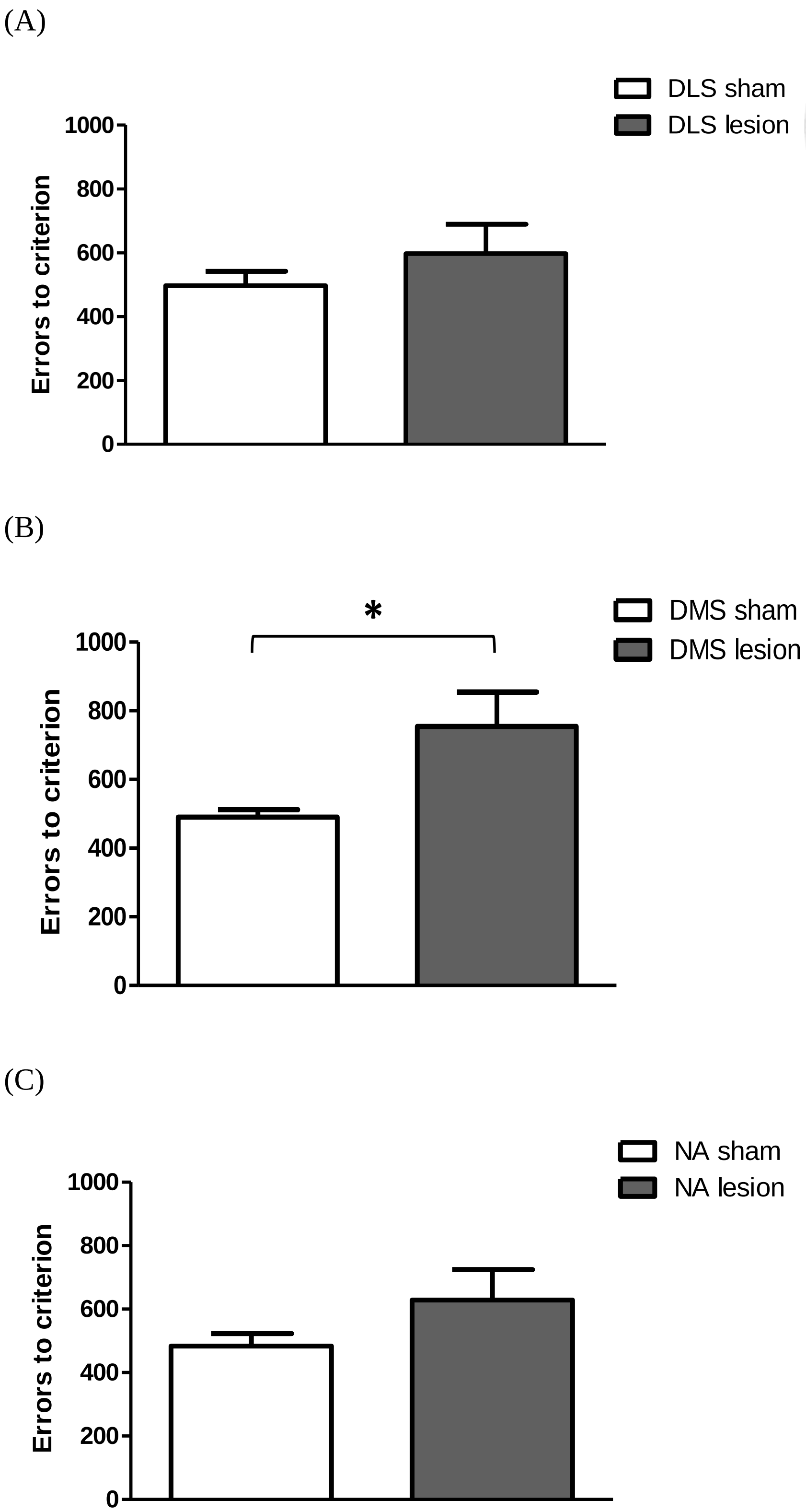


Figure 3. 10. Cumulated errors in overall testing.

Note. Animals' cumulated errors to reach the set criteria were recorded. (A) DLS, (B) DMS and (C) NA group. These figures were depicted as mean + SEM. White bar: sham group; gray bar: lesion group. For cumulated errors in overall testing, no significant difference was found in DLS and NA group. Compare to sham mice, DMS lesion mice cumulated more errors to reach the criteria in overall testing. * represented $p < .05$.

