

Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Heejae Choi

March 23, 2017

Investigating Associations between Model-based Reinforcement Learning and Model-based
Navigation

by

Heejae Choi

Michael Treadway
Adviser

Department of Psychology

Michael Treadway
Adviser

Phillip Wolff
Committee Member

Samiran Banerjee
Committee Member

2017

Investigating Associations between Model-based Reinforcement Learning and Model-based
Navigation

By

Heejae Choi

Michael Treadway

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Department of Psychology

2017

Abstract

Investigating Associations between Model-based Reinforcement Learning and Model-based Navigation By Heejae Choi

Model-based and model-free reinforcement learning and boundary-based and landmark-based learning are conceptually similar in that model-based and boundary-based systems pay attention to the overall structure and environment, while model-free and landmark-based systems focus on a reward or landmark when making a decision. The brain regions that are activated by the two reinforcement learning systems are also in parallel with the two spatial learning systems. Model-based learning involves prefrontal cortices and hippocampi, which are also activated by boundary-based learning. Model-free learning induces activity in the dorsolateral striatum, ventral striatal projections and putamen activities, while landmark-based learning induces activity in the dorsolateral striatum. In the current study, we examined the behavioral correlation between model-based/model-free reinforcement learning and boundary/landmark based spatial learning, in order to investigate whether or not there is a domain general cognitive system that supports both model-based/boundary-based learning and model-free/landmark-based learning. Model-based and model-free learning was assessed with the two-stage decision task, and boundary and landmark-based learning was assessed with the boundary-landmark task. We tested 26 participants, and no significant correlation was found between model-based decision-making characteristics and boundary-based spatial learning. There was no significant correlation between model-free decision-making characteristics and landmark-based spatial learning. However, model-free learning indicators showed negative correlation with the average error rate in the boundary-landmark task, and the model-based indicator also showed a positive correlation with the average error rate in the boundary-landmark task. This indicates that increased reliance on the model-free decision making was associated with better performance on the spatial learning task.

Keywords: Model-based learning, Model-free learning, Boundary-based learning, Landmark-based learning

Investigating Associations between Model-based Reinforcement Learning and Model-based
Navigation

By

Heejae Choi

Michael Treadway

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Arts with Honors

Department of Psychology

2017

Acknowledgements

This honors thesis would not have been possible without the help and support from Translational Research in Affective Disorders laboratory at Emory University. I wish to express my sincere thanks to Dr. Michael Treadway, the principal investigator of the laboratory, for providing me an opportunity to implement the study, all the necessary facilities for the research, and thorough comments that greatly improved the manuscript. I also want to thank Dr. Jessica Cooper, post-doctoral fellow of the laboratory, for helping me with every step of the research including statistical analysis and interpretations. I also thank her for her comments on an earlier version of the manuscript which provided me a great help figuring out the direction of the paper. I would also like to appreciate Chelsea Leonard, the research specialist, and Dan Cole, the technician, for helping me recruiting the participants and implementing the experiments.

I would also like to show gratitude to Dr. Daniel Dilks and Frederik Kamp in Dilks laboratory at Emory University, for providing Boundary-landmark task and Scene-face attention task and for helping with the analysis.

Table of Contents

I.	Introduction	1
II.	Methods	7
	A. Participants	7
	B. Behavioral Tasks	7
	a. Two-stage Task	7
	b. Boundary-Landmark Task	8
	c. Scene-face Attention Task	9
	d. Self-report Questionnaires	9
	C. Procedures	10
III.	Results	11
	A. Demographics	11
	B. Two-stage task	11
	C. Boundary-Landmark Task	14
	D. Between-Task Effects	15
	a. Correlation between stay probabilities after four trial conditions in the two-stage task & errors and influence in the boundary-landmark task	15
	b. Correlation between model-based characteristics in the two-stage task & errors and influence in the boundary-landmark task	16
	c. Correlation between model-free characteristics in the two-stage task & errors and influence in the boundary-landmark task	17
	E. Scene-Face Attention Task	18
IV.	Discussion	18
	A. Limitations	21
	B. Future directions	22
V.	Conclusion	24
	References	
	Appendices	
	A. Two-stage task instruction	40

Introduction

A large body of research from cognitive science, neuroscience, and behavioral economics suggest that decision-making is primarily based on two systems: controlled and automatic processing (Daw, Niv & Dayan, 2005; Daw, Gershman, Seymour, Dayan & Dolan, 2011; Dickinson 1985; Kahneman & Frederick, 2002; Kahneman, 2003; Loewenstein & O'Donoghue, 2004). In reinforcement learning, decision-making is based on the predictions of the value of an action, or the expected amount of reward that the action would bring. Consistent with controlled and automatic distinctions, reinforcement-learning theories distinguish between two algorithms of learning: model-based and model-free reinforcement learning (Daw et al., 2005). Model-based learning refers to having a cognitive map of the action-outcome contingencies in a given environment, while “model-free” denotes a more pavlovian form of learning in which one is only sensitive to recent history of rewards and punishments (Doll, Simon, & Daw, 2012; Huys et al., 2012).

Prior work has shown that these two systems are analogous to the psychological constructs of goal-directed vs habitual behavior. In reinforcement learning terms, ‘goal-directed’ is also called ‘model-based’, and the ‘habitual’ system is often compared to ‘model-free’ system. One of the key distinctions between habitual and goal-directed behavior is outcome devaluation (Daw et al., 2005). In habitual behavior, there is a strong association between an action and the situation that the action was performed in (Dickinson, 1985). Habitual behavior repeats actions that were rewarded. When the value of the outcome changes, the habitual system fails to adjust to the new value and thus favors the selection the previously realized value. Model-free system is similar to habitual system since both systems are very sensitive to reward availability, but it is different on the aspect that model-free system can learn that a previously rewarded option is not

rewarded anymore, and it leads to make a different choice to get rewards. This notion is based on the “law of effect,” which states that an action is more likely to be repeated in the future when it is rewarded. (Thorndike, 1911).

Investigations into the neurobiological underpinnings of model-based and model-free systems has revealed that model-free learning involves activity of dopamine neurons and their ventral striatal projections (Schultz, Dayan, & Montague, 1997; Yin, Knowlton, & Balleine, 2004; Yin, Ostlund, Knowlton, & Balleine, 2005). A lot of research suggest that the posterior lateral putamen is a key region that is involved in model-free learning, as putamen is part of dorsal striatum (Dayan & Dolan, 2012; Balleine & O’Doherty, 2010; Tricomi, Balleine, & O’Doherty, 2009; Wunderlich, Smittenaar, & Dolan, 2012). These findings follow the conventional idea that the dorsolateral striatum is associated with habitual or reflexive control (Packard & Knowlton, 2002).

On the other hand, the goal-directed or model-based system is based on the instant reevaluation of the new outcome and is able to adjust its choices (Doll et al., 2012). The model-based system generates behavior based on an ‘internal map’ of actions and its outcomes (Culbreth, Westbrook, Daw, Botvinick, & Barch et al., 2016). Some studies suggest that animals can also make model-based choices, by having representations of the possible outcomes of their candidate actions (Dickinson & Balleine, 2002). Model-based learning primarily involves prefrontal cortex activity (Daw et al., 2005; Sutton & Barto, 1998). Research by McDannald, Lucantonio, Burke, Niv, and Schoenbaum (2011) showed that the ventromedial prefrontal cortex (vmPFC) and orbitofrontal cortex (OFC) plays a significant role in learning in response to the change of the value and learning driven by changes in reward identity, which is necessary for model-based learning. Some studies suggest that it is reliant on the lateral and dorsolateral

prefrontal cortex (dlPFC; Smittenaar, FitzGerald, Romei, Wright & Dolan, 2013), anterior caudate (Balleine & O'Doherty, 2010), and the ventral striatum (Daw et al., 2011).

Some literatures suggested that model-based learning is affected by hippocampus activity. According to Corbit & Balleine (2000), rats with hippocampal lesions showed severely impaired performance in value-degrading and impaired ability to update the changed action-outcome contingency. Research of Simon & Daw (2011) and Johnson and Redish (2007) suggested that hippocampus may be related to cognitive map and model-based learning.

