RELPH: A Computational Model for Human Decision Making

by

Nazanin Mohammadi Sepahvand

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Arts
in
Psychology

Waterloo, Ontario, Canada, 2013

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Nazanin Mohammadi Sepahvand

Abstract

The updating process, which consists of building mental models and adapting them to the changes occurring in the environment, is impaired in neglect patients. A simple rock-paper-scissors experiment was conducted in our lab to examine updating impairments in neglect patients. The results of this experiment demonstrate a significant difference between the performance of healthy and brain damaged participants. While healthy controls did not show any difficulty learning the computer's strategy, right brain damaged patients failed to learn the computer's strategy. A computational modeling approach is employed to help us better understand the reason behind this difference and thus learn more about the updating process in healthy people and its impairment in right brain damaged patients. Broadly, we hope to learn more about the nature of the updating process, in general. Also the hope is that knowing what must be changed in the model to "brain-damage" it can shed light on the updating deficit in right brain damaged patients. To do so I adapted a pattern detection method named "ELPH" to a reinforcement-learning human decision making model called "RELPH". This model is capable of capturing the behavior of both healthy and right brain damaged participants in our task according to our defined measures. Indeed, this thesis is an effort to discuss the possible differences among these groups employing this computational model.

Acknowledgements

I would like to thank my supervisor, Dr. Britt Anderson, for his continuous guidance and patience. I am deeply grateful to my family for their unwavering support and also my friends for their encouragement. Special thanks to my friends Nadine Quehl, Elisabeth Stottinger and Elnaz Barshan; without their help this thesis would not have been possible. Thank you to all the members of Dr. Anderson's and Dr. Danckert's labs for helping to shape my research ideas.

Dedication

To my grandma for her kindness and encouragement…

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Humans are surprisingly efficient in making decisions (Griffiths & Tenenbaum, 2006; Green, Benson, Kersten, & Schrater, 2010). This efficiency is remarkable because it occurs in the face of uncertainty, relies on imperfect knowledge (must utilize ambiguous cues), and takes place in an environment of variable risk and non-deterministic outcomes (Trimmer et al., 2011; Fellows, 2004; Bland & Schaefer, 2012; Bach & Dolan, 2012; Payzan-LeNestour & Bossaerts, 2011). The underlying mechanism of decision making however is not yet clearly understood. Our lab is particularly interested in understanding how people improve their decisions based on their experiences. Our approach is to investigate this mechanism in patients who have brain injury and difficulties in adaptive sequential decision making tasks. The contribution of this thesis is to employ computational methods to compare the results of brain damaged patients and healthy people. This comparison can shed light on the decision making mechanism in the brain. On the other hand, it can help more understand the possible reasons behind the brain damaged people's failure in this process.

To facilitate their decision making process humans build mental models of their environment. Optimal decision making requires these models to be updated constantly based on concurrent feedback from the environment. Any failure in this updating process leads to poor decisions. To investigate this nature of updating in people with brain damage we have engaged our participants in playing the Rock, Paper, Scissors (RPS) game against a computer opponent. The computer opponent can employ various strategies, and may shift its strategy during the course of play. An intact updating mechanism is necessary for participants to learn the computer's strategy after each switch. The data collected from this task provides a global view of the updating process, but does not give details on the necessary steps and sub-processes involved. To begin to address these limitations, I employed a computational modeling approach. The purpose of this thesis is to offer a computational model of the updating process in the brain which replicates the results of the RPS experiment for both healthy and brain injured participants, and offers a more detailed understanding of the updating process.

This thesis is organized as follows: for the rest of this chapter, I elaborate on the idea of updating in brain damaged patients, the RPS task we developed to examine the idea and the preliminary results of this task. In chapter 2, I describe the computational model we selected as our initial potential candidate of human decision making and investigate the performance of this model in our RPS task. The remainder of this thesis describes our efforts to improve this model based on well-known related facts about the human neural systems involved in learning and decision making.

Motivation and Summary of Neglect

Hemineglect, also known as Hemispatial neglect or simply neglect, is a common consequence of injury to the right side of the brain. The injury is often caused by a stroke and patients typically fail to notice or respond to salient or novel stimuli appearing in the contralesional visual field, although their primary sensory and motor processing systems may be intact (Driver & Mattingley, 1998; Heilman, Bowers, Valenstein, & Watson, 1987). The performance of these patients is relatively preserved for stimuli located ipsilaterally to the hemispheric lesion. Neglect can follow injury to either the right or left side of the brain but it is much more frequent in patients with right brain damage. It has been traditionally associated with the right posterior parietal cortex, particularly the inferior parietal lobe (IPL) or temporoparietal junction (TPJ) (Vallar & Perani, 1986; Karnath, Ferber, & Himmelbach, 2001). Recent studies however demonstrate the involvement of other brain areas, mainly prefrontal cortex, basal ganglia and insula (Danckert, Stöttinger, Quehl, & Anderson, 2012; Karnath, Himmelbach, & Rorden, 2002; Karnath, Berger, Küker, & Rorden, 2004).

Three classical tasks to examine hemineglect in the visual modality are figure copying, line bisection and object cancellation (Bisiach, Capitani, Nichelli, & Spinnler, 1976; Schenkenberg, Bradford, & Ajax, 1980; Albert, 1973; Mesulam, 1985; Wilson, Cockburn, & Halligan, 1987). In a typical line bisection task, participants are asked to bisect a horizontal line. Since the neglect patients ignore the left side of their visual field, their selected middle point tends to be placed toward the right of center. Similarly, neglect patients are likely to ignore the objects appearing on the left side in object cancellation tasks, and the left side of the figures in figure copying tasks.

The result of these tasks demonstrates that an impairment of spatial cognition is one of the deficits in neglect. In fact it is the most striking deficit observed in neglect. Spatial impairments in neglect have been reported in several other tasks such as the double-step saccade task (Duhamel et al., 1992), the covert visual attention task (Posner & Cohen, 1984; Losier & Klein, 2001; Striemer & Danckert, 2007) and the motor imagery task (Danckert et al., 2002). Recent studies however have demonstrated that RBDs' deficits are not restricted to spatial impairments only; non-spatial impairments also observed in some neglect patients (Robertson et al., 1997; Husain et al., 1997; Shapiro et., 2002; Shaqiri & Anderson, 2012; Danckert et al., 2007; Merrifield, Hurwitz, & Danckert, 2010). In 1997, Robertson and colleagues indicates that auditory sustained attention is impaired in neglect patients. Abnormal temporal attention in visual domain is also reported in different studies (Husain et al., 1997; Shapiro et., 2002). Some of the "higher-order" non-spatial deficits such as deficits in theory of

mind (Happé, Brownell, & Winner, 1999; Griffin et al., 2006) empathy (Shamay-Tsoory, Tomer, Berger, & Aharon-Peretz, 2003), appreciation of humor (Brownell, Potter, Bihrle, & Gardner, 1986), and the ability to differentiate between lies and jokes (Winner, Brownell, Happé, Blum, & Pincus, 1998) are also reported in neglect as well. In this section, I will describe examples of spatial and non-spatial deficits reported in RBDs.

Spatial and non-spatial deficits in RBDs

Spatial updating is impaired in neglect patients. A common task to investigate spatial updating is the double-step saccade task (Duhamel et al., 1992). In this task, participants are required to make two successive saccades to rapidly extinguished consecutive targets. The target can appear on any location on the screen. While control participants can perform this task with high accuracy; neglect patients show difficulty when a target appears first on the contralesional side and then ipsilesionally (Duhamel et al., 1992; Heide, Blankenburg, Zimmermann, & Kömpf, 1995). Given the fact that the ipsilesional side of the visual field is not neglected, this result suggests that the problem cannot be simply due to some primary sensory damage such as retinotopic, spatial coding or saccade generating deficit (Duhamel et al., 1992; Heide, Blankenburg, Zimmermann, & Kömpf, 1995).

Poor spatial motor imagery is also evidence of spatial representational impairment in neglect patients. In 2002, Danckert and colleagues showed that imaginary motor movement is impaired in a neglect patient (Danckert et al., 2002). In this task, the participant was asked to first point to a target and then to imagine the same motor action. For the actual motor action the participant showed the same movement pattern as healthy controls: increased movement duration for smaller targets. Humans' movements are faster to a large target than a small one. This pattern of speed-accuracy trade-off is even observed for imaginary movements (Fitts, 1954; Decety, Jeannerod, & Prablanc, 1989; Jeannerod, 1997). While the patient showed the same pattern as healthy controls for the real movement, his movement for the imaginary part didn't follow this pattern.

But cognitive disorders in neglect patients are not limited to spatial tasks; they also involve non-spatial deficits as well. Position priming and statistical learning are two frameworks that have been employed in our lab to investigate the non-spatial impairment in neglect patients. The priming effect refers to how the repetition of a feature of a target, such as position, facilitates human visual search. This effect has been widely studied in visual search (Kristjánsson & Campana, 2010; Kristjansson, 2008; Neely, 1991). It has been shown that when a feature of a target remains fixed from one trial to another, it influences participants' reaction time; they get faster in detecting that target for those trials

(Maljkovic & Nakayama, 1996). The effect of color and position priming on neglect patients has been recently tested in our lab (Shaqiri & Anderson, 2012). The task was to report the color of the stimulus (which was either white or black); the stimulus had a 75% chance of appearing in a high probability region or so-called "hot spot" on the left side of the screen, the neglected space, and 25% chance of appearing on rest of the screen. Neglect patients showed preserved color priming, but their result for location priming was quite different. Although the neglect participants were a bit faster to respond when the target repeated in the same position; this priming benefit was not significant compared to non-primed trials.

Statistical learning is another suitable framework to study non-spatial updating (Turk-Browne, Isola, Scholl, & Treat, 2008; Aslin & Newport, 2012). Statistical learning helps humans to learn the regularities distributed in space and time. Thus both statistical learning and the priming effect are quite similar. In fact, some researchers believe the concepts are so closely related that priming might be a form of statistical learning (Walthew & Gilchrist, 2006). To investigate statistical learning in neglect patients, Shaqiri and Anderson conducted the "hot spot" experiment with some variations (Shaqiri & Anderson, 2012). In order to separate statistical learning from the priming effect, target location in this experiment was varied throughout the screen. We also removed the potential primed trials by excluding the trials in which the previous target location was within 5 degrees of visual angle. Considering only the trials in which the target appeared on the left side of the screen, neglect patients showed some benefit from the regularity since they were faster to respond for the hot spot. Their RTs however were slower compared to their RTs to right sided targets. Thus our result confirms difficulties in learning statistical distribution for those participants.

Updating: a unified theory?

Putting all these results together, we hypothesize that impairment in "representational updating" is a unified theory that can explain most of these deficits observed in neglect patients (Danckert et al., 2012). According to this hypothesis, humans build mental models and improve them based on new experiences. These internal representations facilitate decision making. This facilitation can be a great help especially in cases that humans must make efficient decisions quickly. Humans build such mental models based on their experiences and through their interactions with the environment. To evaluate the accuracy of their mental model, people compare its predictions to subsequent observations. Mismatches between what we expect to see and what is actually observed are employed to improve our internal representations. Any failure in this updating process leads to an inability to

learn from the past experiences, which in turn leads to poor internal model representations and therefore poor decisions.

This hypothesis can explain neglect patients' spatial and non-spatial updating deficits. As long as patients have difficulties updating their internal representations from observations, one should expect poor performance in related tasks such as those mentioned above. In the double-step saccade task, in order to make the correct second saccade, one must anticipate the outcome of the first saccade to update the spatial representation; any failure in updating this representation after the first saccade can lead to a poor result. Failure in imagining the right pattern of movement in the motor imagery task can be another instance of poor internal spatial representation in neglect patients. In both non-spatial tasks, priming and statistical learning, participants seem to be impaired in learning the high probability spot in the screen. In contrast to healthy people, participants with neglect failed to update their initial belief about the task based on the oncoming information

Updating seems to be a suitable framework to explain neglect patients' failure in a variety of tasks. To have more insight into updating deficits in neglect patients, we need to have a more accurate definition of updating. The updating process consists of two main sub-processes: (1) finding mismatches between the current model's predictions and actual observations and (2) using these mismatches to adapt the model. While the definition seems straightforward, there are several questions that arise about this mechanism in humans.

The first question I will address involves the uncertainty about the source of the mismatch. Although mismatches between a model's prediction and real observation always show a discrepancy, given the fact that our environment is in flux and noisy, not every mismatch reflects model inaccuracy. Even for the ideal condition of having a perfect model of the task, there is still a possibility of making errors in prediction. This type of failure is due to the probabilistic nature of the process. For example, imagine a task in which taking a certain action 80% of the time leads to a win and 20% of the time to a loss. While you know the rule, still you cannot be 100% sure about the next outcome of taking this action. This uncertainty however doesn't reflect your lack of knowledge about the rule governing the task; it represents the inherent non-deterministic nature of the rule. This means that not all the mismatches are worth considering as model failures and the question is how humans manage to distinguish these different sources of errors. The other main question is that even if humans manage to differentiate between two different types of error, how do they use this

information to update their internal model? The process must be consistent with the evidence of human adaptive decision making and also be able to explain the updating failure in neglect patients.

To find the answer to these questions and several other questions about the updating mechanism and its deficit in neglect patients, we needed to find a simple task that was still rich enough to help us have a better insight into the problem. For these reasons we selected the classic children's game, rock-paper-scissors (RPS). We have used RPS as our experimental procedure because this game was used in prior studies of decision making (e.g., Kangas et al., 2009; Danckert et al., 2012) is a well-known game, has rich dynamics (Sato, Akiyama, & Farmer, 2002), and is simple enough to be used in human populations with brain injury or in animal studies (Danckert et al., 2012; Lee, McGreevy, & Barraclough, 2005; Lee & Seo, 2007; Lee et al., 2005). An RPS experiment was conducted in our lab with 12 healthy controls, 15 patients with right brain injury and 10 patients with left brain injury (Danckert et al., 2012). The results show that updating in the RPS task was selectively impaired after brain injury. In the rest of the chapter, the task, the results and my computational approach are described in three separate sections.