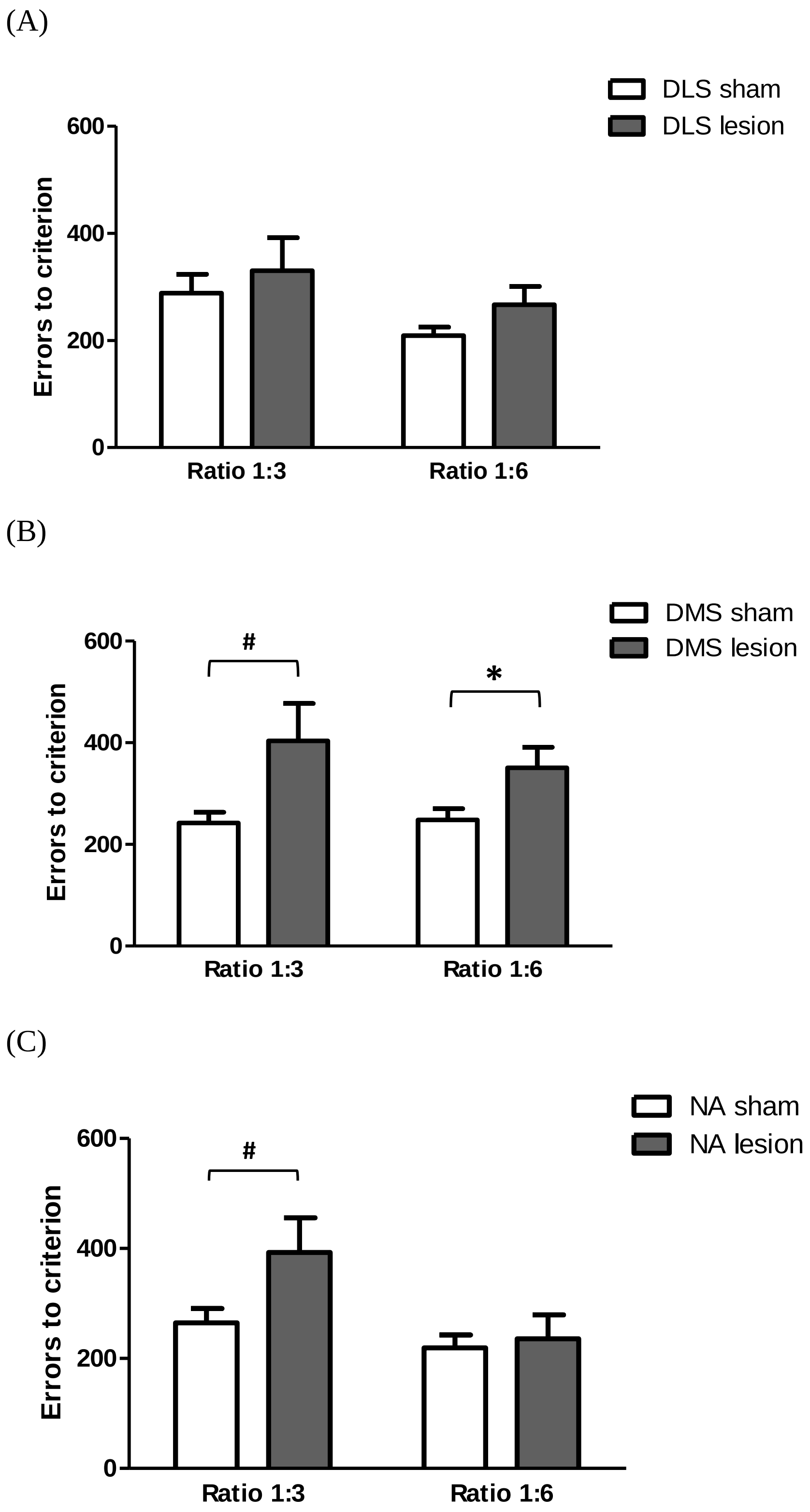


Figure 3. 11. Cumulated errors in 1:3 reward ratio and 1:6 reward ratio learning.

Note. Animals' cumulated errors to reach the set criteria were recorded. (A) DLS, (B) DMS and (C) NA group. These figures were depicted as mean + SEM. White bar: sham group; gray bar: lesion group. "Ratio 1:3" represents 1:3 reward ratio learning and reversal; "Ratio 1:6" represents 1:6 reward ratio learning and reversal. For cumulated errors in 1:3 reward ratio and 1:6 reward ratio, no significant difference was found in DLS group. Compare to sham mice, DMS lesion mice cumulated more errors to reach the criteria in 1:6 reward ratio, with a marginal effect in 1:3 reward ratio. There is a trend that NA lesion mice cumulated more errors to reach the criteria in 1:3 reward ratio compare to sham mice. * represented $p < .05$, # in DMS group represented $p = .06$, and # in NA group represented p = .086.