To evaluate habitual/goal-directed behaviors and to observe how they readjust the value of new outcomes, studies with rats were done by using outcome devaluation paradigms (Balleine & Dickinson, 1998; Balleine & O'Doherty, 2010; Yin et al., 2005). Outcome devaluation paradigms primarily use foods pellet as reinforcers, and most widely used one is lever-pressing task. One example is that experimenters train rats to press the lever when they are hungry, and then observe their behavior when they are full. If they are full, the value of lever-pressing should be degraded. If the rats revalue the lever-pressing, we would see it as a goal-directed behavior. If behavior of the rats are merely based on the previous reward, we would refer it as a habitual behavior (Doll et al., 2012). Human studies use similar tasks with fMRI scanning, or also involve more complex decision-making tasks associated with rewards. For example, in study by Gläscher, Daw, Dayan and O'Doherty (2010), participants were given with two choices and the reward availability according to their choices changed over time. However, in recent literatures, the most widely used task for model-free/ model-based behavior study is the two-stage task.

The two-stage task is one of the most widely used decision making tasks that is used for investigating model-based or model-free reinforcement learning. The task was first used in Daw et al.'s study (2005) and has been validated by many other studies (Daw et al., 2011; Otto,

Gershman, Markman, & Daw., 2013; Wunderlich et al., 2012; Voon et al., 2015). Using this task, Daw and his colleagues studied the neural signatures of model-based and model-free based strategies (Daw et al., 2011). Before this study, the standard view was that model-based and model-free work separately and in parallel, supported by lesion studies with rat (Daw et al., 2011; Killcross & Coutureau, 2003; Yin et al., 2004; Yin et al., 2005). However, Daw et al.,'s functional MRI study with humans (2011) suggested that people demonstrate both strategies in a mixed way, shown by overlapping neural signals, when they make a choice. Other computational modeling studies with the two-stage task also found that participants demonstrate a mixture of model-based and model-free strategies when performing the task (Daw et al., 2011; Gershman, Markman & Daw, 2013; Gillan, Otto, Phelps, & Daw, 2012; Otto et al., 2013).

Otto and his colleagues combined the model-based and model-free decision making task (two-stage task) with working memory task to see whether working memory task affects the performance of the two-stage task (Otto et al., 2013). They showed that increased working memory load led participants to be more reliant on the model-free learning strategy and that participants could actively trade-off the cognitive demands of the environment with their choice strategies by trial by trial. When there was no working memory load, participants showed a mixture of the two strategies. These findings posit that decision makers exhibit both strategies when they make choices, and that implementation of model-based processes depends on the availability of working memory and executive functioning. They also showed that model-based and model-free system can be dissociated.

Certain clinical populations show distinctive model-based or model-free reinforcement learning behaviors that may also be associated with differences in executive functioning. Population of obsessive-compulsive disorder (OCD) and disorders that involve both natural

(binge eating) and artificial (methamphetamine) rewards showed more of model-free learning in a decision task (Voon et al., 2015). Schizophrenia patients display reduced model-based decision making (Culbreth et al., 2016), and they show intact model-free reinforcement learning processes (Weickert et al., 2002) and implicit reinforcement learning (Heerey, Bell-Warren, & Gold, 2008). It has been reported that schizophrenia patients also show reduced spatial working memory (Fleming et al., 1997; Glahn et al., 2003; Goldman-Rakic, 1994; Park & Holzman 1992). Similarly, obsessive-compulsive disorder patients also exhibit spatial working memory deficits (Purcell, Maruff, Kyrios, & Pantelis, 1998; Van der Wee, Ramsey, Jansma, & Denys, 2003), and spatial cognitive dysfunction (Nakao et al., 2009).

In the current work, we compare performance in model-based and model-free decision-making to spatial processing. Interestingly, some studies have shown that the brain regions that are involved in spatial cognition work in parallel with the regions that are involved in reinforcement learning in normal populations. Doeller and Burgess (2008) observed the increased activation of the right dorsal striatum when participants were learning of landmark-related locations, whereas the right posterior hippocampus activation was associated with the learning of boundary-related locations. It is significant to note that striatal regions that are involved in model-free or habitual learning (dorsal striatum, posterior putamen) are also activated in landmark-based learning, and hippocampus was also related to model-based learning in previous literatures. Horne et al. also showed that rats with impaired hippocampi performed worse in boundary-based learning compared to control group, supporting that boundary-based learning is dependent on hippocampus (Horne, Iordanova & Pearce, 2010).

The goal of the current work is to examine the relationship between spatial cognition (boundary-landmark learning) and model-based and model-free decision-making. To assess

model-based and model-free learning, we will utilize the two-stage task. To assess spatial learning we will use the boundary-landmark task. The Boundary-Landmark Task is an object-location memory task which was first developed and used by Doeller and Burgess (2008). It has been used in several studies to examine spatial memory and incidental learning of location (Bullens et al., 2010; Doeller & Burgess, 2008). Doeller and Burgess used this task to investigate which brain regions are involved while people learn the locations of objects in relation to landmark or boundary. We chose to compare performance on these tasks to determine whether internal cognitive representations of reward environments are related to the ability to create cognitive representations of physical environments.

If an internal cognitive representation of reward environment is similar to that of physical environment, similar systems will be activated by the two-stage task and the boundary-landmark task, and thus, the performance of two tasks will be correlated. We hypothesize that model-free learners in the two-stage task will demonstrate landmark-based spatial learning, indicated by fewer errors in landmark-based learning compared to boundary-based learning. If boundary-based learning is related to model-based processing, we expect to see that individuals who show model-based learning in the two-stage task will show more of a boundary-based spatial learning. We will also examine the relationship between model-based and model-free learning and self-report measures of personality. We expect that people who had higher score at BIS-reward and fun-seeking criteria would show more model-free behavior and significantly affected by reward experience in the two-stage task.

This study is significant in that no study to date has directly compared the performance on these two tasks (the two-stage decision making task and the boundary-landmark task) to determine whether similar processes are involved decision-making and spatial learning. Through

this present study, we expect to see whether the internal representation of reward environment involve similar system of that of physical environment.

Methods

Participants

The study sample was consisted of 26 subjects, ages 18-45 ($M_{age} = 22.32$, $SD = 5.86$), recruited through advertisements at the Emory University main campus. The sample was unequally distributed in terms of gender: 54% female and 46% male. Every subject has read and signed the informed consent before the experiment. Participants were told that they will be receiving the compensation based on their performances on the task, but everyone received \$20 for their participation after study completion and a post-evaluation, regardless of their performance. This study was approved by the Institutional Review Board at Emory University.

Instructions for each task were verbally given to the participants in the beginnings of each tasks. The two-stage task had additional written instructions on the computer screen before they start the task. The experimenter stayed with the participant in the same room while they were reading the instructions and until the end of the practice trials.

Behavioral Tasks

Two-stage Task. The two-stage task was developed based on Markov decision task (Daw et al., 2011) where subjects are given with two sequential choices. Each choice image had a single Tibetan letter on a colored background. Different background colors indicated that they were in a different stage from the last set of choices; two choices at the same stage had the same background color, while different stage choices had different background colors (Figure 1). The task was consisted of 200 trials, and every subject had additional 10 practice trials to be familiar with the task before they start the main trials. There was no time limitation for choices.

In the first stage of a trial, two choices appeared side by side, and participants were asked to choose between two choices using a keyboard (stage-1 choice). After a choice was made, participants were moved to the second-stage where the image of their choice moved to the top of the screen, and another set of two choices appeared below the first choice. There were two possible set of choices in stage-2. Each choice in the first stage led to one set with 70% of probability and sometimes led to the other set with 30% of probability. The transition from the first stage to the second stage to one set with 70% of probability was called as ‘common transition’, and the transition to the other set with 30% of probability was called as ‘rare transition.’ After a choice at the second stage was made, either ‘+1’ or ‘+0’ was displayed, indicating that they were rewarded for the trial or not. The reward probabilities of each of the four stage-2 choices were all independent, ranging from 25% to 75%. They were manipulated with random walk; the probabilities started at a random value, but they were slowly changed over time with the addition of a random noise. This made sure that participants continue to sample the options, not getting to the same state and repeat their responses.

Boundary-Landmark Task. Participants were placed in the virtual reality circular arena on a first-person perspective, and they were able to navigate the arena and move the viewpoint by pressing arrow buttons. The task began with the learning phase, where four objects (vase, gift, cake, and champagne) were placed in the arena one at a time. Then they were asked to move around the arena, find and learn the location of the object, and asked to collect the object by walking over the object. After the learning phase, one of four objects was presented on a white background for 2 seconds (the cue phase), and participants were asked to replace the object within the arena, on the location they thought the object was (the replace phase). After their response, the object was appeared in its correct location. To make sure participants were learning

from feedback, they were required to collect the object before proceeding to the next trial (the feedback phase).