Description of the RPS Task

To examine the idea of updating impairment in RBD patients, we conducted an experiment with three different groups of participants in our lab. RPS is a competitive game in which two opponents must play against each other. Each one has to play rock, paper or scissors at each trial. The rule is simple: rock beats scissors, scissors beats paper and paper beats rock. In order to win at each trial one must anticipate the opponent's strategy. The players are also aware of the fact that the opponent can change his strategy after each trial.

In our RPS study, healthy controls, as well as people with right or left brain injury played 600 rounds of RPS against a computer. Figure 1 shows the schematic representation of one trial of the experiment. A trial began with 2 red squares, vertically aligned in the center of the screen. The top square represented the computer's choice, and the bottom one represented the participant's choice. Participants were told that once the computer made its choice on a given trial, the square representing the computer would change from red to green. Participants knew that at this point, the computer's choice would not change. They were then free to make their own choices and change their choices until they reported that their selection was finalized. After that, the computer's and player's choices for that trial would be revealed by replacing the colored squares with images of the choices. When the

participant was ready, another button press began a new trial. No explicit feedback was given regarding the outcome of a given trial.

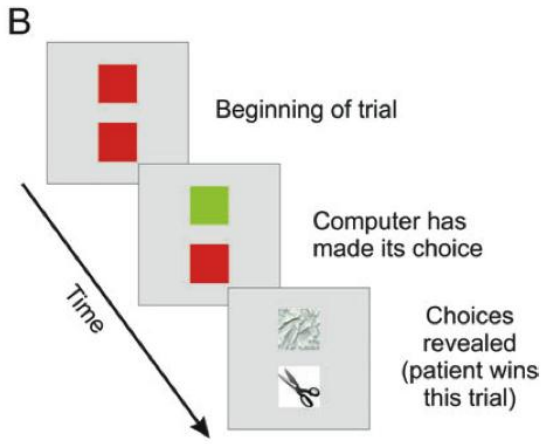

Figure 1. A schematic representation of one trial of the RPS experiment

In this experiment, participants played 600 rounds of RPS against a computer starting with a "random condition" for 200 trials. At each trial in the 'random' condition, 'rock', 'paper' or 'scissors' had the same probability of being chosen by the computer. After 200 trials, the computer strategy switched to a moderate bias of 'rock', called 'lite' condition, for another 200 trials. In the 'lite' condition, rock was chosen half of the time, and each of the two other options had 25% chance of being picked. Finally, for the last 200 trials the computer strategy switched from 'lite' to a 'strong' condition, in which paper was played 80% of the time and the two other options had the same probability of 10% to be chosen.

Right Brain Damages Impaired to Update in the RPS experiment

As we hypothesized, the average result of each group in our task differed. Studying the result of the experiment indicates that, on average, RBDs have an impaired updating process compared to both the healthy controls and LBDs. Figure 2 shows the average win rate for the three groups versus time. The win rate measures the proportion of wins, calculated from a moving window of the last 20 trials. Participants were not informed about the computer strategy and they were expected to deduce this during the course of play from their own observations; thus any difference in the results of our groups represents a potential difference in the updating process. It is evident from the figure that control

subjects rapidly recognized the computer's transition to the heavily biased strategy and adapted their play. Mean proportion of optimal play (averaged over blocks of 10 trials) for the strong condition was submitted to a mixed design repeated-measures ANOVA with within subjects factors of block (block 1 to block 20). The result of this analysis suggests that although HCs didn't manage to reach the maximum possible win rate which is 80%, the proportion of scissors (averaged over blocks of 10 trials) in their play for this condition is statistically greater than chance ($F(1,16) = 14.732, p < 0.001, \eta^2 = 0.479$). The same analysis calculated for the mean proportion of wins (again averaged over blocks of 10 trials) revealed that wins increased significantly over the course of strong condition (block X condition interaction: $F(19, 418) = 5.14, p < .001, \eta^2 = .20$).

While LBDs also recognized this transition, their results differ from HCs. LBDs reached a better maximum win rate but took longer to do so. However, this difference was not significant in terms of either steady state win rate ($F(1,20) = 0.027, p > 0.1, \eta^2 = 0.021$) or rate of learning ($F(1,20) = 1.890, p > 0.1, \eta^2 = 0.086$). Steady state win rate refers to the stable win rate that each participant managed to reach after learning the computer's strategy. This is clear in Figure 2 as the win rate that each group reaches after the transient steep increasing slopes in their plots. Even after this transient behavior, there are still some fluctuations in win rate, but in general the value is quite stable. For the sake of simplicity, this measure is called "max win rate" for the rest of this thesis. Learning rate refers to how fast a participant learns the computer's new strategy, or in other words, how long it takes him/her to reach to his/her max win rate. To compare the learning rate and max win rate of these groups, the first 10 blocks of the strong condition is considered as the transient behavior and used to compare the learning rates. The second 10 blocks are employed to compare the max win rate.

RBDs typically fail to note the transition, and continue to play varying strategies (Danckert et al., 2012). A repeated-measures ANOVA with within subjects factors of block (block 1 to block 20) shows that the proportion of scissors in RBDs' plays in the strong condition, on average, was not statistically better than chance (main effect: $F(1,18) = 1.822, p > 0.1, \eta^2 = 0.092$) and also the win rates didn't change significantly over the course of strong condition (block X condition interaction: $F(19, 418) = .523, p > 0.1, \eta^2 = 0.22$). This means that RBDs didn't get any closer to the proper strategy which is playing scissors more than chance.

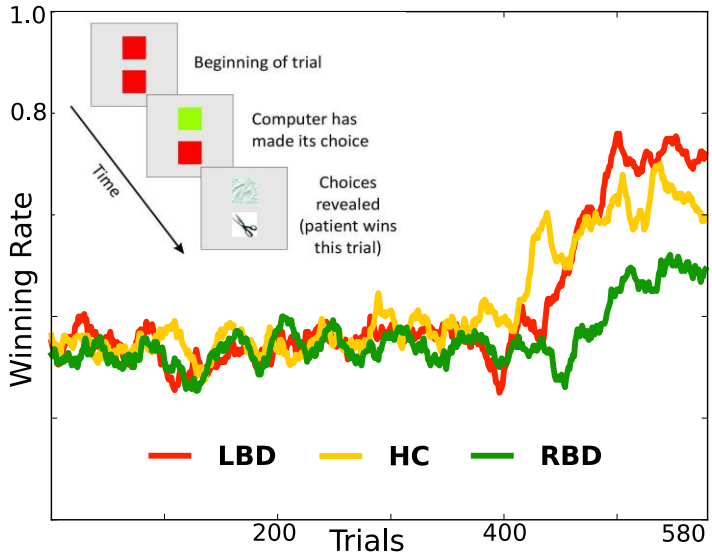

Figure 2. The average win rate for our three different groups against the number of trials. The average win rate is calculated from a moving window of last 20 trials. The orange line shows the average win rate for healthy controls. The red line represents the average win rate for left brain damaged patients (LBD) and the green line is the average win rate for right brain damaged patients (RBD).

Our RPS task is designed in such a way that the computer opponent changed its strategy twice; first from random to a moderate bias and secondly from a moderate bias to a strong bias. The purpose was to have two so-called change-points in this task, which gives us the opportunity to compare the performance of HCs and RBDs more accurately. Our analyses however demonstrate that the proportion of rock being played in the lite condition for HCs is not statistically better than chance (HCs: $F(1,16) = 0.509, p > 0.1, \eta^2 = 0.031$). In addition HCs' and RBDs' win rate are not statistically different in this condition ($F(1,24) = 0.002\ p > 0.1, \eta^2 = 0.001$). The same holds true for LBDs as well; they are also not statistically different from HCs in the lite condition ($F(1,20) = 0.002, p > 0.1, \eta^2 = 0.002$). This suggests that unfortunately all HCs, LBDs and RBDs failed to notice this change in the computer strategy; they both missed this change-point. Even if they noticed this change-point, it did not reflect in their sequence of play; thus investigating their plays would not give us any helpful information. Thus for the remainder of this thesis I restrict myself to analysis of the strong condition.

Discussion

Considering the many deficits reported in neglect patients, we hypothesized the idea of an "updating impairment" as a general impairment in neglect. Updating is defined as the process of (1) detecting a change in the environment and (2) adapting to this change using previous experiences. To evaluate the idea of an updating impairment in neglect patients and also to have a better understanding of the updating process per se, we conducted the RPS task reported in the last section. The results generally confirmed our hypothesis, but failed to provide evidence for the reason why they failed to master the task. For example, it was possible that RBDs failed to notice the change in the computer strategy or they may have noticed the change, but were unable to discover the computer's new strategy.

To investigate the reasons for this failure more thoroughly I employ a computational modeling approach. This approach can help us to be more specific about our definition of updating. The main advantage of computational modeling, in our case, is to give us the opportunity to replicate the result of healthy participants in our experiments. Any failure in this replication enables us to improve our model and therefore helps us to have a better understanding of the updating process in the brain. Another important advantage of this method is any effort to "lesion" the model could shed light on the possible impairments in updating processes in the RBDs. "Lesioning" would happen by change-the parameters of the model to replicate the results of RBD patients in our RPS task. This is the main logic behind this thesis. I will first start with a simple computational model which has been suggested in computer science as a novel temporal sequence detection method in a nonstationary environment, and which is an excellent RPS player. I will employ this model to reproduce the healthy controls' behavior in our task in general. Using these preliminary results, I will adapt my computational model to reproduce the main aspects of our healthy controls' results. In the next step, I model brain-damaged participants' performances by change-parts of the model, which are thought to be impaired in RBDs.

Chapter 2: ELPH: A pattern recognition method

The prior chapter reviewed evidence that patients with RBD have an "updating" impairment. Updating problems were seen in the RPS task reported by Danckert and colleagues in 2012. The purpose of this thesis is to find a computational model which is able to reproduce the behavior of both healthy people and people with RBD for the same task. To begin our computational investigations, I start this chapter with a search among possible models suitable for settings similar to the RPS game. Subsequently, I describe the selected candidate (ELPH) and give the primary results for this model.

Potential Candidates

There are many machine learning algorithms adaptable to RPS (Dempster, Laird, & Rubin, 1977; Sutton & Barto, 1998; Chen, Cowan, & Grant, 1991; Poupart, Vlassis, Hoey, & Regan, 2006; Rabiner & Juang, 1986), but all have some limitations. In RPS, similar to other sequential decision making tasks, participants have to make their decision while they are playing (i.e., online). This requires developing a strategy within relatively few trials. Further, in RPS, as we used it, and for human decision making in general, the decision making environment is non-stationary[1], and strategies change over time. Unfortunately, most common learning algorithms are designed for stationary processes only. Although this does not exclude the possibility that they could be used for non-stationary environments, there is no proof of convergence in such environments. Even if they converge, they may be too slow to adapt. Furthermore, these algorithms often require large amounts of data for learning and most of the time this amount of data is not available for human learners. Lastly, many learning methods are best fitted to offline learning processes; they digest a batch of data in order to determine their optimal policy. Since our task, and updating in general, is challenged with the need to learn online under the assumption of nonstationarity of the environment, many available algorithms are not applicable.

One interesting model which has claimed to overcome some of these limitations is the one suggested in 2005 by Jensen and colleagues. Named *Entropy Learned Pruned Hypothesis Space* (ELPH) (Jensen et al., 2005) this is a sequence learning method for an agent that must (1) be able to learn from sparse data, (2) learn online while it is interacting with the environment, and (3) anticipates the nonstationarity of the environment. It is therefore able to adapt to changes in the

[1] Stationary processes are those that have probability distributions that do not change over time.

environment efficiently. More importantly, the algorithm was informed by ideas from human psychology (e.g., the size of its memory) which makes it a good candidate for our attempt to model human performance. Lastly, this model has been shown to be an excellent RPS player. It competes favorably with many of the popular machine learning methods and it out-performs typical human players (Jensen et al., 2005). It is not likely to be a perfect model for human decision making since it is a pattern recognition method, as opposed to a strategy learning one; how we adapt it may help characterize the nature of updating and how deficits occur. By comparing ELPH's performance to the results of our three participant groups we will gain a better understanding of updating. In this chapter I will develop the ELPH model and report its straightforward application to our participant RPS data.

ELPH: Details of its pattern recognition method

ELPH proposes a two-compartment approach to sequence learning. The Short Term Memory (STM) component contains the *n* most recent temporally ordered observations and the Hypothesis Space (HS) contains individual hypotheses about what is likely to be observed given what has been seen. Individual hypotheses are ordered subsets of prior observations taken from STM. As an example, imagine that this method is exposed to the pattern below:

rock, rock, paper, rock, rock, paper, rock, rock, paper,…

Let's say the length of STM (*n*) is equal to 2. *n* =2 means that at the end of each trial the two last observations are saved in STM. Imagine that at time *t-1* the last two observations were 'paper' and 'rock'. Then the current content of STM would be the tuple ('paper', 'rock'). The goal is to make a prediction about the next observation according to related hypotheses stored in the HS. Related hypotheses are those which predict what is likely to be observed next based on the current content of STM.

In order to make a prediction from hypotheses in HS we need to generate those hypotheses in the first place. The logic behind hypothesis generation in ELPH is that every subset of STM might be a predictor of the next observation. In our example it might be that rock at *t-1* is always followed by rock at *t*. It is also possible that rock at each trial must be preceded by a paper and a rock respectively. The last possible option is that paper at *t-2* is an indicator of next observation regardless of whatever item appeared at *t-1*. So all the possible hypotheses in this case are:

Hyp_1: {(paper, *t-2*), (rock, *t-1*)} => rock

Hyp_2: {(rock, *t-1*)} => rock

Hyp_3: {(paper, *t-2*)} => rock

These hypotheses are stored for later reference. The task is to learn over time which of these hypotheses actually predicts the next observation and what is that prediction. To learn the predictive hypotheses, a "prediction set" is associated with each hypothesis consisting of all the events that have immediately followed that hypothesis and a count of the number of times each outcome has actually happened. For a particular hypothesis, *Hyp*, a prediction set is as follows:

$$\{[e_1, c_1], [e_2, c_2], \ldots, [e_m, c_m]\}$$

here e_1, e_2 … e_m is the list of possible outcomes (rock, paper, or scissors in the present case) and c_1, c_2... c_m represents the counts of how many times each has been observed. For example, after 20 trials, each hypothesis will be updated to:

Hyp_1: {(paper, *t-2*), (rock, *t-1*)} => [rock, 7]

Hyp_2: {(rock, *t-1*)} => [rock, 3], [paper, 4]

Hyp_3: {(paper, *t-2*)} => [rock, 7]

This means the sequence of paper and rock has been observed 7 times so far and it always leads to rock. Observing rock at *t-1*also happens 7 times but 3 times it has been followed by rock and 4 times by paper. The idea behind the learning process in this method is that if a subset of observations has been followed consistently by an event, it is most likely to be followed by the same item in the future. The learning process is simply an updating of these prediction sets based on incoming data.