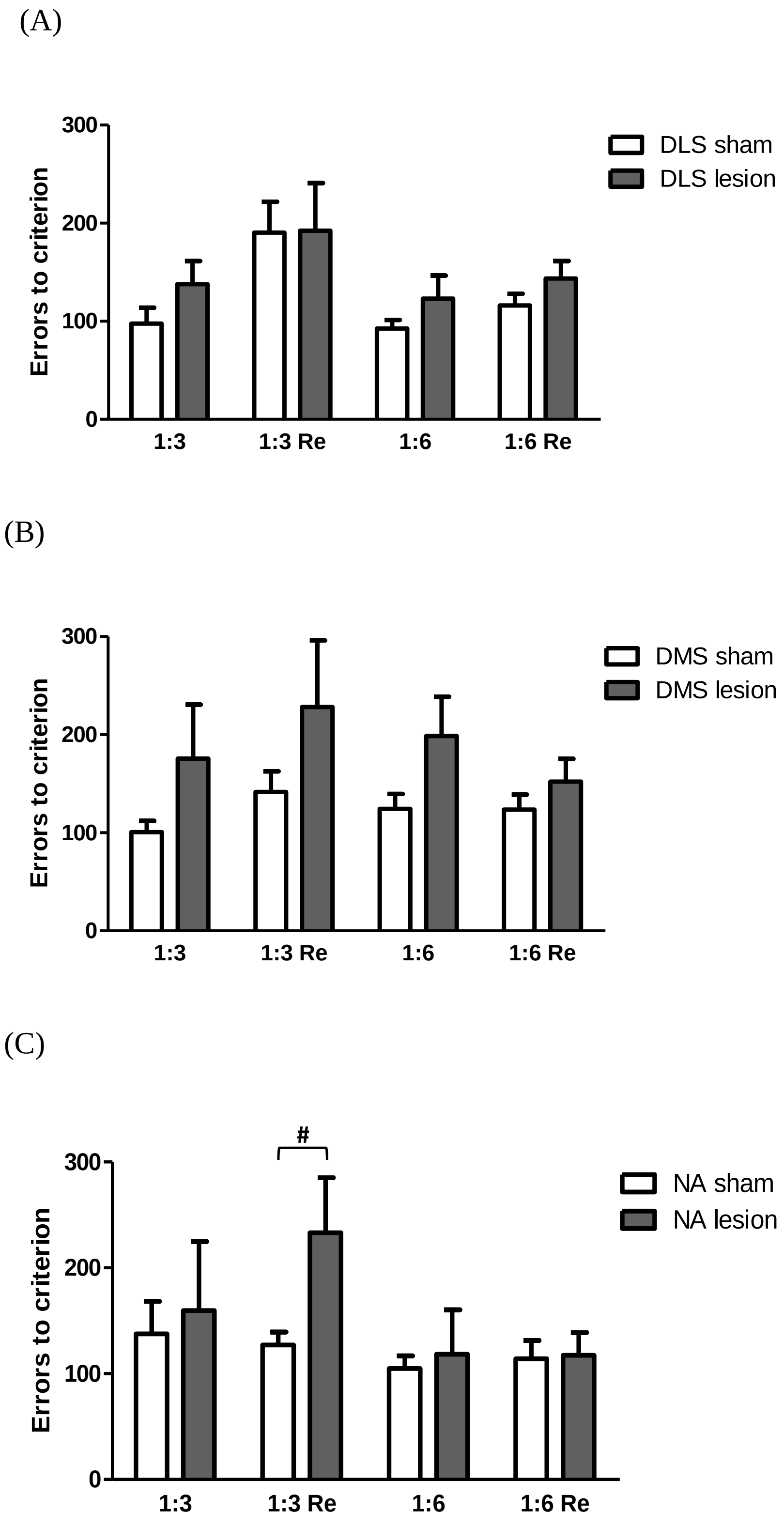


Figure 3. 12. Errors in each section of the 2-choice dynamic foraging task.

Note. Animals' cumulated errors to reach the set criteria were recorded. (A) DLS, (B) DMS and (C) NA group. These figures were depicted as mean + SEM. White bar: sham group; gray bar: lesion group. "1:3" represents 1:3 reward ratio learning; "1:3 Re" represents reversal of 1:3 reward ratio; "1:6" represents 1:6 reward ratio learning; "1:6 Re" represents reversal of 1:6 reward ratio. There is a trend that NA lesion mice cumulated more errors to reach the criteria compare to sham mice. # represented *p* = .077.