Each set had 16 trials with 3 blocks in total. The location of the landmark (traffic cone) was changed between blocks. The location of two objects were presented in relation to the boundary, whereas the other two objects were placed relative to the landmark. Participants were given with feedback every time they placed the object; participants would have learned the relationships between the object and landmark or boundary.

Scene-Face Attention Task. To control for individual differences in attention during the tasks, subjects also completed a matching to sample task. The task was adapted from Weigelt et al.'s study (2013). It was originally designed to measure the discrimination threshold. This task is used as a controlled task; we hypothesized that we would see no correlation between participants' performance on the other tasks and this task. On each trial, a fixation cross appeared on the center of the screen for one second. After that, a picture of face or scene (sample item) was presented on the center of the screen, and after one second, the picture disappeared and a test pair of either a face or scene were presented side-by-side. The test pair was consisted of the sample item and a distractor. Participants were asked to choose which item they had just seen. This task requires the ability to hold face or scene information in memory for a few hundred milliseconds.

Self-Report Questionnaires

After completing three tasks, all participants were asked to complete the demographic survey and two self-report questionnaires. They were informed that they could skip any questions they did not want to respond to.

BIS/BAS. Behavioral Inhibition, Behavioral Activation, and Affective Responses to Impending Reward and Punishment (Carver & White, 1994). This was designed based on the Gray's reinforcement sensitivity theory in which he suggested that behavior and affect are based on mainly two systems: a behavioral inhibition system (BIS) and behavioral activation system (BAS; 1981,1982). Twenty-four questions were asked in total.

TEPS. Temporal Experience of Pleasure Scales. (Gard, Gard, King, & John, 2005). It is designed to separately measure the individual differences in anticipatory pleasure which derives from the motivated approach to the goal and consummatory pleasure which comes from goal achievement. Anticipatory pleasure scale is related to reward responsiveness, and consummatory pleasure scale is related to appreciation of positive stimuli and receptiveness of diverse experiences (Gard et al., 2005). It has 18 items in total: 10 items for anticipatory pleasure scale and 8 items for consummatory pleasure scale (Gard et al., 2005).

Demographic Survey. Information on age, race, ethnicity, occupation, gender, education level, marital status, handedness, income level, and highest degree earned was collected.

Procedures

All experiments took place in the Translational Research in Affective Disorders laboratory at Emory University (TReAD lab, PAIS room 450). Participants were informed about possible harms, estimated time, compensations and brief descriptions of tasks via informed consent. After the subject had read and signed the informed consent, he/she was given with the first task which was one of the three main tasks (Two-stage task, boundary-landmark Task, and Scene-Face attention Task) in a quasi-random order. The participants were told that the amount of compensation will be based on their performance on the task. After finishing three tasks, participants were asked to complete the demographic survey and two self-report questionnaires

(BIS/BAS and TEPS). Then, participants were given with \$20 and debriefed that everyone has got \$20 as a compensation. All tasks and questionnaires were administered on an Apple desktop. The two-stage task was administered in Python 2.7 using Pygame 1.9. The boundary-landmark task was administered UnrealEngine2 Runtime software (Epic Games, NC) and the Scene-face attention task was administered using Matlab software (Mathworks, MA). Self-report measures were collected using Inquisit (Millisecond Software, WA).

Results

Demographics

We analyzed the two-stage and the boundary-landmark tasks separately to determine whether the performances in these tasks showed effects of gender, age, or education. For the two-stage task, there was no significant effect of age, gender, and years of education on stay probabilities in each trial type. We also analyzed effects of gender, age, or education on self-report questionnaires. BAS total score was significantly correlated with age [$r(24) = -0.580, p = 0.002$]. Among BAS criteria, fun-seeking subscale score [$r(24) = -0.421, p = 0.036$] and reward subscale score [$r(24) = -0.673, p < 0.001$] were negatively correlated with age. For TEPS, anticipatory pleasure subscale score was significantly correlated with age [$r(24) = -0.513, p = 0.009$], but consummatory pleasure subscale score did not show any significant correlation with age. The self-report questionnaire responses did not differ by gender.

Two-Stage Task

In the two-stage task, participants encountered four different trial types: trials on which the transition from stage one to stage two was common (70%) and the trial was rewarded (Common-Rewarded; CR), trials on which the transition was common but the trial was not rewarded (Common-Unrewarded; CU), trials on which the transition was rare (30%) but the trial

was rewarded (Rare-Rewarded; RR), and trials on which the transition was rare and the trial was not rewarded (Rare-Unrewarded; RU). We calculated the stay probabilities for each of the four trial types. Stay probability is the percentage of trials, for each type, where the participant selected the same stage 1 choice that they selected on the previous trial. For example, if the participant selects option A in stage 1, experiences a common transition, and receives a reward, this would be a Common-Rewarded trial. If the participant then selected option A in stage 1 of the next trial, it would be considered a stay trial for the Common-Rewarded category. The mean stay probabilities for each trial type are displayed in Figure 2 and Table 1.

There was no significant difference between the stay probability after the common transition and the stay probability after the rare transition [$t(25) = 0.802, p = 0.430$]. Participants showed significantly higher stay probability when they were previously rewarded on the choice [$t(25) = 3.978, p = 0.001$]. The mean probability of staying when rewarded in the previous trial was 0.6892, while mean probability of staying when not rewarded was 0.6102.

Consistent with previous studies, we used mixed effect logistic regression to assess effects of model-based and model-free learning at the group level (Daw et al., 2011; Gillian et al., 2012; Otto et al., 2013; Smittenaar et al., 2013). We used the lme4 linear mixed effects package in the R statistical environment to conduct the analysis and to estimate the regression coefficients. The equation that we used for mixed effect logistic regression is shown below:

$$\log(odds) = \beta_0 + \beta_1 (reward) + \beta_2(transition) + \beta_3(reward * transition)$$

The beta and p -value of the predictors are shown in Table 2.

As previous literatures showed, we observed a significant effect of reward, which is the marker of model-free learning. Also, consistent with previous work, we did not observe a significant effect of transition. However, we did not observe a significant reward-by-transition

interaction, providing lack of evidence of model-based learning, which is inconsistent with previous findings. In other studies that used the two-stage task, participants demonstrated the mixture of model-free and model-based learning processes, showing both significant effect of reward and significant interaction between reward and transition (Daw et al., 2011; Gillan et al., 2015; Otto et al., 2013). Since our sample showed significant effect of reward but no significant interaction between reward and transition, we can say that our sample population demonstrated more of a model-free reinforcement learning than a model-based reinforcement learning. Also consistent with previous work, we used the lme4 package to calculate individual beta values for each participant for reward, transition, and reward-by-transition to conduct to examine the indicators of model-free and model-based behaviors for each individuals (Bates et al., 2012; Culbreth et al., 2016; Daw et al., 2011; Gillan et al., 2015; Otto et al., 2013; Smittenaar et al., 2013; Voon et al., 2015).

In addition, we analyzed effects of gender, age, or education on the two-stage task. There were no significant effect of age, gender, and years of education on stay probabilities in each trial type. We also analyzed the correlation between self-report measures and the task. The beta value of reward-transition interaction was positively correlated with BIS scale [$r(24) = 0.426, p = 0.030$]. Higher BIS score mean more responsiveness to uncertain, nonrewarding stimuli. However, there was no difference in BIS score between rewarded trials and unrewarded trials.

For Temporal Experience of Pleasure Scale (TEPS), we did not observe a correlation between anticipatory pleasure and decision-making performance, but did observe a marginally significant correlation between consummatory pleasure subscale score and staying probability after unrewarded trials [$r(24) = 0.332, p = 0.098$]. Additionally, total TEPS score and its subscale consummatory pleasure score were positively correlated with the stay probability after Common-

Unrewarded trials, although the significance was marginal [$r(24) = 0.339, p = 0.090$; $r(24) = 0.349, p = 0.081$]. The other TEPS subscale - anticipatory pleasure scale – score was not correlated with stay probability after CU trials.

Boundary-Landmark Task

23 out of 26 participants completed 3 blocks of the task. Rest of the three participants completed only 2 blocks, but the statistical results including them were not significantly different from the results excluding them. Our statistical analysis results below also include the participants who completed only 2 blocks.