The next step is to generate a prediction from these hypotheses and their sets of observations. In our example, Hypotheses 1 and 3 are always followed by rock. Hypothesis 2 however is followed half of the time by paper and half of the time by rock. This means that the second hypothesis is not as worthy, because it is not able to make as reliable a prediction about the subsequent item. Desirable hypotheses are those whose predictions are consistent and trustworthy or, more accurately, less random. The final outcome of the algorithm (prediction of next observation) is one of the most predictive hypotheses. Thus the first step to make a prediction about the subsequent observation is to determine which hypothesis has the least random event set and therefore is the most predictive hypothesis.

To mathematically evaluate how trustworthy the predictions of a hypothesis are, the entropy of that hypothesis is calculated at each trial. Entropy is a measure in information theory which calculates the amount of randomness and uncertainty associated with an event (Cover, Thomas, & Kieffer, 1994). The higher the amount of entropy, the more randomness exists for that event. If x is the random variable and $P(x)$ is the probability distribution over x, then the entropy for this random variable is calculated as follows:

$$H(P) = -\sum_x P(x) log_2 P(x) \qquad (1)$$

In our method for a particular hypothesis, *Hyp*, with a prediction set of $\{[e_1, c_1], [e_2, c_2] \ldots [e_m, c_m]\}$, the entropy is calculated below[2]:

$$P(e_i) = \frac{c_i}{\sum_{j=1}^{m} c_j} \rightarrow H\big(P(Hyp)\big) = -\sum_{i=1}^{m} \frac{c_i}{(\sum_{j=1}^{m} c_j)+1} \; log_2 \frac{c_i}{(\sum_{j=1}^{m} c_j)+1} \qquad (2)$$

Therefore ELPH tracks the observations that follow each sequence in the hypothesis space and from this data it computes the entropy. Among all related hypotheses, the most predictive one is the one with the lowest amount of entropy, which is chosen to predict the next data. This hypothesis has a prediction set which consists of all the predictions of that hypothesis. Next step is to find the best candidate to predict the next observation among all the available events in the prediction set of this

[2] This definition is slightly different from the exact definition of entropy, for more details see (Jensen, Boley, Gini, & Schrater, 2005)

hypothesis. This candidate is the most frequent event. For instance the prediction of *Hyp* with the same prediction set as above is e_j in which c_j (the count of the number of times e_j has been predicted by this hypothesis) has the maximum value among all c_k where $k = 1,\ldots m$.

Prediction in ELPH requires recording all the items have been observed so far. There are two principal problems with simply accumulating all past observations. First, the number of hypotheses grows too large, too quickly, as STM capacity increases. If n equaled 5, there would be 2^5-1 related hypotheses stored in HS. The second problem is that the history of observations lengthens. To compensate for this growth in the hypothesis space ELPH prunes the hypotheses based on how useful they are for prediction.

To remove the *inconsistent* hypotheses, entropy is used again. The entropy of all the stored hypotheses in the HS is computed at the end of each trial. Hypotheses with an entropy value more than a threshold (denoted as H_{thr}) are deleted. The idea is the same as for prediction: the hypotheses with high entropy are the ones with conflicting predictions; they are less useful and therefore less valuable. They are not worth the storage space. Pruning HS is necessary not only because of a limited storage space, but also because of nonstationarity; pruning facilitates adaptation. In this case the hypotheses which used to be predictive, but are not anymore, are replaced more quickly with new and better hypotheses.

In summary, ELPH consists of three main functions: **hypothesis generation**, **updating** and **pruning**. At the beginning of each trial, the hypothesis generation function creates all the probable subsets of the current content of STM as potentially new hypotheses. To make a prediction about the next observation, the HS is asked to look at all the existing hypotheses with at least a subset of observations recently observed. Generated hypotheses may either exist in HS or may be novel. The ones that already are in HS are used to make a prediction. Among all those, the one with the lowest amount of entropy is selected and among all its predictions the item which has been observed most often is chosen as the prediction of the next observation. After the prediction new data is observed an updating function updates all the related hypotheses according to this new data. If this new observation is consistent with one of the predictions of that hypothesis, the corresponding count is updated; it is incremented by one. If not, this event, which has happened for the first time (c = 1), is added to associated prediction set. Updating function also augments the generated hypotheses which were not in the HS before. The prediction set for those hypotheses would consist of this new event (observation) followed by the event count set to 1. At the end, the pruning function calculates the

entropy of all the existing hypotheses in HS and deletes those with entropy value more than a threshold.

Results

In this section the preliminary performance of ELPH in our RPS task is discussed. To evaluate ELPH performance as a human strategy learning algorithm we played the ELPH algorithm against the same RPS sequences as the participants reported in (Danckert et al., 2012). To compare the participants' behavior to the one of ELPH, we need to select some metric of performance. We used the win rate and learning rate. These two measures are defined in the previous chapter (page 18). The rate of learning estimates how fast ELPH adapts to changes and the max win rate represents the win rate after adaptation. These two measures are selected because, as discussed, RBDs' and HCs' differ for these two measures.

To find the best-matched version of ELPH for our participants we need to find the optimal parameters of the algorithm (n and H_{thr}). Since our participants most likely employ various policies to learn the computer's strategy, we estimate the best parameters for each participant separately. Best parameter sets were computed using maximum likelihood (ML) estimation. According to this method, the optimal parameters for each participant, denoted by $\hat{\theta}$, are those that make the participant's sequence of plays (D) the "most likely" one among all the possible sequences of play (Myung, 2003):

$$\hat{\theta} = arg\,max_{\theta}\ P(D|\theta, ELPH) \quad (3)$$

There are several methods suggested to find $\hat{\theta}$ in ML estimate. Due to the peculiar characteristic of this model not having a specific input-output mapping, traditional optimization methods were not applicable. To overcome this challenge, we searched through a lattice of possible values for the parameters that resulted in ELPH playing the exact sequence of choices that that participant had played.

In the rest of this chapter, I will represent the performance of optimum ELPH and compare its results to our participants' results. Each group of participants is presented separately considering the fundamental differences observed in their results. The final goal is to have a computational model

with similar outcomes to our healthy controls, which we can then modify to match the performance of right or left brain damaged participants

Comparison to healthy participants

Despite optimizing the parameters for each participant, ELPH performance consistently *outperforms* HCs both in terms of max win rate and learning rate. Parameter fits suggested that ELPH, on average, is faster in learning and reaches a better win rate. Table 1 represents the estimated optimal parameters for each participant in healthy control (HC) group. For each participant I associate a corresponding ELPH; the parameter of this ELPH is estimated so as to maximize based on ML estimate. The second column in this table is the optimal value of H_{thr} for the participant with the ID mentioned in the first column (Mean of H_{thr} = 1.57; SD = 0.6) and the third column denotes the optimal STM length, n, for that participant (Mean of n = 1.6; SD = 0.9).

Table 1. Optimal parameters of ELPH for each participant in HC group computed using ML estimation approach. The first column is the number representing each participant. The second column shows the best-matched value of H_{thr} and the third column is the optimal value of STM length.

Participant Number	Best H_{thr}	Best n
1	1.2	1
2	1.2	2
3	1.4	1
4	2.2	2
5	2.6	2
6	2.3	2
7	1.1	1
8	2.2	4
9	1.1	1
10	0.8	1
11	1.3	1
12	1.4	1

Figure 3 indicates the average of optimal ELPH win rate and that of healthy controls'. It is evident that ELPH performance consistently outperforms HCs both in terms of max win rate and learning rate. Not only does ELPH achieve a higher max win rate on average ($F(1,22) = 11.944, p < 0.005, \eta^2 = 0.974$), but it is also faster to learn the opponent's strategy compared to our HCs

($F(1,22) = 10.311, p < 0.005, \eta^2 = 0.319$). A mixed design repeated-measures ANOVA with within subjects factors of block (block 1 to block 10) is used to compare the learning phase of HCs and optimal ELPH. Each block is an average over ten consecutive trials; thus each condition in our task consists of 20 blocks. The same analysis is used to compare the max win rate but this time the within subjects factors are blocks from 11 to 20. As is explained in the first chapter, I assume that learning occurs over the first 10 blocks due to the changing win rate in these blocks and that in the second 10 blocks where learning rates are stable I assume participants are mainly exploiting what has been learned in the learning phase.

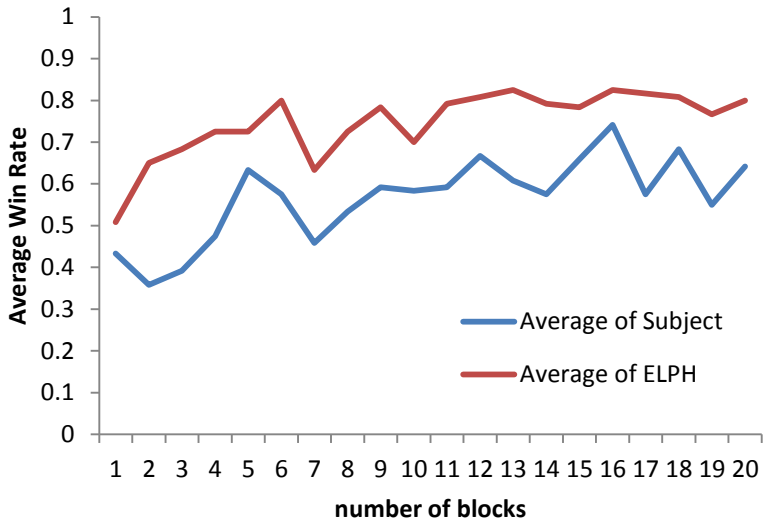

Figure 3. ELPH average win rate and that of healthy controls'. The average win rate of ELPH is calculated as the average of all win rates of the associated ELPHs. The red line in the figure represents the average win rate of ELPH and the blue line shows the average win rate of all the participants in HC group.

I investigate the average behavior of our participants since the behavior of all our healthy participants is quite similar; at least in terms of these measures. This suggests that the average behavior of the group can be a good indicator of the behavior of all the participants. For almost all the HCs, ELPH outperforms the participants. The possible reasons behind this outperformance are discussed at the end of this chapter. For now I will describe the performance of ELPH compared to our patient groups, LBDs and RBDs.

Comparison to patient groups

Comparing ELPH performance with the one of brain damaged participants not only can help us better understand the functional nature of the impairment in these patient groups; but also can enable us to recognize the potential explanations about the difference between ELPH results and the results of HCs.

ELPH plays a maximizing strategy similar to LBD participants

Table 2 shows the optimal parameters of the algorithm for each participant: the second column is the best-matched value found for H_{thr} (Mean of H_{thr} = 1.32; SD = 0.7) and the third column is the one for n (Mean of n = 1.85; SD = 0.9).

Table 2. Optimal parameters of ELPH calculated for each participant in LBD group using ML estimation approach.

Participant Number	Best H_{thr}	Best n
1	1.2	2
2	2.4	2
3	1.2	2
4	1.0	4
5	1.2	2
6	2.1	1
7	1.4	2
8	1.4	1
9	0.4	2
10	1.4	1

Figure 4 demonstrates the ELPH average win rate compared to the average win rate of our LBD patients. Average win rate is calculated the same as for HCs: the average win rate of all corresponding ELPHs. The red line in the figure is the average win rate of ELPH and the blue line is that of LBDs.

It is evident that ELPH is more successful in capturing the behavior of LBDs. ELPH and LBDs reach the same max win rate ($F(1,18) = 2.332, p > 0.1, \eta^2 = 0.155$). The learning rate of ELPH however is faster than LBDs ($F(1,18) = 21.177, p < 0.001, \eta^2 = 0.540$). This result suggests that I might be able to do reverse engineering to examine the possible reasons behind the failure of ELPH in

replicating the behavior of HCs. By reverse engineering, I mean instead of having a model that captures the behavior of HCs, we have a model with an outcome similar to LBDs. Comparing the observed differences between the results of our behavioral experiment for HCs and LBDs can offer some solutions; it can help us to realize what is missing in the model. This possible solution is discussed shortly at the end of this chapter and with more detail in the next chapter.

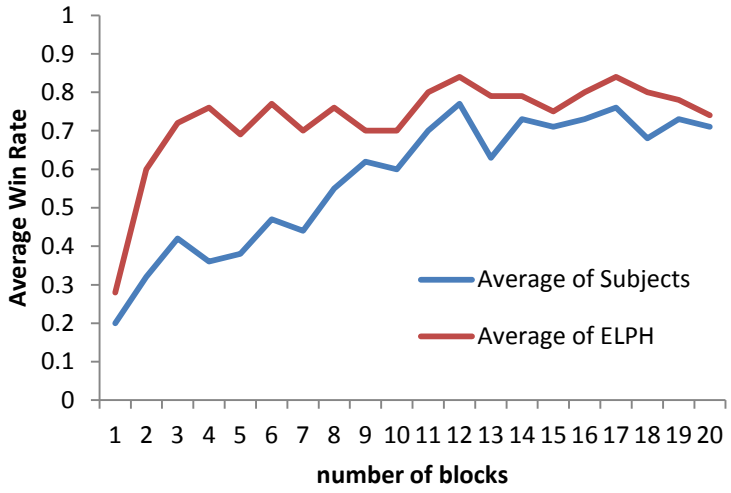

Figure 4. ELPH average win rate and that of LBDs. The red line represents the average win rate of ELPH and the blue line shows the average win rate of all the participants in LBD group.