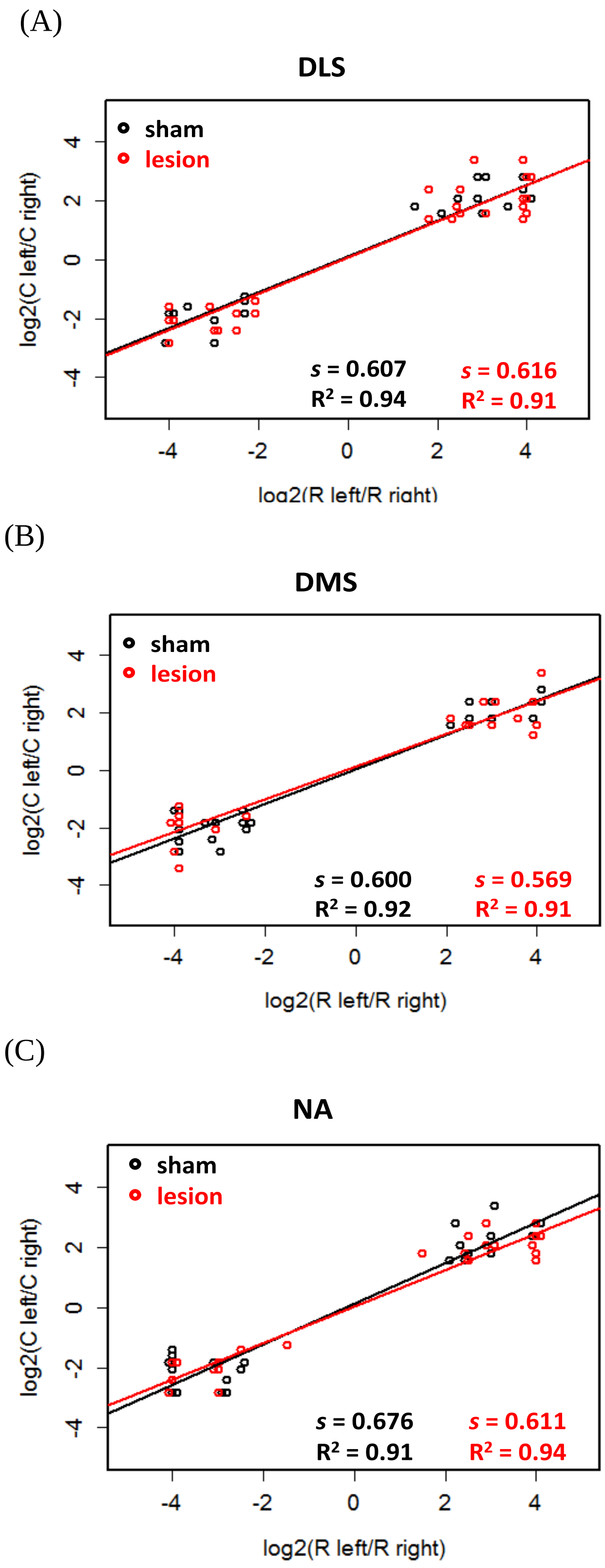
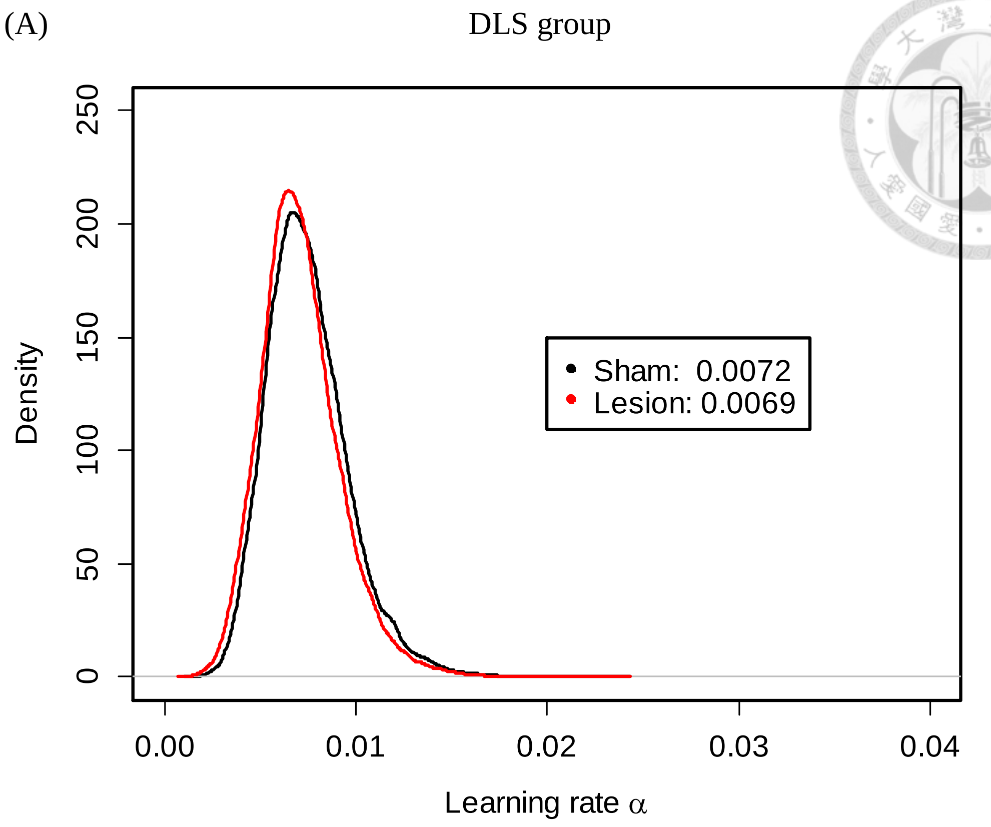