We first analyzed boundary errors and landmark errors by calculating the distance between the responses of participants and the actual object location. Boundary error was the distance between the participant’s response and the actual location of the object that was placed in relation to the landmark, and landmark error was the distance between a participant’s response and the actual location of the object that was placed in relation to the boundary. Then, using the individual’s error values, we also calculated the influence scores. Relative influence of boundary versus landmark was calculated by dividing the landmark errors by the sum of landmark errors and boundary errors using equation 1:

$$[Influence = \frac{d_{landmark}}{d_{landmark} + d_{boundary}}].$$

where $d_{landmark}$ is a landmark error, and $d_{boundary}$ is a boundary error.

The influence values ranged from 0 to 1. The value closer to 1 means that a participant was using the boundary more when they were learning the location of the object, while a value closer to 0 means that a participant used the landmark more for location learning. The overall performance of the boundary-landmark task is shown in Table 3. An influence score over 0.5 indicates a greater influence of boundary-based learning relative to landmark-based learning. In

our sample, 11.538% of participants (3 out of 26) had influence scores greater than 0.5, indicating more boundary-based learning, while 88.462% (23 out of 26) had scores lower than 0.5, indicating more landmark-based learning. In other words, our sample showed more of a landmark-based learning ($M_{influence} = 0.4141$, $SD = 0.0169$).

For the boundary-landmark task, boundary error and overall influence did not significantly differ in age, gender, and years of education. However, landmark error and age were positively correlated [$r(24) = 0.577$, $p = 0.003$]. Average error rate was not significantly differ by age, gender, and years of education, although age was positively correlated with marginal significance [$r(24) = 0.348$, $p = 0.088$]. The boundary-landmark task performances did not show any correlations with self-report responses.

Between-Task Effects

Correlation between stay probabilities after four trial conditions in the two-stage task & errors and influence in the boundary-landmark task. The average error value (average of boundary error and landmark error) was negatively correlated with stay probabilities after all trial conditions except Rare-Unrewarded condition [CR: $r(24) = -0.559$, $p = 0.003$, CU: $r(24) = -0.400$, $p = 0.043$, RR: $r(24) = -0.390$, $p = 0.049$]. We examined the correlation between CR stay probability and average error values in boundary-landmark task. CR stay probability is a general marker of performance in two stage task, since the best strategy at common-rewarded trial was to stay/repeat their first choice regardless of model-based strategy or model-free strategy. We found negative correlation between stay probabilities at CR trials and average error value in boundary-landmark task; people who performed better in two-stage task also performed better in boundary-landmark task, making less errors [$r(24) = -0.559$, $p = 0.003$].

In regards to transition without considering the effect of reward, people who chose to stay in their first choice after common transitions tend to make less errors in the boundary-landmark task [landmark error: $r(24) = -0.387$, $p = 0.051$, boundary error: $r(24) = -0.518$, $p = 0.007$, average error: $r(24) = -0.514$, $p = 0.007$]. Disregarding the effect of transition, people who had higher stay probability after rewarded trials tended to make less error in the boundary-landmark task [landmark error: $r(24) = -0.406$, $p = 0.040$, boundary error: $r(24) = -0.529$, $p = 0.005$, average error: $r(24) = -0.529$, $p = 0.005$]. Stay probability after rewarded trials was positively correlated to the influence value, even though its significance is marginal [$r(24) = 0.351$, $p = 0.079$]. This would mean that people who were more sensitive to reward showed less error rate in the boundary-landmark task, and at the same time they showed more boundary-based learning behavior.

Correlation between model-based characteristics in the two-stage task & errors and influence in the boundary-landmark task. According to previous literatures that used the two-stage task, model-based learners show lower staying probability after RR trial condition compared to RC trial condition, and lower staying probability in UC trial condition than UR trial condition (Figure 2). This is because model-based learners would attribute a reward after a rare transition to the stage 1 choice that is most likely to lead to the rewarding stage 2 choice, the unchosen stage 1 option. Alternatively, a model-free learner would attribute the stage 2 reward following a rare transition to the stage 1 choice, despite the fact that repeating the choice is unlikely to lead the participant back to the rewarding state. Based on these two distinctive characteristics of model-based learning, we defined two new variables: Rewarded-difference ($CR - RR$) and Unrewarded-difference ($RU - CU$). Rewarded-difference did not correlate with any of the variables in the boundary-landmark task. Unrewarded-difference was significantly

correlated with landmark error [$r(24) = 0.402, p = 0.042$] and correlated with average error with marginal significance [$r(24) = 0.370, p = 0.063$]. This means the bigger the difference between stay probability of unrewarded-rare and unrewarded-common, the higher the landmark error rate. However, Rewarded-difference did not show any correlations with the boundary-landmark task behavior, and the direction of correlation between Rewarded-difference and the boundary-landmark task error values and the direction of correlation between Unrewarded-difference and the error values in the boundary-landmark task were different. Hence, the result was not supportive for our hypothesis. Moreover, both Rewarded-difference and Unrewarded-difference did not show any significant correlations with the influence value in the boundary-landmark task. This data does not support the hypothesis that increased model-based characteristics are correlated with increased reliance on boundary-based spatial learning.

Furthermore, we analyzed the correlation between the influence value in the boundary-landmark task and individual's reward-by-transition interaction coefficient which is another variable that other studies have used as an indicator of model-based learning (Daw et al., 2011; Otto et al., 2013; Gillan et al., 2015). There was no significant correlation between individual's reward-by-transition interaction coefficient and influence value. The reward-by-transition interaction coefficient was not correlated with any of the error values in the boundary-landmark task (landmark error, boundary error & average error).

Correlation between model-free characteristics in the two-stage task & errors and influence in the boundary-landmark task. To examine correlations between model-free learning characteristics and landmark-based learning, we analyzed the correlation between the influence in the boundary-landmark task and individual's effect of reward (beta) which is an indicator of model-free learning in the two-stage task. While there was no significant correlations

between individual's reward beta and the boundary-landmark task influence value, there was marginally significant negative correlation between the individual's reward beta and landmark errors [$r(24) = -0.340, p = 0.062$], and also average error [$r(24) = -0.363, p = 0.068$].

In addition, we already mentioned that people who stayed more after rewarded trials tended to make less errors in the boundary-landmark task. However, to eliminate the possibility of baseline effect and to see the effect of reward more directly, we created a new variable by subtracting the stay probability after unrewarded trials from stay probability after rewarded trials (Rewarded stay prob – Unrewarded stay prob). It did not show any significant correlation with the boundary-landmark influence value, but there was a marginal significance in correlation with average error in the boundary-landmark task [$r(24) = -0.346, p = 0.084$]. Overall, the statistical analyses could not support that model-free learning is correlated to landmark-based learning.

Scene-face attention task analysis

24 out of 26 participants could complete the scene-face attention task. Two participants did not do the scene-face attention task due to time constraint and were excluded from the analysis. As we expected, the result was not correlated with any of the variables from the two-stage task and the boundary-landmark task (all p values greater than 0.3). This lack of correlation suggests that general level of attention and engagement in laboratory tasks did not generate our observed effects between spatial navigation and decision making performances.

Discussion

We studied human choice behavior in sequential decision-making task and spatial learning task. More specifically, we aimed to see behavioral correlation between model-free/model-based reinforcement learning assessed using the two-stage task and landmark/boundary-based incidental learning in the boundary-landmark task. The goal of this

project was to investigate whether there is a domain general cognitive system that supports both model-based/boundary-based learning vs. model-free/landmark-based learning. Our hypotheses were: 1) Model-free learning characteristics in the two-stage task will be correlated with landmark-based learning in the boundary-landmark task, 2) model-based learning characteristics in the two-stage task will be correlated with boundary-based learning in the boundary-landmark task.

We found that several measures of reward sensitivity were correlated with reduced errors in the boundary-landmark task. Specifically, the correlation study showed that people who stayed after rewarded trials performed better in the boundary-landmark task, indicated by significant negative correlations with landmark error, boundary error, and average error. They also showed more boundary-based learning (positive correlation with the influence value), even though the significance was marginal. People who had higher reward beta value in regression analysis had smaller value of errors with marginal significance, but the beta value did not show any significant correlations with the influence value with the boundary-landmark task. In summary, there were insufficient evidence of relationship between model-free learning and landmark-based learning, even though reward-sensitive people performed better making less errors in the boundary-landmark task. We did not find any significant relationship between model-based decision-making characteristics and boundary-based spatial learning. Increased reliance on model-based strategies was not correlated with influence of boundary-based learning.