ELPH cannot replicate the performance of RBD patients

I fit the best-matched sets of parameters for each RBD patient; Table 3. (H_{thr}: Mean = 1.32 and SD = 0.7; n: Mean = 1.85 and SD = 1.18). Figure 5Figure 5demonstrates the ELPH average win rate compared to the average win rate of our RBD patients. The red plot is the average win rate of ELPH and the blue plot is that of LBDs. ELPH performance is significantly different from that of RBDs ($F(1,26) = 42.436, p < 0.001, \eta^2 = 0.620$); it obviously out performs our participants. These results are obtained by employing the best-matched parameters for each associated ELPH. Thus, this result is the best that ELPH can provide in reproducing the behavior of RBDs. While ELPH clearly fails to replicate the behavior of RBDs; our initial focus is on modeling the behavior of HCs. Having a realistic model which can capture the healthy participants' behavior will allow for a more principled approach to reproducing the behavior of RBDs.

Table 3. Optimal parameters of ELPH for each participant in RBD group calculated using ML estimation approach.

Participant Number	Best H_{thr}	Best n
1	0.1	3
2	1.2	1
3	1.4	1
4	2.4	1
5	0.8	3
6	1.0	1
7	0.5	4
8	1.7	1
9	1.3	1
10	2.1	1
11	2.2	1
12	1.4	1
13	1.9	4
14	0.5	2

ELPH performance in replicating the result of RBDs is worse than its performance for HCs in both terms of learning rate and win rate. Thus, there is not any characteristic of the RBDs' behavior that is captured better by ELPH compared to HCs. If that was the case, as it is for LBDs, this result could be employed to improve our model for HCs. Since the result doesn't offer such an option, I do not discuss these results further.

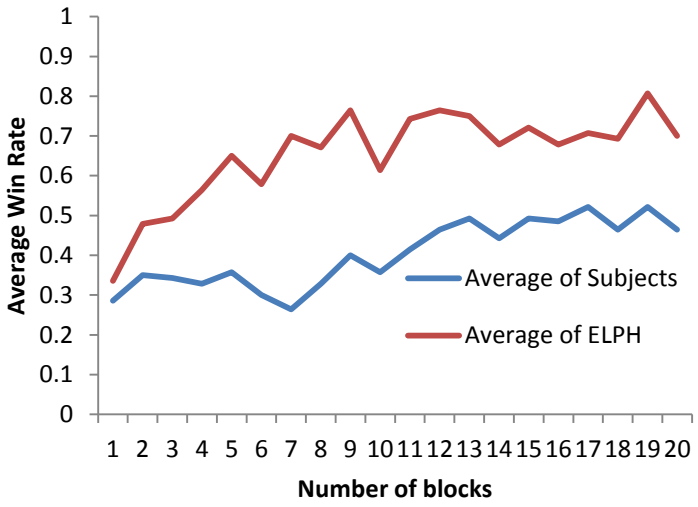

Figure 5. ELPH average win rate and that of RBDs. The red line in the figure represents the average win rate of ELPH and the blue line shows the average win rate of all the participants in RBD group.

Discussion

In this chapter, I described a sequence learning method for an agent that must learn temporal sequences online with limited data. It also must be able to adapt quickly to the variations since the pattern existing in the sequence might change rapidly. These assumptions hold true for our task as well (the RPS game). For this reason, and the limitations of other algorithms, we believe that the ELPH algorithm can be a good first candidate for human decision making. The initial simulations indicate that among all three groups of participants, ELPH results are most similar to LBDs'. ELPH clearly outperforms HCs and RBDs in terms of both learning rate and max win rate.

The failure of ELPH to replicate the results of HCs could be due to two different reasons: (1) there might be some missing components in the ELPH model or (2) there might be some parts of the algorithm that are substantially different from neural algorithms. Both reasons seem likely. While ELPH is inspired by some studies in human decision making, it is not clear if all ELPH's assumptions are applicable for a human learner. Another source of potential problems comes from the fact that there is no examination of whether or not various components of the ELPH algorithm are neurally compatible.

In the rest of this thesis I will investigate these potential problems and possible solutions to them. In chapter 3 I will discuss the first category of problems: what are the missing components? In

chapter 4 I will discuss the second category of problems: the neural plausibility of ELPH's assumptions. Comparing these two issues can shed light on the possible shortcomings of the model and also propose better alternatives.

Chapter 3: A missing component in ELPH?

As discussed in the previous chapter, ELPH fails to replicate the elementary characteristics of the HCs' behavior. I discussed two groups of possible reasons for this failure. The first group includes those suggesting some component might be missing in this method. The second group consists of ones which go further and require more neurally plausible alternative algorithms for some functions in the model. In this chapter I will discuss the first category. Since ELPH is proposed as a sequence learning method, it may not necessarily be limited in the same way as a human brain; it is reasonable to consider the possibility that some constraints in the brain have not been applied to the model. On the other hand, as was shown in the previous chapter, ELPH's results are closer to those of LBDs which supports the idea of an omitted component in ELPH. In this chapter I will describe some of these potential missing compartments and investigate whether adding those can improve our model.

What is missed in LBDs?

As is shown in Figure 4 (the previous chapter), ELPH's performance is similar to that of LBDs compared to the two other groups. ELPH is however unable to capture the behavior of HCs for any feasible set of parameters. This suggests that the reason behind the difference between ELPH's ability to simulate LBDs' and HCs' behavior can be explained, at least in part, by a missing component. This missing component might get clearer by comparing the behavior of LBDs and HCs in our task. In this section, I explain a probable candidate which this comparison offers.

The main difference between LBDs' and HCs' is that the max win rate for LBDs in the strong condition is *larger*. LBDs reach max win rates of around 80% in the strong condition, the maximum win rate possible for our task; for healthy controls, however, average max win rates didn't exceed 70%. In our task, trials are independent of each other and paper always has the higher probability, (80% of the trials), of being chosen by the computer at each trial. Thus, the rational optimal decision is to select scissors all the time in order to maximize the expected pay-off (which is win rate in our task). This seems to be exactly what LBDs do since the only way to reach such a win rate is to select scissors for *all* trials. For this reason I call LBDs "maximizers." To explore this idea, I plotted the proportion of scissors choices for each LBD patient in Figure 6. Almost all the LBDs (except participant 4 whom seems to fail to learn anything) start playing scissors for almost all the trials.

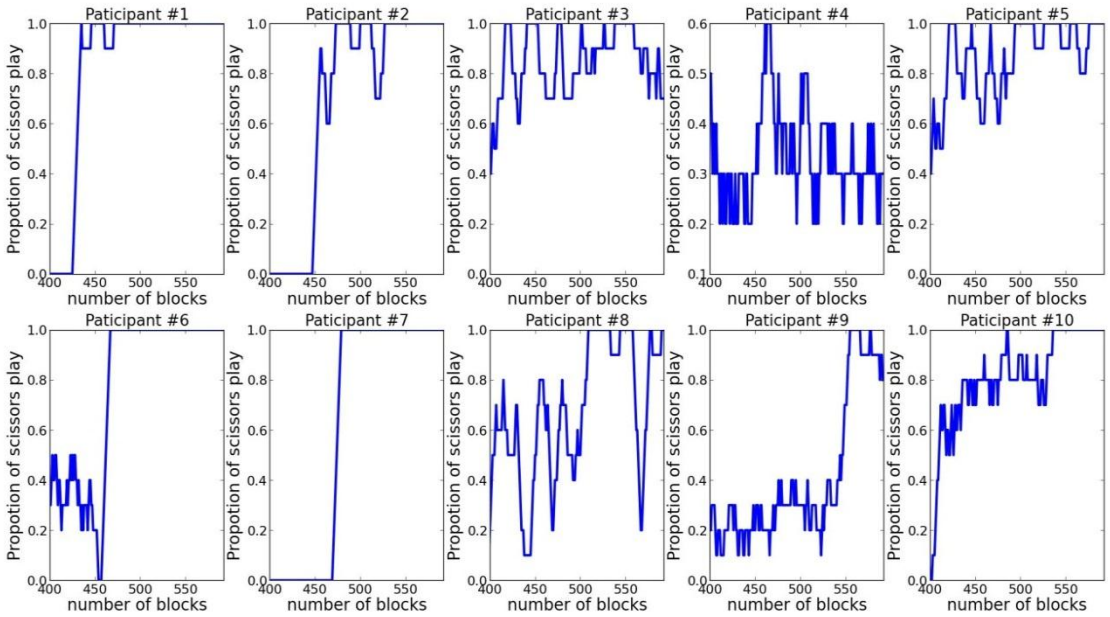

Figure 6. Proportion of scissors being played for each participant in LBD group

Figure 7 plots the percentage of scissors played by HCs, and the graphs show that some of the HCs behave differently from LBDs. While they realized that scissors is the better option to choose (this is evident from the fact that they play scissors more than chance), they decided not to play scissors for all trials at least for a while. This transient, sub-optimal behavior is what is missed in most of LBDs. This behavior is a well-known phenomenon called, "probability matching", confirmed by several experiments in cognitive neuroscience and economic literature (for a review see (Vulkan, 2000)). In the following section I will discuss what has been found so far to explain the rationale behind this phenomenon and how this can be implemented in our model.

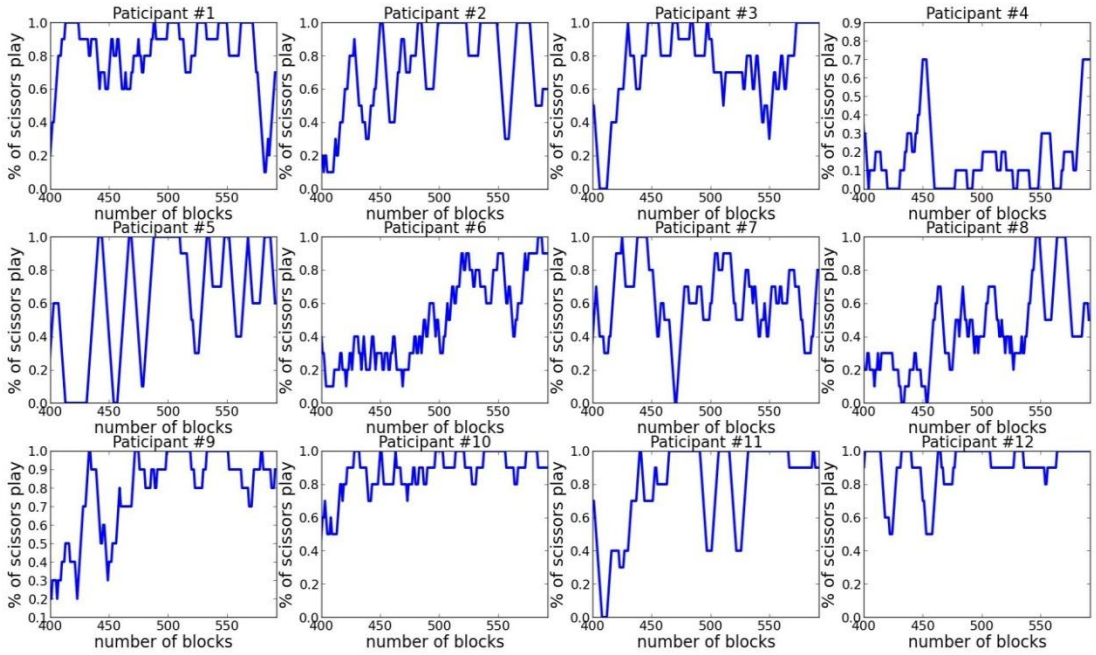

Figure 7. Proportion of scissors being played for each participant in HC group

Probability Matching

In a standard probability matching task, participants are given a repeated choice between two alternatives with different payoff probabilities; one choice always has the higher probability, which is unknown to the participants. Our RPS task is a three choice version of this type of task. Since the outcomes are independent, the optimal decision is to choose the high payoff alternative on every trial. Several studies, however, indicate that healthy participants are not maximizers; they instead match their response probabilities to the relevant outcome probabilities. This suboptimal behavior which leads to lower expected payoff is called "probability matching," and is a stable observation that has been observed in several studies.

The reason behind this behavior is not well understood (Vulkan, 2000). Several hypotheses have been suggested to explain the rationale behind this behavior, ranging from those who believe this strategy is simply a mistake (Koehler & James, 2009), to those who think it is a smart approach (Gaissmaier & Schooler, 2008). Here I only describe one of the main interpretations of probability matching which can be related to our model. The idea is that this behavior is due to humans' effort to find patterns in a sequence of outcomes (Unturbe & Corominas, 2007; Vulkan, 2000; Wolford,

Newman, Miller, & Wig, 2004). In their daily experiences in the world, outcomes often follow patterns; so humans learn to adapt their decision making process. In 2008, Gaissmaier and Schooler demonstrated that those participants that employ a probability matching strategy in a typical probability matching task are more likely to discover the patterns that exist in non-random sequences. Similarly, Green and colleagues in 2010 showed that an optimal Bayesian model-based learner with incorrect initial beliefs about the process can also show the probability matching behavior (Green et al., 2010). They believe that humans are optimal learners and this behavior is only a result of their wrong prior beliefs about the task: the assumption of temporal dependency of outcomes. Green and colleagues also argue that this incorrect belief is more plausible for humans than the independency assumption since most of the time humans are exposed to events with temporal dependency.