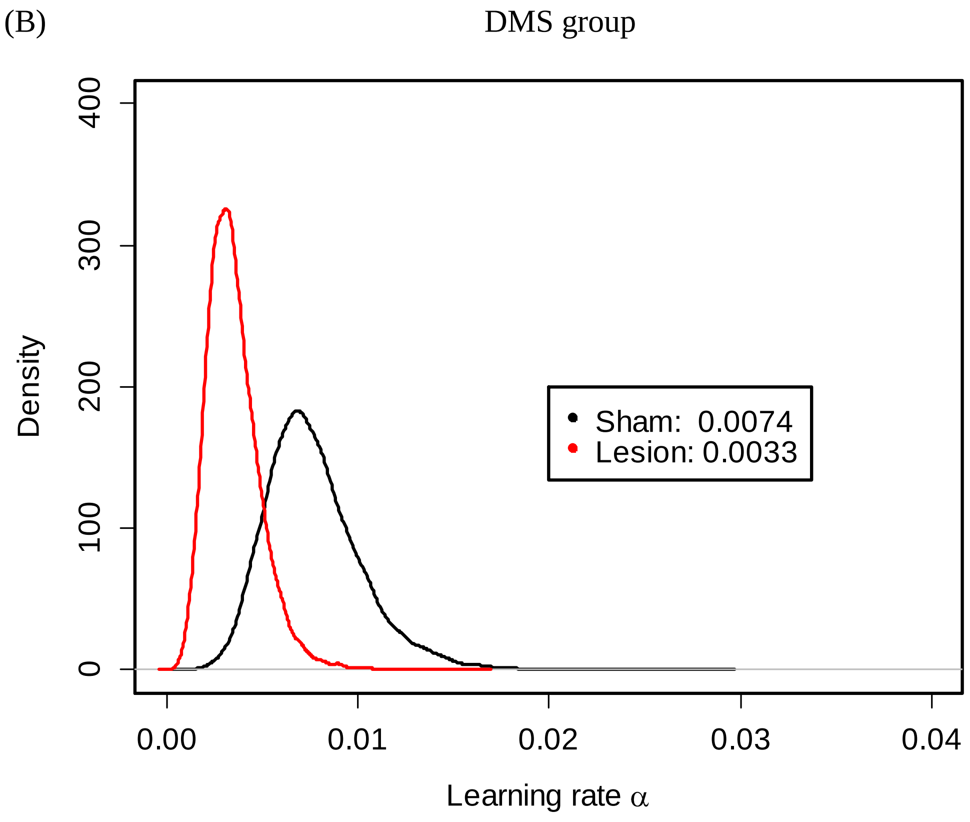