It is possible that we failed to identify a relationship between model-based decision-making and boundary-based spatial learning because of our specific sample. Many previous studies that used the two-stage task found both significant effect of reward and significant reward-by-transition interaction, suggesting that overall their participants demonstrated a mixture

of model-based and model-free learning. Inconsistent with previous work with the two-stage task, our sample showed significantly higher stay probability after rewarded trials and showed lack of transition-by-reward interaction, indicating that the sample was more prone to model-free reinforcement learning. Similarly, the boundary-landmark task performance of our group relied heavily on landmark-based learning. Thus, it is possible that our group lacked individuals with advanced model-based and boundary-based processing, and that a sample with more similarities to previously collected data would exhibit a stronger relationship between these processes. The lack of interaction between reward and transition in our sample suggests that our participants utilized more model-free strategies than other samples, potentially hindering our ability to distinguish any relationship between model-based processing and boundary-landmark influence.

In this work, we also examined the relationships between task performance and self-reported measures of mood and personality. We did not see significant correlation between anticipatory pleasure scale reward effect, but we did observe marginally significant correlations between consummatory pleasure scale and the two-stage task performances. The positive correlation between consummatory pleasure scale and staying probability after unrewarded trials would mean that people who had higher score in consummatory pleasure scale were more willing to stay even after no rewards. However, it was not correlated with any of the model-based learning indicators. Also, in previous findings, women showed higher TEPS score than men (Gard et al., 2005), but our sample did not.

Our result showed that people who showed more model-free behavior (people who are more sensitive to reward availability) performed better in boundary-landmark task by making less errors. This may be due to the structure of the boundary-landmark task. The boundary-landmark task provided a feedback of correct answer after their responses for every trial. We had

16 trials by repeating each trials with four objects, four times. Regardless of boundary-based learning or landmark-based learning, it is plausible that people who are more susceptible to learn from feedbacks and to change their choices made less errors in the boundary-landmark task as the trials were proceeded. In the two-stage task, people who would change their choice easily depending on the result or outcome would be model-free learners. In this way of thinking, it can be one of explanations for our finding that people who are more reliant on model-free learning system performed better in the boundary-landmark task.

Limitations

As mentioned above, the design of the boundary-landmark task has a limitation in that it had feedbacks with the correct location of the object (either boundary-based or landmark-based) after each trial, and the object and its location (whether it is based on boundary or landmark) remained same and repeated four times in one block. However, the two-stage task provided reward based on probabilities with random walk, for each trial after participants made their sequential choices, and it did not show or directly told them what the best strategy to maximize their reward is. For future studies, we suggest to modify the structure of the boundary-landmark task in a way to be more parallel with the two-stage task, such as giving rewards based on their performance on each trial and not giving the correct answer feedback, or providing feedbacks for certain number of trials, not every trial.

Our study decision had several limitations related to sample size and length of our experimental session. Our sample size was small; a bigger sample size might show more varied learning types. It is possible that our sample did not represent the population well; a larger sample size would have more model-based learning components. Furthermore, other studies that used the boundary-landmark task implemented four blocks on each participant, but this study

could only do 3 blocks due to time constraints. It is possible that boundary-based learning emerges in later trial, after more experience.

Many participants reported that they felt fatigue while they were doing the task, since they had to perform three different tasks with multiple trials. Expanding this study into multiple test sessions would allow for increased length of individual tasks and decreased fatigue. Additionally, some participants reported that they felt nausea while they were doing the boundary-landmark task, due to an unfamiliar movement in the virtual reality environment. It is possible that these factors affected their performance level.

It is also possible that stress level would have affected our participants' decision making behavior. Some studies observed increase in habitual behavior following acute stress (Schwabe & Wolf, 2009; Schwabe & Wolf, 2011). More recent study done by Radenbach et al. (2015) used similar two-stage sequential decision task and showed that physiological stress response were associated with model-based behavioral control; cortisol-reactivity and model-based control were negatively correlated. We did not measure a stress level of each participant, but considering that many of our participants reported fatigue and many repeated trials with three different tasks, it is plausible that their increased stress level hinder them from demonstrating model-based learning behavior.

Future directions.

Future work with this data would include fitting the data into computational models. Previous studies that involved the two-stage task did a computational modeling to see whether participants exhibited more of model-based or model-free learning behaviors (Culbreth et al., 2016; Daw et al., 2005; Daw et al., 2011; Gläcier et al., 2010; Voon et al., 2015).

A model should have a parameter reflecting model-based and model-free weight (Daw et al., 2011). Fitting a computational model with a single parameter that captures relative reliance on model-based and model-free strategies would allow for direct comparisons between relative model-based and model-free w parameter and boundary-landmark influence from the spatial navigation task. SARSA Temporal Difference learning model is the most frequently used model for model-free behavior analysis. SARSA TD learning model follows this equation:

$$Q(s, a) = Q(s, a) + \alpha[r' + \gamma Q(s', a') - Q(s, a)]$$

where s and s' refers to the current and next state, a and a' refers to the current and next action, and r is reward. In other words, $Q(s, a)$ means the optimal strategy to maximize the expected value. For model-based learning, most of the modeling studies are based on the Bellman equation:

$$Q(s_t, a_t) = E_{s'}[r + \gamma \max_a Q(s, a) | s, a]$$

which means the optimal strategy to maximize the expected value (Daw et al., 2011; Gillan et al., 2011).

Future work could also implement neuroimaging to directly determine whether neural activation in one task is associated with neural activation in the other task. We would use functional MRI scanning to compare the neural correlates between two tasks, within subjects. We would expect that dorsolateral striatal activation will be correlated with model-free learning. Implementing fMRI methods would provide more convincing data concerning whether two learning behaviors are based on a similar system.

Future work could also utilize multiple measures of decision-making and spatial learning. The two-stage task is not the only task that could examine model-free or model-based reinforcement learning behavior, as mentioned in the introduction. However, most of the studies

that used other tasks were done with rats, using maze and lever-pressing task. Future work could focus on creating human analogs of these tasks to fully examine behavior.

Finally, our study was completed with normal population, but we expect to see studies with participants from different clinical population. The boundary-landmark task has not yet been tested with specific clinical populations, although spatial working memory and location learning were frequently studied with different clinical populations and showed some significant difference from normal population.

Conclusion

The present study has significance in that it is the first within-subject study to directly compare behavioral performances between the two-stage task and the boundary-landmark task. We tried to establish a connection between two different domains (spatial cognition and reward structure environment) and their internal representations, endeavoring to quantify and operationally prove the abstract idea that internal cognitive representation of physical environment and reward environment would be similar, and there would be a domain-general system. We could not find significant correlation between model-based learning indicators and boundary-based learning indicators and nor between model-free learning indicators and landmark-based learning indicators. However, our study suggests that increased reliance on model-free decision-making would be associated with better performance in spatial-learning task. Further studies with modified the boundary-landmark task and computational modeling are expected.

Reference

- 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
- Bullens, J., Nardini, M., Doeller, C. F., Braddick, O., Postma, A., & Burgess, N. (2010). The role of landmarks and boundaries in the development of spatial memory. *Developmental Science*, 13(1), 170–80. doi:10.1111/j.1467-7687.2009.00870.x
- Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology*, 67(2), 319.
- Culbreth, A., Westbrook, A., Daw, N., Botvinick, M., & Barch, D. (2016). Reduced model-based decision-making in schizophrenia. *Journal of abnormal psychology*, 125(6), 777–787.
- Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, 22(6), 1075–1081. doi:10.1016/j.conb.2012.08.003
- Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-Based Influences on Humans’ Choices and Striatal Prediction errors. *Neuron*, 69(6), 1204–15. doi:10.1016/j.neuron.2011.02.027
- Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12), 1704–1711. doi:10.1038/nn1560
- Dickinson, A. (1985). Actions and habits: The development of behavioural

autonomy. *Philosophical Transactions of The Royal Society B Biological Sciences B* 308(1135), 67-78. doi:10.1098/rstb.1985.0010

Dickinson, A., & Balleine, B. (2002). The role of learning in the operation of motivational systems. In *Stevens' Handbook of Experimental Psychology* (3rd ed., Vol. 3, pp. 497-533). New York: John Wiley & Sons.