The Solution

Regardless of the real reason behind this behavior, what is important is to model this behavior mathematically. Based on the "probability matching" hypothesis, the probability of selecting each event matches the probability of relevant outcomes. This decision rule is known as the 'soft-max' rule (Thrun, 1992) and is mainly employed as an action selection rule in most reinforcement learning based methods. With soft-max, the probability of choosing each action is determined on the basis of that action's relative expected value. This idea is proposed against the "greedy rule" in which the chosen action is always the most probable one. If the value of action a_i is denoted by $V(a_i)$ where V refers to the value function and i=1,... m where m is the total number of possible actions, the most common soft-max method defines the probability of choosing action a_i as follow:

$$P(a_i) = \frac{e^{V(a_i)}}{\sum_{j=1}^{m} e^{V(a_j)}} \qquad (4)$$

This method still favors the option with higher probability; but it also provides a chance for other options to be tried in proportion to their values. This rule is suggested as a solution for the exploration-exploitation trade-off in reinforcement learning (Cohen, McClure, & Angela, 2007). In reinforcement learning, as will be discussed in chapter 4, the leaner must learn while it is acting. To maximize the payoff (known as expected reward in reinforcement learning literature), he must always select the choice with the highest outcome (exploitation), but since learning happens online, this optimal choice must be learned by exploring all the possible choices. Soft-max rule is a balance

between exploration and exploitation processes. This decision rule has been employed in several models suggested for human decision making. In one of the most related ones, Daw and colleagues in 2006 studied the effect of different decision-rules for their model. Their results demonstrate that their model with soft-max provided the best fit to human behavioral results in an n-arm bandit task (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006).

The result of applying this exploration method to ELPH is shown in Figure 8. I refer to this version of ELPH as soft-max ELPH. All other parameters and algorithms remain the same as those used in the previous chapter. Soft-max rule is applied at two different levels: selecting the hypothesis to predict subsequent observations and also selecting the best event of the chosen hypothesis as the next prediction. In the previous version, the hypothesis with the least entropy was always selected for prediction. In this version a soft-max rule applied over all related hypotheses, Hyp_is, with entropy values, H_is, to determine the probability of selection:

$$P(Hyp_i) = \frac{e^{-H_i}}{\sum_{j=1}^{m} e^{-H_j}} \qquad (5)$$

The minus sign reflects the fact that the lowest entropy must have the highest chance to be chosen. Previously, the prediction of a particular hypothesis, *Hyp*, with the prediction set of {[e_1, c_1], [e_2, c_2]…[e_m, c_m]}was e_j with the maximum c_j where j=1,…m. In this version, the probability of selecting each event is again calculated based on soft-max rule:

$$P(e_i) = \frac{c_i}{\sum_{j=1}^{m} c_j} \qquad (6)$$

I omitted the exponentials in this soft-max formula. The reason is all the c_js are positive numbers; thus there is no need for an exponential function. This also keeps the formula simpler, although it changes the results compared to the case of soft-max rule with exponential function. In our case, removing the exponential function favors the exploration phase even more than the soft-max rule with exponential.

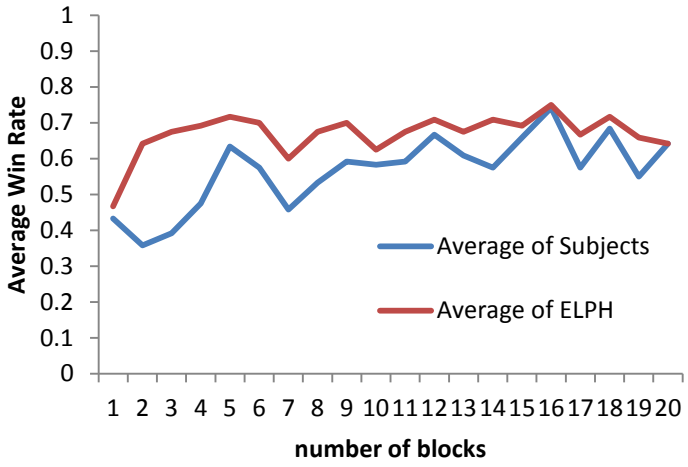

Figure 8. Average win rate for Soft-max ELPH and that of HCs. The average win rate of soft-max ELPH is calculated as the average of all win rates of the associated soft-max ELPHs. The red line in the figure represents the average win rate of soft-max ELPH and the blue line shows the average win rate of all the participants in HC group.

As is evident from the graph in Figure 8, soft-max ELPH's max win rate drops to that of HCs. The max win rate of soft-max ELPH and HCs are compared employing a mixed design repeated-measures ANOVA. The result revealed that the difference is not statistically significant ($F(1,22) = 1.135, p > 0.1, \eta^2 = 0.049$). The learning rate of soft-max ELPH, however, remains faster compared with HCs ($F(1,22) = 6.289, p < 0.05, \eta^2 = 0.222$). Putting these together suggests that while this adaptation improves our model's capability to model our normal participants' patterns of performance; the difference between these two is more than merely difference in action-selection rules. The other probable missing components will be described in the following section.

What about Limitations?

As was mentioned earlier, some of the assumptions in ELPH may not be reasonable considering the well-known restrictions observed in the brain. One of the most obvious of humans' limits is capacity limitations to store and process information or so-called memory restrictions. Memory also has a critical role in our method; STM and HS can both be considered as memory compartments in which the former stores the most recent observations and the latter records the most recent and related hypotheses. The only memory constraint applied to this method is the number of hypotheses in HS.

The pruning function is defined to remove the inconsistent hypotheses to restrict the number of hypotheses. It is, however, not clear if this limitation is sufficient. Several studies demonstrating limitations in different types of memory systems have been identified. In this section of the thesis, I review some of these limitations and investigate how they can be added to our model. It is important to note that memory on its own is a large topic in cognitive neuroscience and a systematic review is beyond the scope of this thesis. What is described in this section are only very basic and prominent limitations which have been reported frequently.

Psychologists recognize at least three different kinds of memory: working memory, short term memory and long term memory. Long term memory refers to memory of knowledge and prior events. This type of memory is recognized to be structurally and functionally different from short term memory, which is a temporary storage that can hold a limited amount of information in a very reachable state for a very limited period of time. Working memory however is not completely distinct from short-term memory. Working memory was introduced by Miller and colleagues (G. A. Miller, Galanter, & Pribram, 1960) to refer to a construct that keeps task-relevant information in an active state for behavioral guidance. This type of memory and attention (refers to the mechanism in which the information in the environment is processed selectively) are two very close concepts which work together to guide behavior. In this sense, working memory differs from the passive nature of short term memory which is merely a transient storage of a limited amount of data. Based on this definition, a STM component in our method can be considered as a simple model of working memory. This component keeps track of recent observations to guide our decision making process to the related hypotheses. HS on the other hand is a really basic model of short term memory.

There are several behavioral tasks and neuroimaging studies that indicate limitations of both short term memory and working memory (Callicott et al., 1999; Cowan, 2008; Todd & Marois, 2004; Marois & Ivanoff, 2005). It is however important to mention that since the differences between working memory and short term memory are still not clear, conducting a task that can study one while excluding the other is not very straightforward. For the rest of this section I describe some of the more famous and less controversial results.

Working Memory

As previously mentioned working memory is a structure representing the ability to transiently store information to be manipulated online and be used to guide cognitive processes (Baddeley, 1986; Goldman-Rakic, 1996). A key aspect of working memory is its capacity limitation, usually measured

in cognitive tasks that attempt to decrease performance by increasing working memory load (Miller, 1956; Fuster, 2008; Shallice, 1988; Just & Carpenter, 1992). There are several measures employed to examine working memory capacity, including working memory span tasks. One of the most popular measures, of particular interest to us because of the similarity to our task, is the n-back task.

In a typical n-back task, as is shown in Figure 9, the participants are asked to monitor a series of stimuli and specify if the current stimulus matches the one presented n-trials before (Gray & Braver, 2002; Jonides et al., 1997; Owen, McMillan, Laird, & Bullmore, 2005). These stimuli can be anything: different alphabets (Figure 9.A); various shapes (Figure 9.B) and even a fixed object with an altered spatial location (Figure 9.C). For example, Figure 9. A demonstrates a verbal 3-back task in which subjects are required to indicate if the alphabetic character appearing in the current trial is the same as the one that appeared 3 trials previously.

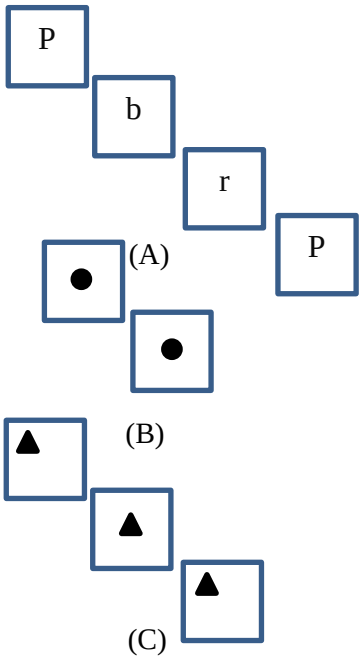

Figure 9. Schematic representation of three different types of n-back task: A. Verbal 3-back, B. Object 1-back and C. Spatial 2-back.

Increasing n in this task increases the cognitive load and consequently decreases the working memory performance. The result of different variations of this task demonstrates that increasing n from 1 to 3 drops the performance rapidly (McElree, 2001; Owen et al., 2005). Although this number can vary from one participant to another, generally speaking it is widely accepted in the literature that

for n greater than 3, the cognitive load is too high, making it impossible for working memory to accomplish the task.

These results are relevant for our task since they suggest that humans' performance drops dramatically when they are asked to keep more than 3 consecutive observations in their working memory. Comparing the paradigm of this task and our model, it can be inferred that there is a high correlation between the capacity of STM in our method and the value of n in an n-back task. In order to find the similarity between the stimulus n trials ago and the current one, participants are required to keep n observations in their working memory. Therefore this result suggests that assuming any value greater than or equal to 4 for STM length, which also is denoted by *n* in our model, is not feasible. A simple way to apply this constraint to our methods is to limit the possible values of *n*, STM length, in our optimization procedure to the appropriate range. Based on the result found for the n-back task, the suitable range for *n* seems to be from 1 to 4. This limitation was already applied to the optimization process which leads to the optimum parameters in chapter 2. Thus these results demonstrate that this restriction of the variation of *n* is not enough since ELPH outperforms our HCs.

Short Term Memory

Short term memory keeps a small amount of data easily accessible over a short period of time. The duration and capacity of short term memory are believed to be limited. HS which contains individual hypotheses is assumed to be a simple representation of short term memory in the brain. Therefore, both duration and capacity constraints must apply to this compartment. In the rest of this part, each of these limitations and their mathematical equivalents in our model will be discussed briefly.

Duration of short-term memory

One of the main streams of research in short term memory shows that its content decays over time due to its limited capacity. It has been shown in several studies that humans remember less and less of their observations as more and more time passes (Jenkins, Earle-Richardson, Slingerland, & May, 2002; Rose, Feldman, & Jankowski, 2004; Ebbinghaus, 1885). However, in our model there is no memory declination. The prediction set of each hypothesis contains the exact number of times that an event follows the observation list of that hypothesis. As long as that hypothesis stays in HS, there is no decline in what has been observed and therefore the exact number of occurrences of each prediction is remembered. This unfeasible assumption might be a possible reason behind the

outperformance of ELPH. To add the forgetting compartment we are required to search for mathematical formulas offered to functionally describe this process.

To be able to functionally describe forgetting, some memory performance measures such as recall or recognition are computed as a function of time. Interestingly, the form of these various forgetting functions is found to be similar (although the time scale might be different from one measure to another); a nonlinear course with an initial rapid decline followed by a long, slow decay (Wixted & Ebbesen, 1991). This apparent similarity motivated researchers to search for a so-called general mathematical law of forgetting. This mathematical relationship between time (*t*) and performance measure (*P*) however is still under debate. Some believe that it follows the power law, $P = At^{-B}$ (J. R. Anderson & Schooler, 1991; Wixted & Ebbesen, 1991; Wixted & Ebbesen, 1997) while others believe it can be described by an exponential function, $P = Ae^{-Bt}$ (R. B. Anderson & Tweney, 1997). In both cases *A* and *B* are parameters of the model. Figure 10 shows both power and exponential function as two models of the performance. As is evident from the figure, the difference between these functions is not major, at least for the purpose of our model. Since the power law is employed in the computational architecture that Anderson and colleagues suggest as a model of human cognition, we also choose this formula to model memory decay in our model.

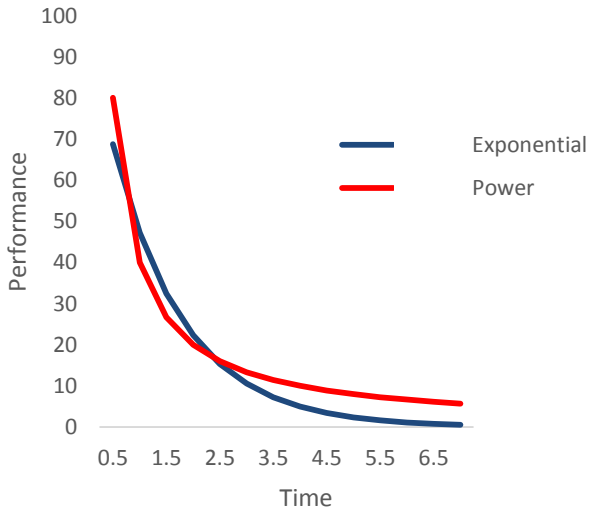

Figure 10. Power and exponential functions plotted against time. The values of A and B parameters are set to 100 and -0.75 for the exponential function and are equal for 40 and -1 for the power function.

To model forgetting I apply the exact idea that Anderson and colleagues suggest (J. R. Anderson & Matessa, 1997; J. R. Anderson & Schooler, 1991). In their model for each chunk of information stored in the memory compartment, a measure called "activity" is defined. A chunk refers to a unit of information stored in memory. The activity of a chunk reflects how frequently that chunk has been retrieved during its lifetime. When requesting a retrieval from memory, among all the related chunks the one with higher activity is retrieved (J. J. R. Anderson & Lebiere, 1998). This activity is calculated by this formula:

$$A = ln\left(\sum_{i=1}^{n} t_i^{-d}\right) \quad (7)$$

where *A* refers to activity level of that chunk, t_i is the time lapse since that chunk was observed for the *i*-th time; *n* is the total number of occurrence of that chunk during the its lifetime and finally *d* is the parameter governing the rate of decay. Anderson and Schooler believe that this formula captures both the power law of forgetting (Rubin & Wenzel, 1996) and power law of learning (Newell & Rosenbloom, 1981). In a power function the effect of learning is related to elapsed time (or the number of presentations), where the amount of time is raised to some power (the "*d*" in equation (7) above; J. R. Anderson & Schooler, 1991). The parameter *d* adjusts how fast this activation decays. Anderson and Schooler argued that the value of *d* is always set to 0.5 because they believe this fits nicely with the human data.