Figure 3. 13. Steady state choice behavior of all lesion and sham groups.

Note. Log choice ratios are plotted as a function of log reward ratio. (A) DLS, (B) DMS, and (C) NA group. The slope represented the reward sensitivity. Steady state choice behaviors of all sham and lesion groups obey the matching law. There is no difference in reward sensitivity between lesion and sham groups.


(A)
DLS group
Density
0
50
100
150
200
250
0.00
0.01
0.02
0.03
0.04
Learning rate α
● Sham: 0.0072
● Lesion: 0.0069


(B)
DMS group
Density
0
100
200
300
400
0.00
0.01
0.02
0.03
0.04
Learning rate α
● Sham: 0.0074
● Lesion: 0.0033

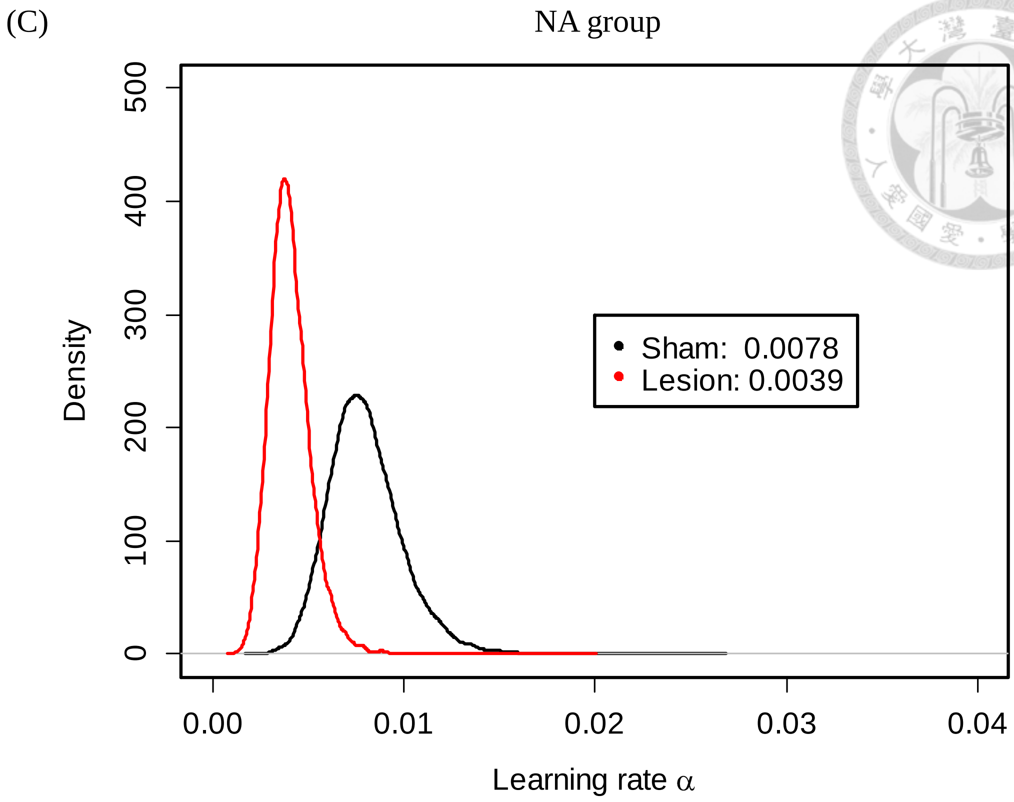
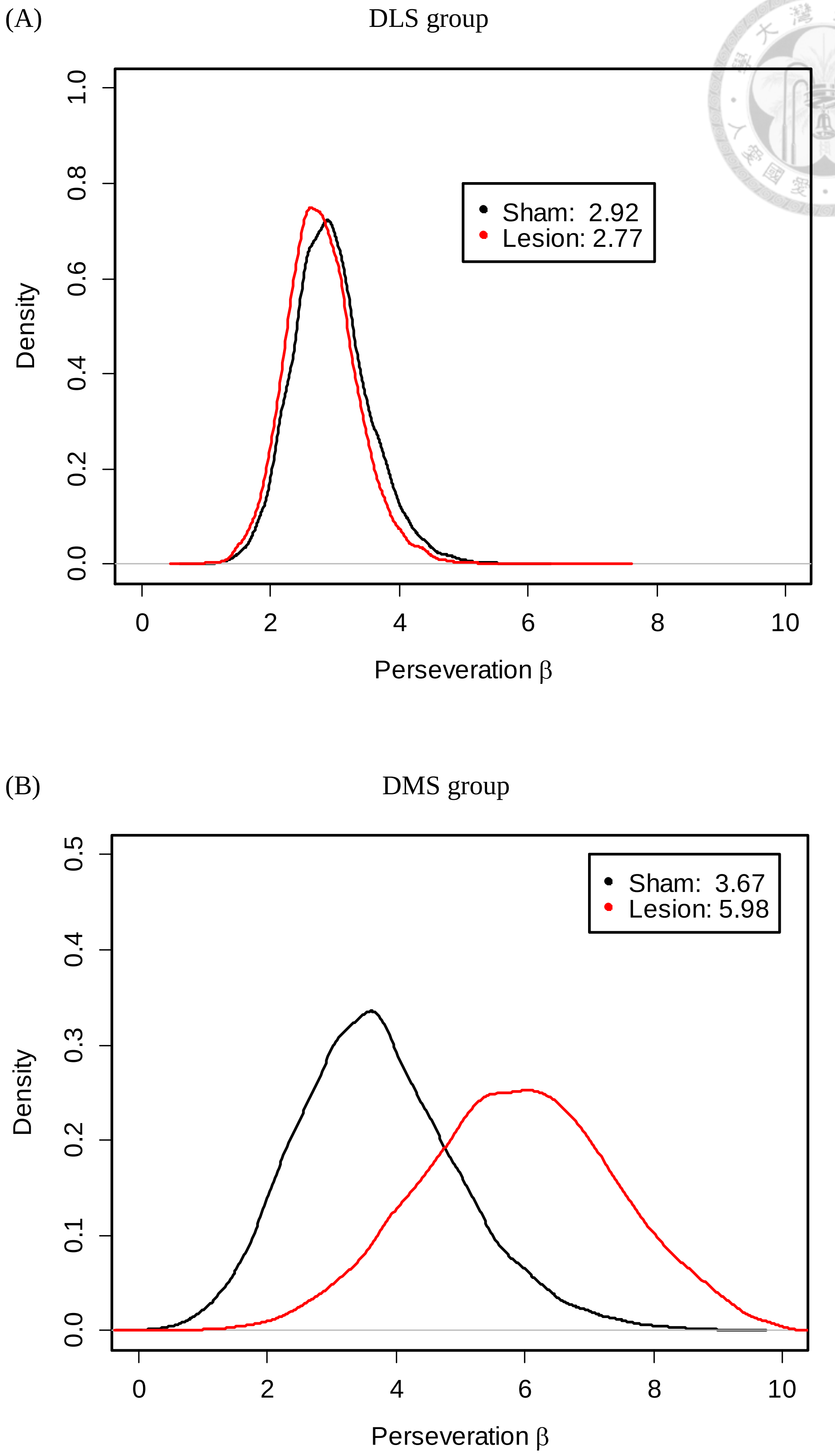