Doeller, C. F., King, J. A., & Burgess, N. (2008). Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. *Proceedings of the National Academy of Sciences of the United States of America*, 105(15), 5915–5920. doi:10.1073/pnas.0801489105

Doeller, C. F., & Burgess, N. (2008). Distinct error-correcting and incidental learning of location relative to landmarks and boundaries. *Proceedings of the National Academy of Sciences of the United States of America*, 105(15), 5909–5914. doi:10.1073/pnas.0711433105

Epic Games (2016). UnrealEngine2 Runtime [Software]. Available from <https://www.unrealengine.com/download>

Fleming, K., Goldberg, T. E., Binks, S., Randolph, C., Gold, J. M., & Weinberger, D. R. (1997). Visuospatial working memory in patients with schizophrenia. *Society of Biological Psychiatry*, 41(1), 43-49. doi: 10.1016/S0006-3223(03)01641-4

Gard, D. E., Gard, M. G., Kring, A. M., & John, O.P. (2005). Anticipatory and consummatory components of the experience of pleasure: A scale development study. *Journal of Research in Personality*, 40(6), 1086-1102. doi:10.1016/j.jrp.2005.11.001

Gillan, C. M., Otto, R. A., Phelps, E. A., & Daw, N. D. (2015). Model-based learning

- protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(3), 523–536. doi:10.3758/s13415-015-0347-6
- Glahn, D. C., Therman, S., Manninen, M., Huttunen, M., Kaprio, J., Lönnqvist, J., & Cannon, T. D. (2003). Spatial working memory as an endophenotype for schizophrenia. *Biological Psychiatry*, *53*(7), 624–626.
- Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus Rewards: Dissociable Neural Prediction Error Signals Underlying Model-based and Model-free Reinforcement Learning. *Neuron*, *66*(4), 585-595. doi:10.1016/j.neuron.2010.04.016
- Goldman-Rakic, P. S. (1994). Working Memory Dysfunction in Schizophrenia. *The Journal of neuropsychiatry and clinical neurosciences*, *6*(4), 348–57. doi:10.1176/jnp.6.4.348
- Heerey, E. A., Bell-Warren, K. R., & Gold, J. M. (2008). Decision-making impairments in the context of intact reward sensitivity in schizophrenia. *Biological Psychiatry*, *64*(1), 62-69. doi:10.1016/j.biopsych.2008.02.015
- IBM Corp (2016). IBM SPSS Statistics for Mac (Version 24.0) [Software]. Available from <https://www.ibm.com/analytics/us/en/technology/spss/#spss-featured-products>
- Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*(9), 697-720. doi:10.1037/0003-066x.58.9.697
- Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In *Heuristics and Biases: The Psychology of Intuitive Judgment* (pp. 49-81). New York: Cambridge University Press.
- Killcross, S., & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cerebral Cortex*, *13*(4), 400–408. doi:10.1093/cercor/13.4.400
- Loewenstein, G. F., & O'donoghue, T. (2005) Animal spirits: Affective and deliberative

- processes in economic behavior. *SSRN Electronic Journal*. doi:10.2139/ssrn.539843
- McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y., & Schoenbaum, G. (2011). Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement Learning. *The Journal of Neuroscience*, *31*(7), 2700–2705. doi:10.1523/jneurosci.5499-10.2011
- Mathworks (2017). Matlab (Version R2016b) [Software]. Available from <http://www.mathworks.com>
- Millisecond Software (2017). Inquisit (Version 5.0) [Software]. Available from <http://www.millisecond.com>.
- Nakao, T., Nakagawa, A., Nakatani, E., Nabeyama, M., Sanematsu, H., Yoshiura, T., . . . Kanba, S. (2009). Working memory dysfunction in obsessive–compulsive disorder: A neuropsychological and functional MRI study. *Journal of Psychiatric Research*, *43*(8), 784-791. doi:10.1016/j.jpsychires.2008.10.013
- Otto, A. R., Gershman, S. J., Markman, A. B., & Daw, N. D. (2013). The curse of planning: Dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychological science*, *24*(5), 751–761. doi:10.1177/0956797612463080
- Packard, M. G., & Knowlton, B. J. (2002). Learning and memory functions of the basal ganglia. *Annual Review of Neuroscience*, *25*(1), 563-593. doi:10.1146/annurev.neuro.25.112701.142937
- Park, S., & Holzman, P. S. (1992). Schizophrenics show spatial working memory deficits. *Archives of General Psychiatry*, *49*(12), 975–82.
- Purcell, R., Maruff, P., Kyrios, M., & Pantelis, C. (1998). Cognitive deficits in obsessive–

compulsive disorder on tests of frontal–striatal function. *Biological Psychiatry*, *43*(5), 348-357. doi:10.1016/s0006-3223(97)00201-1.

Radenbach, C., Reiter, A. M., Engert, V., Sjoerds, Z., Villringer, A., Heinze, H., . . .

Schlagenhauf, F. (2015). The interaction of acute and chronic stress impairs model-based behavioral control. *Psychoneuroendocrinology*, *53*, 268-280.

doi:10.1016/j.psyneuen.2014.12.017

Schwabe, L., Wolf, O.T. (2009). Stress prompts habit behavior in humans. *Journal of*

Neuroscience, *29*(22), 7191-7198. doi: 10.1523/jneurosci.0979-09.2009

Schwabe, L., & Wolf, O.T. (2011). Stress-induced modulation of instrumental behavior: from goal-directed to habitual control of action. *Behavioural brain research*, *219*(2), 321-328.

doi: 10.1016/j.bbr.2010.12.038

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599. doi:10.1126/science.275.5306.1593

Simon, D. A., & Daw, N. D. (2011). Neural correlates of forward planning in a spatial decision task in humans. *Journal of Neuroscience*, *31*(14), 5526–5539.

doi:10.1523/JNEUROSCI.4647-10.2011

Smittenaar, P., FitzGerald, T., Romei, V., Wright, N., & Dolan, R. (2013). Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in

humans. *Neuron*, *80*(4), 914–919. doi:10.1016/j.neuron.2013.08.009

Starc, M., Murray, J. D., Santamauro, N., Savic, A., Diehl, C., Cho, Y. T., . . . Anticevic, A.

(2016). Schizophrenia is associated with a pattern of spatial working memory deficits consistent with cortical disinhibition. *Schizophrenia Research*. *181*, 107-116.

doi:10.1016/j.schres.2016.10.011

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction*. Cambridge, MA: The MIT Press.

Thorndike, E. L. (1911). Animal intelligence; experimental studies,. doi:10.5962/bhl.title.55072

Tricomi, E., Balleine, B., & O'Doherty, J. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *The European Journal of Neuroscience*, *29*, 2225–2232. doi:10.1111/j.1460-9568.2009.06796.x

Vanderwee, N., Ramsey, N., Jansma, J., Denys, D., Vanmeegen, H., Westenberg, H., & Kahn, R. (2003). Spatial working memory deficits in obsessive compulsive disorder are associated with excessive engagement of the medial frontal cortex. *Neuroimage*, *20*(4), 2271–2280. doi:10.1016/j.neuroimage.2003.05.001

Voon, V., Derbyshire, K., Rück, C., Irvine, M. A., Worbe, Y., Enander, J., ... Bullmore, E. T. (2015). Disorders of compulsivity: a common bias towards learning habits. *Molecular Psychiatry*, *20*(3), 345–52. doi:10.1038/mp.2014.44

Weickert, T. W., Terrazas, A., Bigelow, L. B., Malley, J. D., Hyde, T., Egan, M. F., ... Goldberg, T. E. (2002). Habit and skill learning in schizophrenia: evidence of normal striatal processing with abnormal cortical input. *Learning & memory*, *9*(6), 430–442. doi:10.1101/lm.49102

Wunderlich, K., Smittenaar, P., & Dolan, R. J. (2012). Dopamine enhances model-based over model-free choice behavior. *Neuron*, *75*(3), 418–24. doi:10.1016/j.neuron.2012.03.042

Wunderlich, K., Dayan, P., & Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience*, *15*(5), 786–91. doi:10.1038/nn.3068

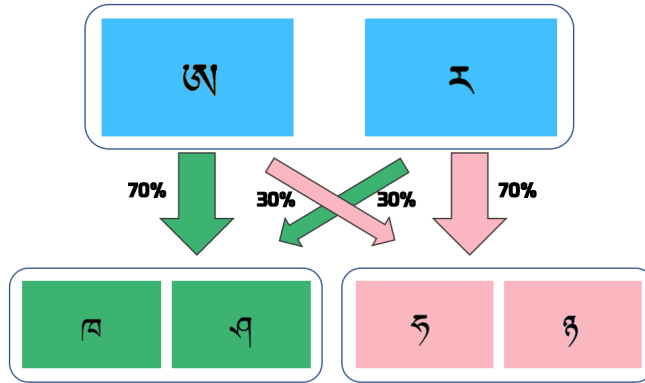
Weigelt, S., Koldewyn, K., Dilks, D., Balas, B., McKone, E., & Kanwisher, N. (2014). Domain-

specific development of face memory but not face perception. *Developmental Science*, 17(1), 47–58. doi:10.1111/desc.12089

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), 513–23. doi:10.1111/j.1460-9568.2005.04218.x

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), 181–9.