Similarly I define an activity for each prediction of a hypothesis. An activity of a prediction reflects the activity of that event during the lifetime of that particular hypothesis. This definition of activity is more biologically plausible compared to simply saving the count of number of times. Thus in our new model of ELPH, instead of recording the count of number of times that each event happened, the activity of that prediction is saved. For example, the prediction set of *Hyp*, which used to be in the form of $\{[e_1, c_1], [e_2, c_2]\ldots[e_m, c_m]\}$, will be in the new form of $\{[e_1, A_1], [e_2, A_2],\ldots,[e_m, A_m]\}$ in which *A* for each event is the activity of that prediction and calculated according to equation (7).

This new version of ELPH is a modified version of soft-max ELPH with a forgetting mechanism added. The new parameter of this new version, in addition to *n* and H_{thr}, is *d*, the decay rate. Although the *d* parameter is assumed to be fixed in Anderson's model, we decided to find the optimal value of this parameter for each healthy participant separately. It is important to note that to facilitate the optimization procedure, it is assumed that the best values for two other parameters, *n* and H_{thr}, are still

the ones found in the previous chapter. Again maximum likelihood is employed to estimate the optimal values of this new parameter.

Figure 11 shows the average win rate for the last version of ELPH and healthy controls. Same as soft-max ELPH, the max win rate is the same for both ELPH and HCs ($F(1,22) = 1.209, p > 0.1, \eta^2 = 0.052$) and ELPH is significantly faster in learning the computer's strategy compared to HCs ($F(1,22) = 7.135, p < 0.05, \eta^2 = 0.245$).This result suggests that adding this component did not solve the problem of outperformance for ELPH. In some cases ELPH's performance is actually nominally better than the original version. The reason might be due to the fact that forgetting in a non-stationary environment is actually helpful to adapt faster. This will be discussed in the discussion section.

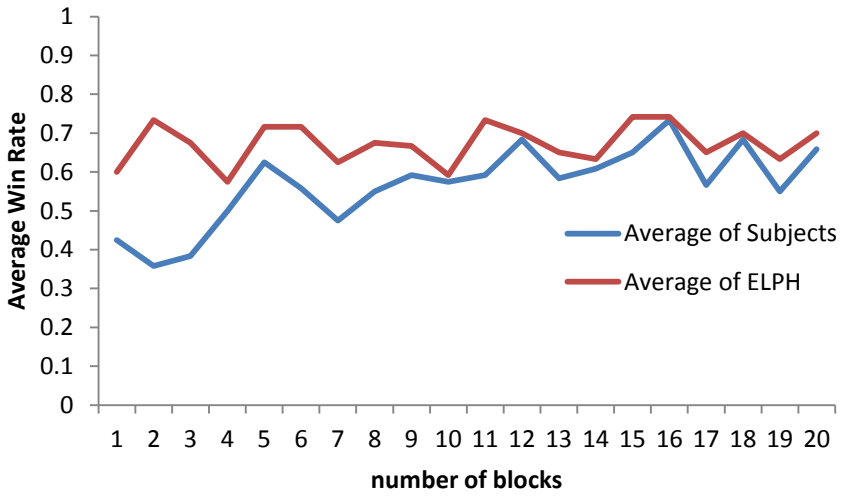

Figure 11. Average win rate for soft-max ELPH with forgetting compartment and that of HCs. The red line in the figure represents the average win rate of this version of ELPH and the blue line shows the average win rate of all the participants in HC group.

Capacity of short-term memory

The idea of a limited capacity for short term memory goes back to 1956 when Miller introduced the magic number of 7 ∓ 2 as the capacity of short term memory (G. A. Miller, 1956). He reported that people could, on average, keep around seven chunks of information in their short term memory. It is

still not clear if storage capacity limits are something different from the timing limits discussed in the previous section. Even if we accept that these are two different concepts, it is not obvious that there is a single capacity limit or whether capacity limits are task dependent (Alvarez & Cavanagh, 2004). For ELPH I borrow the idea that the size of short term memory can be limited and try two approaches to implement this limit.

There is already one mechanism in ELPH to remove poorly predictive hypotheses, which is the pruning function. It calculates the entropy of all the hypotheses and deletes the one with entropy more than H_{thr}. As discussed already, I was not able to find any value for this parameter for which ELPH can replicate the result of HCs.

Deleting hypotheses based on their entropy is only one possible way to restrict the number of hypotheses. Another method is to restrict the size of HS, since the purpose of short term memory is to store information for a *short* period of time (seconds to a few minutes). Longer retention demands other techniques, one of which is rehearsal. By repeating information it is reintroduced to short term memory. Therefore it seems that the hypotheses which have not been observed or repeated within a few trials might be forgotten. This is another method for limiting the size of HS in our model: the deletion of hypotheses that have not been used for more than a specific number of trials. If a particular sequence of observations has not been observed for many trials, its entropy remains unchanged. However, a sequence of observations that is no longer observed, is probably not a useful one, and may well deserve to be forgotten.

To explore this idea we repeated our simulations with an additional parameter, T_{thr}. T_{thr} determined a time threshold for how long ELPH could keep a hypothesis without seeing further examples. For example, T_{thr} = 2 means that if ELPH does not use a specific hypothesis for two successive trials, it deletes that hypothesis. Figure 12 shows ELPH simulations for the optimal values of T_{thr} determined for each participant separately. Again, for the sake of tractability I used the optimal values of n and H_{thr} computed previously. The best values for d and T_{thr} are estimated based on ML estimate. Soft-max decision making rule and forgetting mechanism are also applied for this version of ELPH. As is clear from this plot, adding this parameter makes our plots more closely approximate each other. This change however is not significant in any sense; the max win rate is not surprisingly the same ($F(1,22) = 0.094, p > 0.1, \eta^2 = 0.004$) and learning rate is still significantly different ($F(1,22) = 6.395, p < 0.05, \eta^2 = 0.225$). This result indicates that although all these limitations implemented in our model improves our model; the problem of learning rate might be something different.

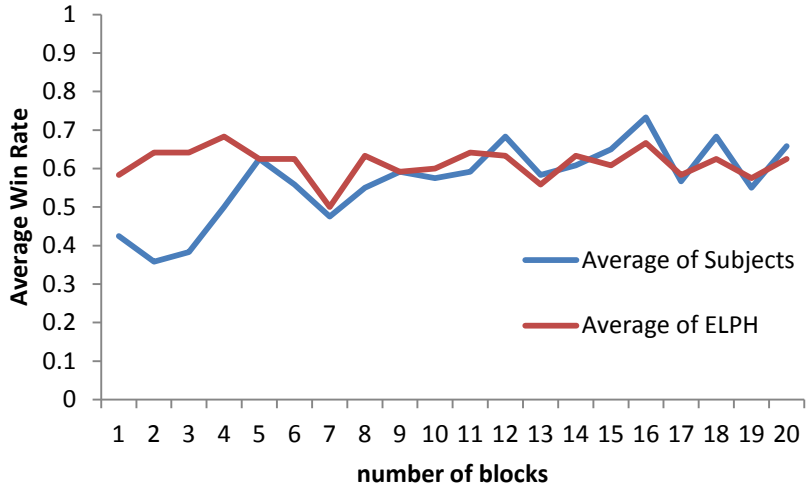

Figure 12. Average win rate for the last version of ELPH (including forgetting, soft-max and HS capacity limit)and that of HCs. The red line in the figure represents the average win rate of this version of ELPH and the blue line is the one of all the participants in HC group.

There are other ways to limit HS size, such as restricting the number of hypotheses. If the number of hypotheses exceeds a certain threshold, then some must be removed. This technique and several other ways have been examined in our model and are not discussed here, because all lead to the same basic result (Figures 11 & 12). Implementing a system of short term memory decay does not improve the resemblance between HCs and ELPH; it sometimes even worsens it. Thus, it is unlikely that STM components are the limiting factor in the ELPH model.

Discussion

Two main families of potential problems have been detected for the ELPH failure in replicating the result of HCs. A potential missing component -the first category- is discussed in this chapter. A soft-max component and a forgetting mechanism are the two missing components resulting from our investigation in this chapter. The soft-max component is suggested after comparing LBDs' and HCs' strategies in our task. Healthy people employ a soft-max decision rule to balance between exploration and exploitation processes. Since they are typically exposed to outcomes with some sort of temporal dependencies, our participants keep exploring for a hypothetical pattern that might exist in the computer's strategy in our task. Although they realized that scissors has been played more than two

other options, they assumed that this is due to a hypothetical pattern that they have not yet learned. This might explain the reason behind the suboptimal behavior observed in HCs but not LBDs.

On the other hand, the forgetting mechanism is a result of applying memory constraints to our model. Two main types of memory limitation are outlined in this chapter: working memory limit and short term memory limit. The equivalent of working memory limit in our model is restricting the range of variation for STM length which was already applied in our model in chapter 2. Short term memory limit is modeled as a forgetting mechanism in ELPH. This version of ELPH still outperformance HCs in terms of learning rate. The result demonstrates that adding the forgetting mechanism not only did not make ELPH any slower to learn, but also it makes it a bit faster. A possible explanation for this observation is that forgetting makes adaptation faster. In change- points, the previous experiences are not only not helpful anymore; they even make the updating slower.

Although adding these components help us to improve the model in terms of capturing our HCs behavior, it does not solve the outperformance problem completely. Win rate and learning rate are the two main characteristics of the participants' behavior considered for comparison in this thesis. ELPH outperforms our HCs in both senses: it is faster and reaches a higher win rate. The problem of win rate is solved by adding a soft-max decision rule; however, the learning rate problem still exists. This result suggests that although there might be some compartments which we are still missing, it is most likely that some functions in the model have not been modeled properly. This requires us to investigate the second group of potential problems. This group which is discussed in the next chapter argues that some functions in this model might not be neurally plausible. It does not mean they are incorrect; it only reflects the fact that our brains might work differently in that sense.

Chapter 4: Reinforcement-Learning ELPH

In our RPS studies ELPH fails to fully replicate the characteristics of HC participants' behavior. Two groups of solutions are suggested. The first group, which discussed the possibility of a missing component in the model, was examined in the previous chapter, and while it improved the resemblance between ELPH and HCs there remained important discrepancies. In this chapter I describe the second family of solutions, which suggest more neurally plausible alternative mechanisms for some functions in ELPH. It is important to note that all changes applied to ELPH in the previous chapter still exist in the new version of ELPH offered in this chapter.

ELPH's learning process consists of generating hypotheses, updating them and selecting the most suitable one for prediction. The general idea of making hypotheses about the task and updating them based on the incoming information seems consistent with what humans do in similar situations. What is not clear is whether each of these functions is compatible with our knowledge of humans' learning mechanisms and their quantitative limitations. To investigate this idea I briefly review neural learning mechanisms and introduce the basic mathematics behind these mechanisms in the first part of this chapter. Comparing the brain areas that support human learning to the brain areas injured in RBDs which were also discussed in the first part can help us decide the priorities for our modeling. In the second part of this chapter I describe the new version of ELPH based on the review from the first part.

Learning mechanisms in the brain

Our brain is known to have multiple learning mechanisms that parallel the learning algorithms developed in machine learning. Investigating these learning methods from both the mathematical and neural points of view can be a great help for evaluating the ELPH model. In this section I describe three main learning algorithms in machine learning which have also been shown to be brain learning mechanisms.

Machine learning is a branch of artificial intelligence that deals with the learning problem. The main question is how one builds a system that is capable of learning optimal actions from data. One approach to learn the optimality is to define it externally to the learner. In this approach the learner receives guidance about the optimal action from outside. The amount of guidance needed for learning differs from one situation to another. In fact various families of learning methods in machine learning are defined based on this measure: how much information is fed into system to guide learning. This

permits dividing up learning methods into three groups: “supervised learning”, “unsupervised learning” and “reinforcement learning”.

In supervised learning, the optimal action, known as desired output in this literature, is accessible for the leaner. Given some training examples, the system must learn the functions mapping input to desired output. For most methods in this family, the difference between the actual and this desired output is calculated as an error signal that is then used as a training signal (Doya, 1999). The system trains to minimize some measure of this error signal. Having the error signal is a great help for learning; but unfortunately in most real-world problems, having access to the desired output of the system is not feasible. Therefore supervised learning is limited in application. Reinforcement learning and unsupervised leaning were developed as alternatives for these situations. The assumption of reinforcement learning algorithms is that while the desired output is not available there is an accessible training signal in the form of a scalar reward signal. This reward signal does not *explicitly* say anything about the optimal behavior; it *implicitly* tells the system how good its action was. Finally, in unsupervised learning, there is no teaching signal at all. The system does not receive any feedback. The only information given to the system is the input data. The system has to learn without any guidance about either the desired output or any feedback about performance.

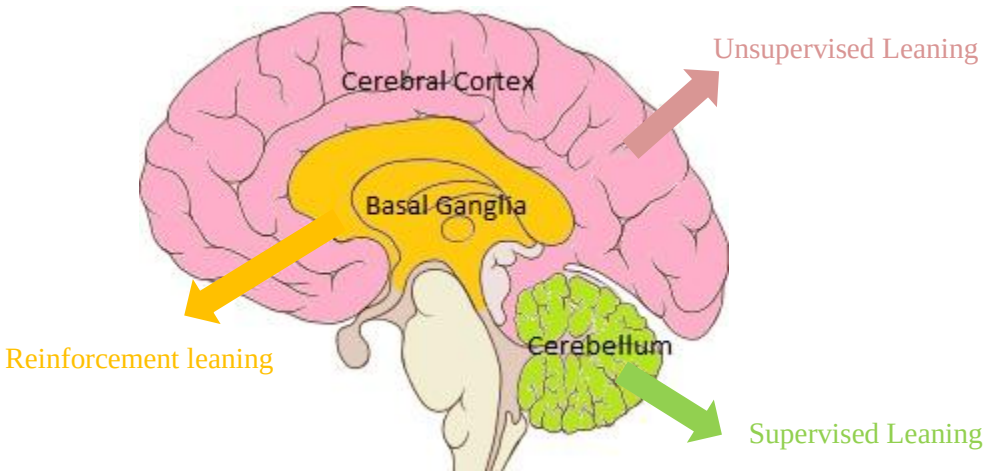

Figure 13. The schematic representation of the different learning mechanisms in the brain

Evidence for all these three learning mechanisms has been observed in different brain areas (Figure 13). Cerebral cortex seems to be most involved in unsupervised learning. There is significant

evidence showing the involvement of basal ganglia in reinforcement learning, and cerebellum is known to be specialized for supervised learning (Doya, 1999; Doya, 2000; Houk & Wise, 1995). For the rest of this section I will briefly discuss each learning mechanism and their relation to specific brain areas.