Figure 3. 14. The model fitting results of learning rate α.

Note. The learning rate of each group was presented. (A) DLS, (B) DMS, and (C) NA group. The posterior distribution of group mean differences of the parameter α between sham and lesion groups (DLS, DMS and NA, respectively) showed a 0.556 (0.957 and 0.978, respectively) probability of being greater than zero. Result of both DMS (b) and NA (c) groups provided marginal evidence favoring the claim that the learning rate of sham group was higher than lesion group. This conclusion of DMS and NA groups are also supported by the Bayesian hypothesis test; we obtained BF = 3.15 (BF = 7.10, respectively), positively in favor of the evidence that the learning rate in the lesion (DMS and NA) groups are lower than their corresponding sham groups.


(A)
DLS group
Sham: 2.92
Lesion: 2.77
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
Perseveration β
(B)
DMS group
Sham: 3.67
Lesion: 5.98
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
Perseveration β

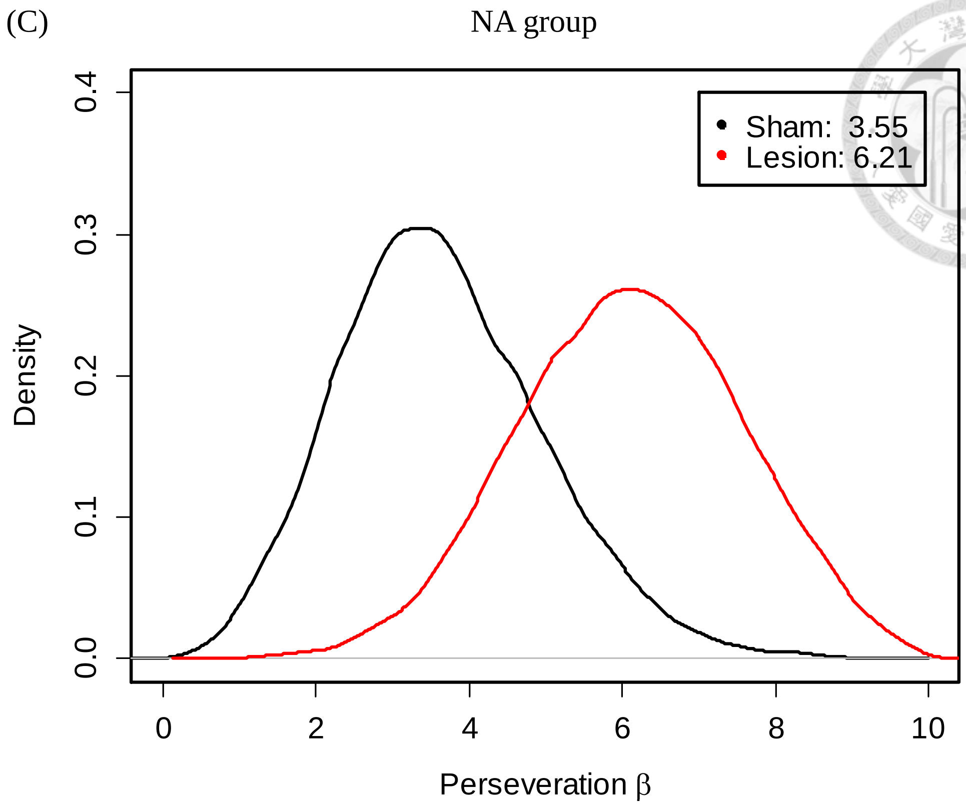


Figure 3. 15. The model fitting results of choice perseveration β.

Note. The choice perseveration of each group was presented. (A) DLS, (B) DMS, and (C) NA group. The posterior distribution of group mean differences of the parameter β between lesion and sham groups (DLS, DMS and NA, respectively) showed a 0.424 (0.876 and 0.902, respectively) probability of being greater than zero. The Bayes factor for testing the hypothesis that choice perseveration is lower in the corresponding sham groups than in the lesion groups (DLS, DMS and NA) showed BF = 0.38 (BF = 1.56 and BF = 1.69, respectively), slightly in favor of the evidence that the choice perseveration in the sham groups of DMS and NA are lower than their corresponding lesion groups.

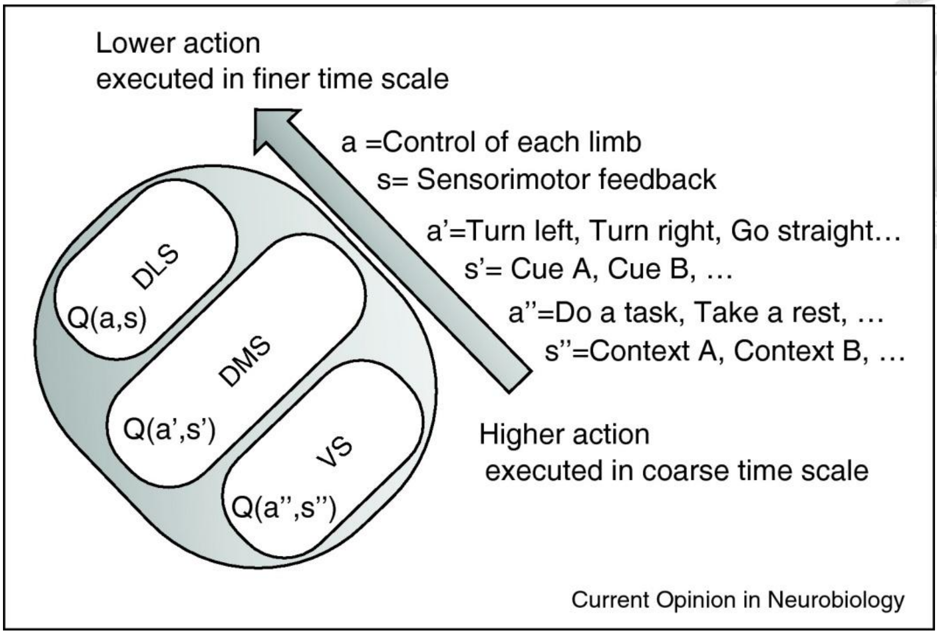

Lower action
executed in finer time scale
a =Control of each limb
s= Sensorimotor feedback
a'=Turn left, Turn right, Go straight…
s'= Cue A, Cue B, …
a''=Do a task, Take a rest, …
s''=Context A, Context B, …
DLS
Q(a,s)
DMS
Q(a',s')
VS
Q(a'',s'')
Higher action
executed in coarse time scale
Current Opinion in Neurobiology

Figure 3. 16. The Hierarchical reinforcement learning in the cortico-striatal loops.

Note. This figure represents a working hypothesis that DLS, DMS, and the ventral striatum (VS, i.e. NS) are parallel and hierarchical Q-learning modules that are in charge of actions at different physical and temporal scales. Adapted from “Multiple representations and algorithms for reinforcement learning in the cortico-basal ganglia circuit,” by M. Ito, K. Doya, 2011, *Current opinion in neurobiology*, 21, 368-73.