A.



B.

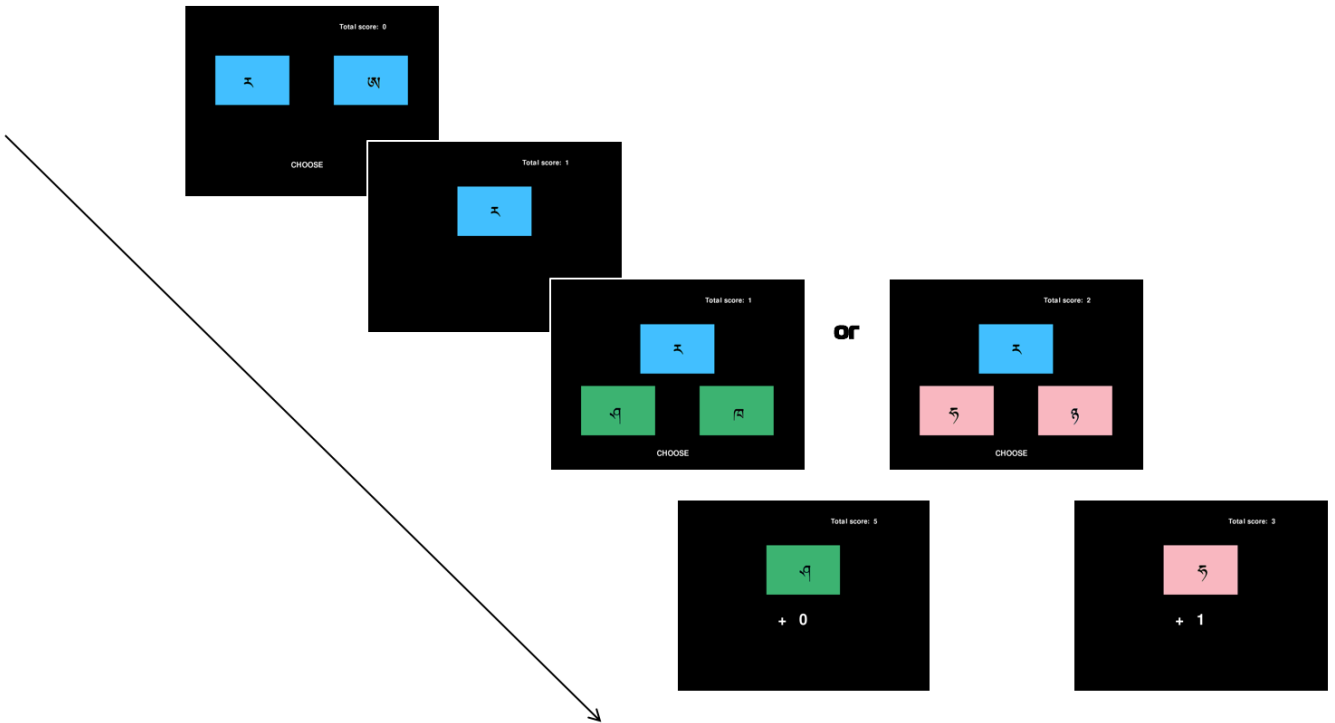


Figure 1. Task design of the two-stage task. (A) On every trial, two choices (blue boxes) are presented in the first-stage, and participants are asked to choose one. Each choice leads to one of two second-stage choices (green or pink boxes) with 70% of probability to one set of choices (common transition), and 30% of probability to the other set of choices (rare transition). Reward probability is random. (B) Timeline of a single trial.

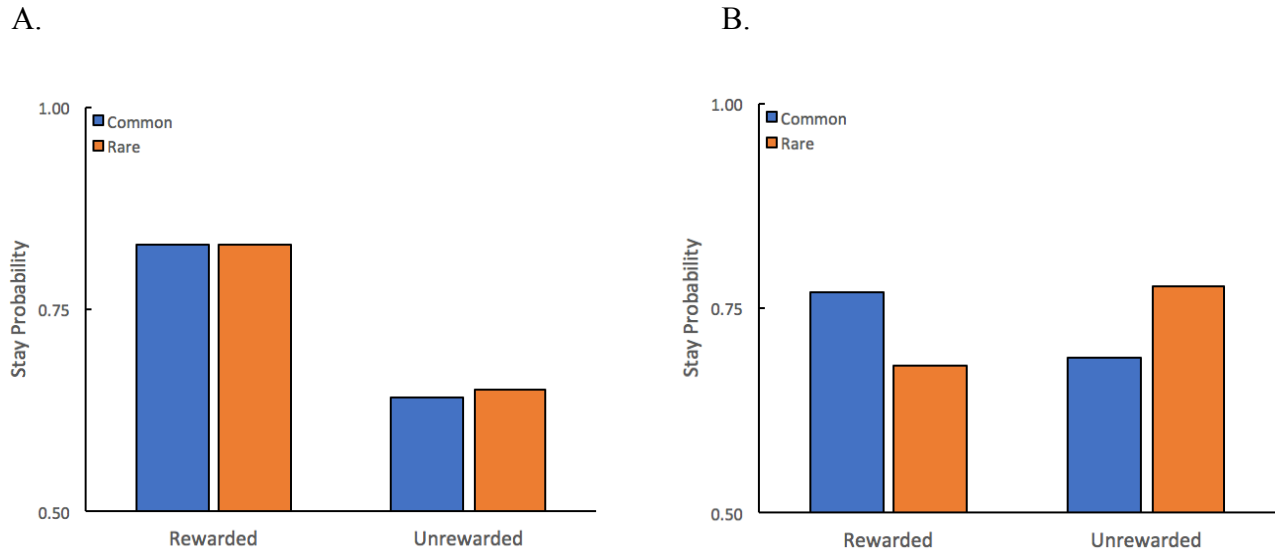


Figure 2. Model predictions for the two-stage task. (A) Model-free reinforcement learning choice strategy. Regardless of transition type, a first-stage choice that was previously rewarded will be repeated. (B) Model-based reinforcement learning choice strategy. An interaction between transition type and reward is expected. After RR trial, model-based learners would change the value of the first-stage option that they did not choose. Adapted from “Model-Based Influences on Humans’ Choices and Striatal Prediction Errors,” by Daw, Gershman, Seymour, Dayan, and Dolan, 2011, *Neuron*, 69, p. 1206. Copyright 2011 by Elsevier.

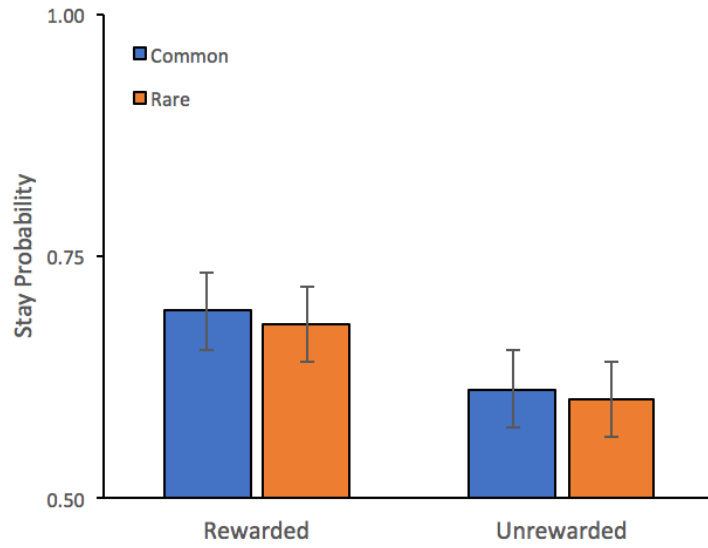


Figure 3. Results from the two-stage task. Averaged across subjects. Error bars are standard errors of the mean.