Supervised learning

Theory

In supervised learning the desired output of the system is given for a set of training data. Training data is a collection of input-output pairs in the form of (*x*,*y*) where *x* is an input sample and *y* is the desired value of output for *x*. The task is to infer the mapping function from input to output from this sample data. To train the system, an error signal is calculated based on the difference between the desired output and the real output of the system (Figure 14). If $y(t)$ is the desired output at time *t* and $\hat{y}(t)$ is the real output of the system at that time, then the error signal can be calculated as:

$$error = |y(t) - \hat{y}(t)| \qquad (8)$$

This error signal is then back-propagated to the system and utilized as the teaching signal. Most of the time the objective is to minimize the sum of squared errors at the sample data points:

$$E = \sum_t \|\hat{y}(t) - y(t)\|^2 \qquad (9)$$

More details about this variety of learning can be found in (Doya, 1999). For our purposes, what is essential is that this form of learning needs training examples provided by a knowledgeable external supervisor.

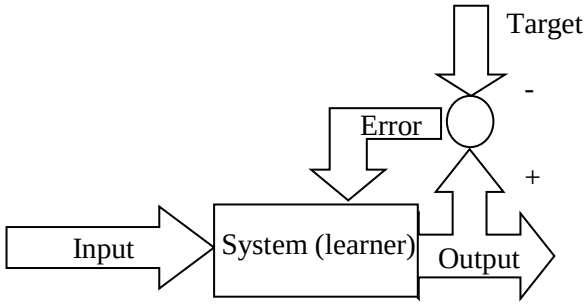

Figure 14. Supervised learning with error signal back-propagating to the system as a training signal

Neuranatomy of Supervised Learning

The cerebellum is traditionally known to be responsible for the control of movement (Stein & Glickstein, 1992; Glickstein, 1992; Ito, 2000; Paulin, 1993). This brain area is responsible for adapting and modifying movements in the presence of sensorimotor perturbations. Adaptation allows humans to improve their motor commands based on the errors from previous efforts. The cerebellum is the brain area that predicts the expected outcome of the motor commands. The difference between actual and predicted outcome is the error signal that is used as a teaching signal. This signal changes the synaptic weights between the posterior parietal cortex (PPC) and M1 and in this way adjusts the movement function (King et al., 2013).

Purkinje cells are the only output neurons in the cerebellum. They receive input from two different sources: climbing fibers and parallel fibers. The error signal, described above, is provided by the climbing fibers to Purkinje cells. These fibers code the reaching error at the end of each movement (Wolpert, Ghahramani, & Flanagan, 2001). Figure 15 shows the schematic representation of cortico-cerebellar connection. Parallel fibers carry the information sent from rest of the brain and mainly cerebral cortex. The error signal guides the learning process occurring in cerebellum by modulating the response of Purkinje cells to the information from the cerebral cortex.

New data show that the cerebellum is involved in more than motor tasks (Allen, Buxton, Wong, & Courchesne, 1997; Ivry & Baldo, 1992; Kim, Ugurbil, & Strick, 1994). The massive connections between cerebellum, basal ganglia and cerebral cortex especially frontal and parietal cortices are one possible route for the contribution of cerebellum for cognitive functions (Ramnani, 2006; Bostan, Dum, & Strick, 2013). It however is not very clear how cerebellum might be involved in such functions. One idea is to generalize the supervised learning computation in cerebellum to more than motor sequence learning (Doya, 1999).

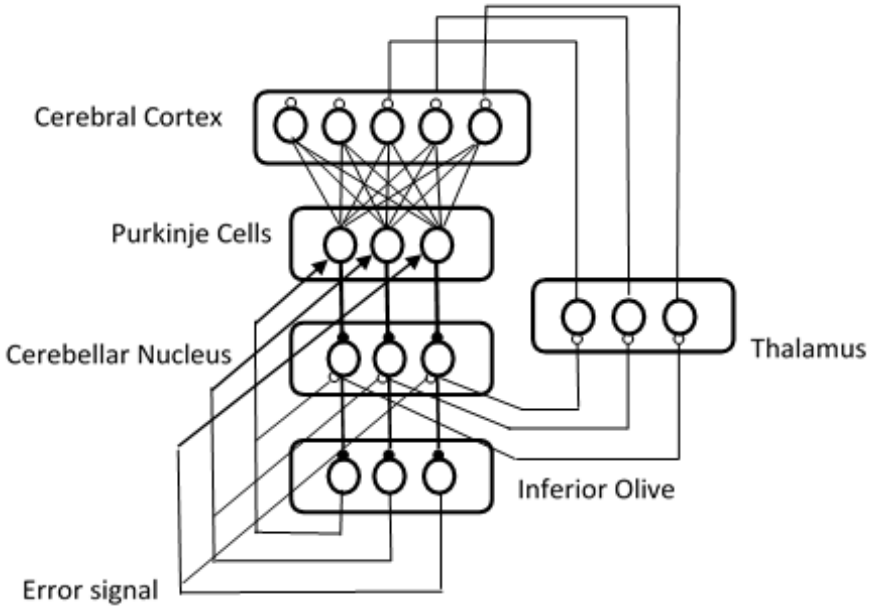

Figure 15. The cortico-cerebellar connection in the brain suggested to be involved in supervised learning in brain

Reinforcement Learning

Theory

For many real-world problems access to input/output pairs is not available. Reinforcement learning relaxes this need. The feedback signal is provided as a scalar reward signal (Figure 16). Learning from this reward feedback is harder than the error signal in supervised learning. This feedback signal represents how good the action was; but it does not say whether this action was the most rewarding one or which action among all the possible actions was optimal. This reward value, given at the end of each trial, also might be influenced by not only the action taken at that trial but also the actions from earlier trials.

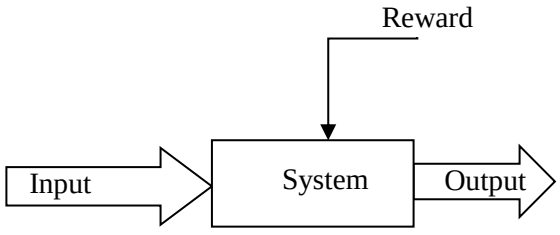

Figure 16. Reinforcement learning with only a reward signal as an external feedback

As the reward signal is the only obtained feedback for a reinforcement-learning learner, the optimal sequence of actions, called the optimal policy, is defined as the one that maximizes the total expected reward in the long run. In the simplest case this expected return is defined as:

$$E\{R_t\} = E\{r_{t+1} + r_{t+2} + r_{t+3} + ... + r_T\} \qquad (10)$$

in which $r_{t+1}, r_{t+2}, r_{t+3}$ … is the sequence of rewards received after time step *t* and *T* is a final time step. *E* refers to expected "return" which is denoted by *R*. This means the leaner must learn the policy which maximizes the expected return which in the simplest case is defined as the total reward that the leaner can receive starting from time step *t* till the end at time step *T*.

Policy in RL corresponds to the stimulus-response rule in psychology. It defines which action should be taken at each state of the environment; it maps the states of the environment to actions to be taken. To find the optimal policy, a value function is defined over all possible states of the environment. The value of a state ($V(s)$) is the expected return an agent can expect to receive over the future starting from that state of environment:

$$V(s) = E\{R_t | s_t = s\} \qquad (11)$$

V in this formula refers to the value of the state *s* and s_t refers to the inferred state of the environment at time step *t*. Since the value of each state is determined by the total expected reward from then on and reward values are dependent on the actions taken in each state, the value of each state depends on the actions taken starting from that state. This means the value of each state varies from one policy to another. The optimal policy is the sequence of actions that brings the highest possible value or expected return for each state of environment.

In summary, the optimal policy is the best sequence of actions at each state of the environment. The leaner must find these actions by maximizing the value of its current state. There are several approaches for estimating the value function. The most important one is temporal difference (TD) learning. In brief, TD starts with a first guess about the value function. At the end of each trial, the error signal is used to improve the estimation of the value function:

$$\delta(t) = the\ new\ estimation\ of\ valuefunction - the\ last\ estimatation \quad (12)$$

This error signal, known as TD error, is the difference between the current estimate of the value function and the estimate before it received the last reward value. If the prediction is perfect, this error signal should be zero and thus no updating is required. Any value other than zero shows an inaccuracy in prediction and a need to improve the estimation. This error signal can be used in various forms for learning. In the simplest case, called delta rule updating, the updating rule obeys the formula:

$$V(s_t)_{new} = V(s_t)_{old} + \alpha\delta(t) \qquad (13)$$

α in this formula is the parameter that determines the learning rate. This updating rule will be discussed with more detail in next section of this chapter.

Neuroanatomy of Reinforcement Learning

Reinforcement learning has been widely studied in the brain. The most important neural participants in RL learning are the dopaminergic neurons of the midbrain (Schultz, Apicella, & Ljungberg, 1993; Schultz, 1998). While the traditional belief about dopaminergic neurons had been that they respond to the reward signal (Wise, Spindler, & Gerberg, 1978; Wise, Spindler, & Legault, 1978); Schultz and colleagues demonstrated that the coding pattern of these neurons in the monkey brain matches perfectly the TD error signal. They recorded the activity of midbrain dopamine neurons of awake monkeys during simple Pavlovian tasks and instrumental tasks. They showed that although the neurons began responding to the presence of the unconditioned stimulus (see Doya, 2007; Figure 4, top row); when the presence of the stimulus was predictable by way of a cue stimulus the response of these neurons to the unconditioned stimulus vanished. Instead the midbrain dopamine neurons started to respond to the cue (see Doya, 2007; Figure 4, middle row). When reward was omitted, the firing rate of dopaminergic neurons became suppressed (see Doya, 2007; Figure 4, bottom row). This reward predicting response pattern was a surprise since it matched perfectly the characteristics of the TD error signal.

In the beginning of the learning process, receiving reward after the unconditioned stimulus is not "*predictable*". This unpredictability means this reward association has not been learned and thus there

is a TD error. Correspondingly dopaminergic neurons show burst firing in this phase of learning. Later on, this association is learned and therefore no dopaminergic neurons fire. In this step, the surprise is the unpredictable occurrence of the conditioned stimulus (CS). Since this random incidence of CS is a sign of subsequent reward that is not being expected by the participant, it causes TD error. Interestingly, dopaminergic neurons also respond to this CS. The unexpected omission of reward is another source of unpredictability and produces TD error. The error value is negative since what happened is worse than what was expected. This justifies the suppression in dopaminergic activity observed in (Doya, 2007; Figure 4, bottom row).

Dopaminegic neurons are mainly found in substantia nigra and ventral tegmental area (VTA), two midbrain areas which are parts of basal ganglia. Substantia nigra is divided into two parts: Substantia nigra pars recticulata (SNr) and Substantia nigra pars compacta (SNc). SNc is suggested to be the area where TD error is calculated (Doya, 2000; Berns & Sejnowski, 1998). The computed error is based on two inputs; one from limbic system which supposedly calculates the present reward and another from striatum which is coding the future reward (Dayan, 1999).

Unsupervised Learning

Theory

Unlike supervised learning and reinforcement learning, in unsupervised learning there is no external feedback: neither any training data nor any environmental evaluation (Figure 17). Unsupervised learning exploits any noticeable statistical regularity of the input data to help it learn the proper action. The task of an unsupervised learner is to detect a pattern. The principal assumption is that the input samples (x_i) are from an unknown probability distribution ($P(x)$) (Dayan, 1999). The main class of unsupervised learning algorithms tries to explicitly estimate this underlying probability function (Moghaddam & Pentland, 1997). Estimating this probability function can help to predict the probability of new input x_t given the previous inputs $x_1, x_2, \ldots, x_{t-1.}$

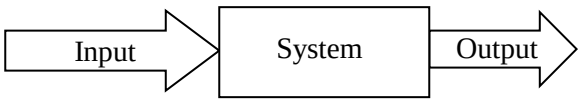

Figure 17. Unsupervised learning with absolutely no external feedback

Unsupervised learning can be investigated from both information theory and Bayesian points of view. The purpose of both methods is to calculate how well a particular probability function fits the observations. In order to understand the logic behind the use of information theory one needs to know the concept of "entropy". As was introduced in chapter 2, entropy measures the amount of randomness in a probability distribution. For a random variable x with probability distribution of $P(x)$, entropy is defined as:

$$H(P) = -\sum_x P(x) log_2 P(x) \qquad (14)$$

in which $H(P)$ is the amount of entropy of $P(x)$. The higher the entropy for a probability function, the more random the corresponding variable. Cross entropy is a metric defined based on the definition of entropy to measure the similarity of two probability functions. This measure computes how close two distributions are:

$$H(P, Q) = -\sum_x P(x) log_2 Q(x) \qquad (15)$$

where $P(x)$ is the real or target probability and $Q(x)$ is the estimated probability. Cross entropy calculates how close the estimated probability, $Q(x)$, is to the real probability $P(x)$. It can be shown that the minimum value of $H(P, Q)$ with respect to Q occurs when $Q(x) = P(x)$; known as Principle of Minimum Cross Entropy. In this case the cross entropy simply equals the entropy of P. The closer two distributions are, the closer the cross entropy is to its minimum value which is the entropy of $P(x)$. The problem is $P(x)$ is unknown for the purpose of unsupervised learning. To overcome this problem, the Monte Carlo estimation of the true cross entropy is defined as bellow:

$$H(Q, C) = -\sum_{i=1}^{N} \frac{1}{N} log_2 Q(x_i) \qquad (16)$$

in which the x_i represent input samples available in data set C where $i = 1, \ldots N$; N is the total number of data and Q is the estimated distribution. It can be proven that as N goes to infinity; this estimation of cross entropy converges to the cross entropy measure for the true distribution.