Table 1.

Result from 2-stage-task: Stay probability after each condition.

	Common transition		Rare transition	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Rewarded	0.6929	0.1578	0.6794	0.1473
Unrewarded	0.6124	0.1335	0.6018	0.1561

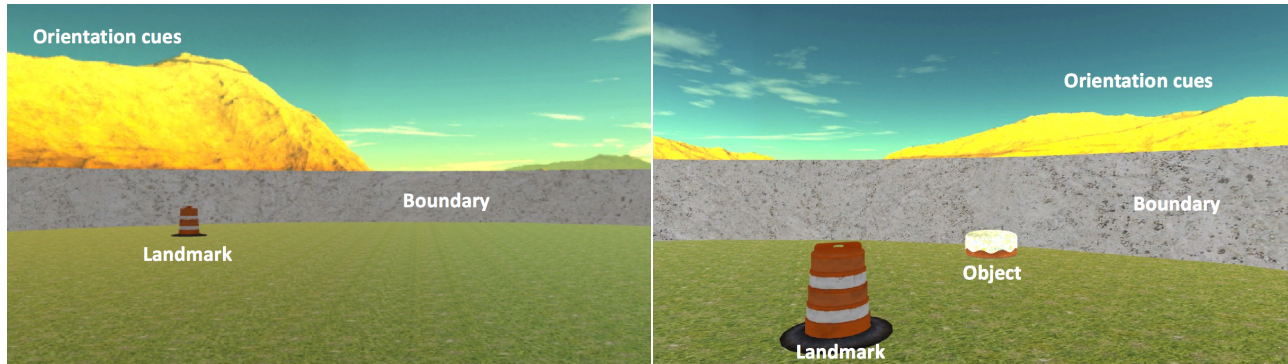
Table 2.

Results of the Logistic Regression Investigating the Influence of Previous Outcome and Previous Transition Type on First-Stage Response Repetition in 2-stage-task.

Predictor	Estimate (SE)	z-value	p-value
(Intercept)	0.478 (0.136)	3.522	0.0004 *
Reward	0.361 (0.135)	2.674	0.0075 *
Transition	0.019 (0.093)	0.205	0.8374
Reward X Transition	0.103 (0.145)	0.709	0.4780

Note. Standard errors are given in parentheses. * indicates *p*-value less than 0.01

A.



B.

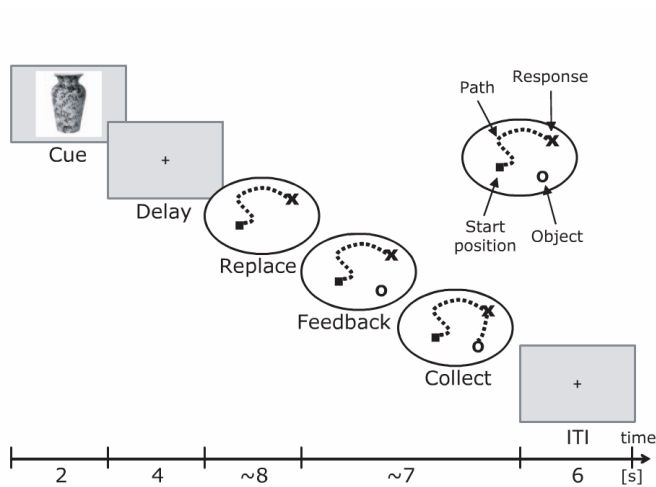


Figure 4. Boundary-Landmark Task. (A) Replace phase and feedback phase in virtual arena of the boundary-landmark task. (B) Trial structure. Adapted from “Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory,” by Doeller, King, and Burgess, 2008, *Proceedings of the National Academy of Sciences*, 105 (15) p. 5916. Copyright 2008 by The National Academy of Sciences of the USA.

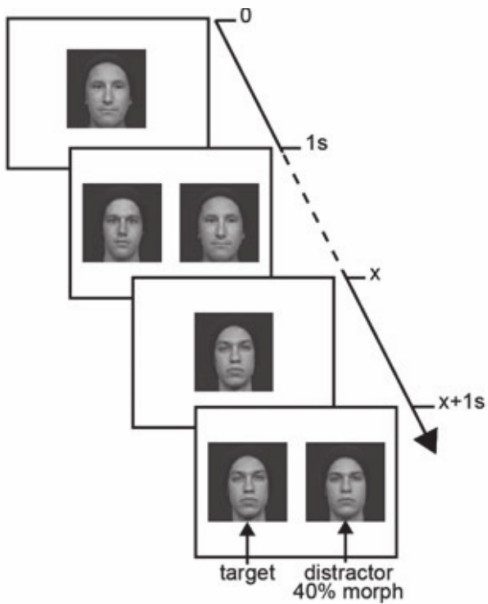


Figure 5. Scene-face attention task. Adapted from “Domain-specific development of face memory but not face perception,” by Weigelt, Koldewyn, Dilks, Balas, Mckone, and Kanwisher, 2013, *Developmental Science*, 17(1), 47–58. Copyright by John Wiley & Sons Ltd. Reproduced with permission.

Table 3.

Boundary-Landmark Task and Two-Stage Task Correlation Result

	Landmark error	Boundary error	Average error	Influence
Rewarded	-0.406 (0.040)*	-0.529 (0.005)**	-0.529 (0.005)**	0.351 (0.079)
Unrewarded	-0.212 (0.299)	-0.350 (0.080)	-0.329 (0.100)	0.221 (0.278)
Common	-0.387(0.051)	-0.518 (0.007)**	-0.514 (0.007)**	0.321 (0.110)
Rare	-0.207 (0.310)	-0.307 (0.127)	-0.296(0.141)	0.222 (0.275)
Common Rewarded (CR)	-0.426 (0.030)*	-0.561 (0.003)**	-0.559(0.003)**	0.349 (0.081)
Rare Rewarded (RR)	-0.314 (0.119)	-0.383 (0.054)*	-0.390 (0.049)*	0.298 (0.139)
Common Unrewarded (CU)	-0.289 (0.152)	-0.409(0.038)*	-0.400(0.043)*	0.268 (0.186)
Rare Unrewarded (RU)	-0.005 (0.980)	-0.162(0.430)	-0.119(0.563)	0.084 (0.682)
Rewarded Difference (CR – RR)	-0.211 (0.300)	-0.322 (0.109)	-0.309 (0.125)	0.112 (0.585)
Unrewarded Difference (RU – CU)	0.402 (0.042)*	0.312 (0.121)	0.370 (0.063)	-0.240 (0.237)
Reward Beta	-0.371 (0.062)	-0.318 (0.114)	-0.363 (0.068)	0.279 (0.168)
Reward X Transition Interaction Beta	0.065(0.754)	-0.086 (0.677)	-0.039 (0.850)	-0.061 (0.769)

Note. *p*-values are given in parentheses. Bolded values without * indicates marginal significance (*p*-value between 0.05 and 0.1). * indicates *p*-value less than 0.05. ** indicates *p*-value less than 0.01.


Appendix A. Instruction for the two-stage task

INSTRUCTIONS:


In this experiment you will be asked to make a series of simple decisions between two alternatives. On each round of the experiment you will have a chance to win points based on your choice. Your goal will be to gain as many points as possible over the course of the experiment.

Click 'continue' to see the next page of instructions.

← CLICK HERE TO GO BACK CLICK HERE TO CONTINUE →



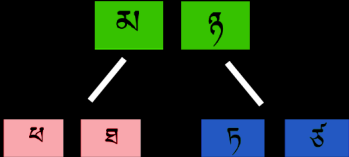
On each round, first you will see two symbols like above. You will make a choice between the two by pressing the 'z' key for the left option and the '/' key for the right option. Your choice will lead you to another stage with a new choice between two options:




Your choice on this second stage will determine your chance of scoring points in the round.

← CLICK HERE TO GO BACK CLICK HERE TO CONTINUE →

There are two possible sets of choices on the second stage. Each option on the first stage usually leads to one set (but sometimes leads to the other).




Each of the four symbols you might see on the second stage has a different probability of earning you a point. Some are better than others. The probabilities will slowly change though, so each option can become better/worse over time.



As an incentive for you to try your best, you will earn a cash bonus based on your performance in this task.

← CLICK HERE TO GO BACK CLICK HERE TO CONTINUE →

Pay close attention to the symbols since their position (left/right) can change from round to round.



← CLICK HERE TO GO BACK CLICK HERE TO CONTINUE →