From a Bayesian perspective, a parametric probability distribution, *m*, is considered as the estimate of target probability. The task is to approximate the unknown parameter set denoted by θ associated with this distribution. Maximum a posteriori (MAP) is a Bayesian method employed to estimate the optimal value of these parameters. This method finds the most probable set of parameter ($\hat{\theta}$) given the observations so far (*D*):

$$\hat{\theta} = arg\,max_{\theta}\ P(\theta|D, m) \qquad (17)$$

The logic behind this method is simple and straightforward: the best set of parameters ($\hat{\theta}$) for the assumed probability distribution (*m*) is the one that is most probable given the observed input samples (*D*). Many learning algorithms are developed to find these optimal values (for more detail see (Gauvain & Lee, 1994) and (Ghahramani, 2004)).

Neuroanatomy of Unsupervised Learning

There are many anatomical and neurobiological reports emphasizing the role of cerebral cortex in unsupervised learning (Wig, 2012; Bostan et al., 2013; Meyer & Olson, 2011). This part of the brain is known to learn the statistics of the environment. The significance of this type of learning is that it seems to be much more common in the brain than supervised learning. The important role of unsupervised learning as been emphasized for language acquisition and visual object recognition (Conway, Bauernschmidt, Huang, & Pisoni, 2010).

In terms of vision, there are around 10^6 photoreceptors in each eye that encode the information about the objects available in the world. The activity of these sensory neurons highly impacts the synaptic properties of neurons in neocortex. One perplexing question for which unsupervised learning seems to have the answers is how invariant objects representations can be developed from heterogeneous visual input in the absence of any external feedback (DiCarlo & Cox, 2007; Riesenhuber & Poggio, 2000).

The literature on language learning suggests that humans learn a language by detecting complex patterns known as statistical syllable sequences (Saffran, Johnson, Aslin, & Newport, 1999; Aslin, Saffran, & Newport, 1998). This type of implicit (unsupervised) statistical (or sequence) learning is shown to be necessary for many different types of language competence such as word segmentation

(Saffran, Aslin, & Newport, 1996), word learning (Mirman, Magnuson, Estes, & Dixon, 2008) and acquisition of syntax (Ullman, 2004).

Humans are shown to be able to discover the spatial structure or temporal statistics in any sensory input (either in visual or auditory modality) without any hint about the existence of a possible pattern. To learn such structures, humans must learn simple, joint and conditional probabilities. Recently d'Acremont and colleagues showed increased activity in BOLD response in medial prefrontal cortex and angular gyri in a probability learning task (d'Acremont, Fornari, & Bossaerts, 2013). Other studies have found activity of the same areas for highly correlated measures such as surprise (Fletcher et al., 2001) and state prediction error (SPE) in reinforcement learning (Gläscher, Daw, Dayan, & O'Doherty, 2010). Fletcher and colleagues showed that the activity of dorsolateral prefrontal cortex (DLPFC) is regulated based on the level of unpredictability of sequence occurrences. Their result indicates the higher bold activity of this area with surprising violations of learned associations. Glascher and colleagues also using fMRI argue that intraparietal sulcus and lateral prefrontal cortex play an important role in estimating state prediction error which is simply defined as the probability of transition from one state of the environment to another.

Statistical learning at the synaptic level can be explained by Hebbian learning. The idea is that synaptic connections of neurons with similar activations strengthen over time. Cortical synaptic strength has been shown to increase when there is an association between the activity of presynaptic and postsynaptic neurons and to decrease when there is no such an association (Artola, Brocher & Singer, 1990; Tsumoto & Suda, 1979).

New Version of ELPH

ELPH's failures in mimicking HC performance prompted us to look for alternative learning mechanisms that are more neurally plausible. Supervised learning is type of learning that learns the input to output mapping from a set of training example pairs. Reinforcement learning learns the best sequence of actions based on a reward signal evaluating the system performance and finally unsupervised learning learns the regularities that exist in the input data without any external guidance. Comparing our task with these learning mechanisms suggested reinforcement learning as the most plausible candidate. There is not any explicit reward signal obtained for the participant at the end of each trial. Nevertheless we believe that the fact that each trial ends up with a win a loss or a tie for the participants can be interpreted as a reward signal by them. This does not, however, preclude a role for

the other two mechanisms; especially supervised learning since at the end of the trial the computer's choice is revealed to the participants. This possibility will be discussed in the next chapter. Comparing the brain areas damaged in our RBD participants to those supporting each of the principal learning mechanism also favored the RL mechanism. Insular cortex and basal ganglia lesions showed the greatest overlap for the poor RPS players in (Danckert et al., 2012). In this section, I describe a modified version of the ELPH algorithm that incorporates a new learning mechanism. I also demonstrate the performance of this new model in replicating the behavior of our participants.

RELPH: a combination of RL and ELPH

Our new model is a combination of the previous version of ELPH and reinforcement learning. I refer to this new model as RELPH. As described previously all RL methods maximize total expected return. One main approach to achieve this is to define a value function over the state space. It can be proven that taking actions that maximize the value function will maximize the total expected return as well. I borrow this idea from RL and apply it to our model.

For our purpose, instead of defining a value function over the state space, a value function is defined over the hypothesis space. In RELPH each hypothesis has a value. The value is the expected reward that the learner receives by playing that hypothesis. The best hypothesis in this case is not the most predictive but the most rewarding one. The true value of each hypothesis is however unknown and must be learned over time. To do so the initial values are set randomly and are updated at the end of each trial after taking each action according to the delta rule, equation (13):

$$V_{t+1}(H) = V_t(H) + \alpha\delta_t \qquad (18)$$

The error signal for each hypothesis is defined as:

$$\delta_t = r_t - V_t(H) \qquad (19)$$

Based on these formulas, if the value of *H* matches the instance reward at time *t*, no updating is needed. If not the difference between the prediction and the observation is defined as an error in prediction and used to update the value. α is the parameter adjusting the weight of previous belief and the instance reward. By replacing equation (19) in equation (18), the updating rule is simplified to:

$$V_{t+1}(H) = (1 - \alpha)V_t(H) + \alpha r_t \qquad (20)$$

The α value can change from zero to one. The closer this parameter to one, the more significant the role of instantaneous reward is in determining the hypothesis value and vice versa; the smaller the value of α the greater the role of prior belief. Based on this definition, this parameter should be big at the beginning of learning and drop gradually by time. But as is discussed later, this strategy leads to better learning, but makes updating harder. Thus, for now, I assume a fixed value for this parameter which differs from one participant to another.

At the end of this section, a brief description of RELPH is presented detailing the several modifications that have been applied to the original model. The two main processes in RELPH are the learning process and the pruning process. The former process is a reward-based process and the latter remains a probability-based process. The most rewarding hypotheses are employed to select the best item to be played in subsequent trial and non- predictive hypotheses are deleted from HS to facilitate adaptation. Each process is briefly described in the rest of this section.

a. Learning Process: This process consists of the same three functions used in original ELPH: hypothesis generation function, updating function and prediction function. Each of them is reviewed in this part.

i. Hypothesis Generation: The main difference between hypotheses in RELPH and ELPH is that instead of the observation being stored, it is the plays that are stored in the prediction set. Consequently, instead of the counts of occurrence of each observation, the total amount of reward for each play is recorded. Each win is assumed to be equal to a reward value of +1, tie is 0 and loss is -1. Recalling the example from the first chapter (see Page 27): the computer plays a simple pattern of rock, rock, paper. STM length, *n*, is 2 and the last two observations at time *t-1* are ’paper’ and ’rock’ and therefore the current content of STM is (’paper’, ’rock’). In ELPH, hypotheses were generated to answer the question of which item is more probable to be observed next. In RELPH however the question is different; this time hypotheses are guesses about which item is more likely to lead to winning in the next trial. In our example, let’s say the player decided to play scissors (initially decisions must be made randomly); since the next observation was rock, it would lose that trial. So the three generated hypothesis would update to:

Hyp1: {(paper, *t-2*), (rock, *t-1*)} => [scissors, -1]

Hyp2: {(rock, *t-1*)} => [scissors, -1]

Hyp3: {(paper, *t-2*)} => [scissors, -1]

meaning that the total amount of reward received so far for playing scissors after observing a sequence of rock and paper is -1 (according to Hyp_1). Thus in RELPH a prediction set is defined as all the items that the player has tried followed by the total amount of reward received for that play. For a particular hypothesis, *Hyp*, a prediction set is defined as follows:

$$\{[e_1, r_1], [e_2, r_2], \ldots, [e_m, r_m]\}$$

in which e_1, e_2 … e_m is the list of plays (rock, paper, or scissors in the present case) and r_1, r_2... r_m represents the total amount of reward gained for playing each item. In our example, if for the next 5 times which the player observed a sequence of paper and rock, he decided to play rock 2 times and paper 3 times, the three above hypotheses would update to:

Hyp1: {(paper, *t-2*), (rock, *t-1*)} => {[scissors, -1], [paper, 3], [rock, 0]}

Hyp2: {(rock, *t-1*)} => {[scissors, -1], [paper, 3], [rock, 0]}

Hyp3: {(paper, *t-2*)} => {[scissors, -1], [paper, 3], [rock, 0]}

Based on these prediction sets, it is clear that paper is the most rewarding option to be played next time that a sequence of paper and rock is being observed since it results in the total reward of +3 which is higher than the total reward of two other options.

As is explained in chapter 2, the content of memory declines over the time and thus make humans unable to remember the exact amount of reward they have received. For this reason r_1, r_2... r_m values in our model must be changed to the total amount of reward that the learner can remember. To model forgetting process, the modified version of the forgetting formula expressed in equation (7) is used to define the activity of each option in RELPH:

$$A = sign\left(\sum_{i=1}^{n} r_i(t_i^{-d})\right) ln\left(\left|\sum_{i=1}^{n} r_i(t_i^{-d})\right|\right) \qquad (21)$$

where r_i is the amount of reward received after *i*-th times of playing that option at t_i. Therefore the final form of prediction set of *Hyp* is {[e_1, A_1], [e_2, A_2],…, [e_m, A_m]} where A_i is the activity of e_i and calculated according to equation (21).

ii. Prediction: While the most eligible hypothesis in RELPH is the most rewarding one, to keep the balance between exploration and exploitation, this hypothesis is not selected all the time. As discussed in the previous chapter, the soft-max decision rule is applied as the action selection rule to improve the estimation of hypotheses value. Thus the probability of selecting H_i is calculated based on its relative value:

$$p(H_i) = \frac{exp(V(H_i))}{\sum_j exp(V(H_j))} \qquad (22)$$

where the sum is calculated over all the related hypotheses. Related hypotheses are those that could be played next because their content matches that in STM. After selecting the best hypothesis, this hypothesis is employed to predict the most rewarding option to play in the next trial. Among all the possible items to play, the one leading to the greatest total reward has the highest probability of selection. To examine the other possible options again the soft-max rule is applied:

$$p(e_i) = \frac{exp(A_i)}{\sum_j exp(A_j)} \qquad (23)$$

in which the sum is computed over all the available options (rock, paper and scissors in our task). It is important to note that even the options that are not in the prediction set of the chosen hypothesis have the chance of being selected. The corresponding reward values for such options are set a random value in the range of (-1, 1). Thus in this case *j* has three values of 1, 2 and 3.

iii. Updating: Updating function also includes three steps per se: **hypothesis updating**, **value updating** and **forgetting**. Hypothesis updating refers to updating each hypothesis according to the result of the last play. This rule stays the same. If the hypothesis generated based on the current content of STM already exists in HS, its prediction set will be updated. If not, this hypothesis will be

added to the HS. For the existing hypotheses, if the last play already existed in the prediction set, the corresponding total reward is updated based on the last value of reward. If not, this new play and the last reward value are added to the prediction set. Value updating refers to updating the value of each hypothesis based on the last reward value, equation (20). At the end of each trial, forgetting mechanism is also employed to update the recalled total reward.

b. Pruning: The idea behind the pruning function in RELPH is quite similar to ELPH. This time instead of choice entropy, outcome entropy is computed. In ELPH, probability of observing each prediction is used to calculate the entropy associated with a hypothesis. In RELPH however the probability of each outcome is used to calculate the entropy. The inconsistent hypotheses in RELPH are those that are not able to make a concrete prediction about the expected reward. All the possible outcomes in our task, RPS game, are win, tie and lose and the count of numbers those happen denoted by o_j where j = 1, 2, 3. The outcome entropy for *Hyp* is then calculated as follows:

$$P(o_i) = \frac{o_i}{\sum_{j=1}^{3} o_j} \rightarrow H\big(P(Hyp)\big) = -\sum_{i=1}^{m} \frac{o_i}{\sum_{j=1}^{3} o_j} log_2 \frac{o_i}{\sum_{j=1}^{m=3} o_j} \quad (24)$$

The hypotheses which are not predictive are removed from HS and non-predictive hypotheses are those which fail to lead a persistent trend of reward. The hypotheses which sometimes lead to wining and sometimes lead to loosing shouldn`t be hired to offer the next play.

Result

In this section the performance of RELPH in replicating our HCs' LBDs' and RBDs' behavior is investigated. This algorithm has four parameters which must be estimated: Entropy threshold for pruning function (H_{thr}), STM length (n), forgetting decay ratio (d) and learning rate (α). Maximum likelihood is employed to estimate the optimal values for these parameters. Our measures still are the max win rate and learning rate and to compare these measures we split the 20 blocks (each block is the average of 10 consecutive trials) in the strong condition into two sets of 10 blocks each. The first 10 blocks in which the response is more transient are selected to compare the learning phase of RELPH and our participants. The last 10 blocks in which the response in more stable are used to compare the max win rate. The result of this comparison for each group (HCs, LBDs and RBDs) is presented the rest of this section.

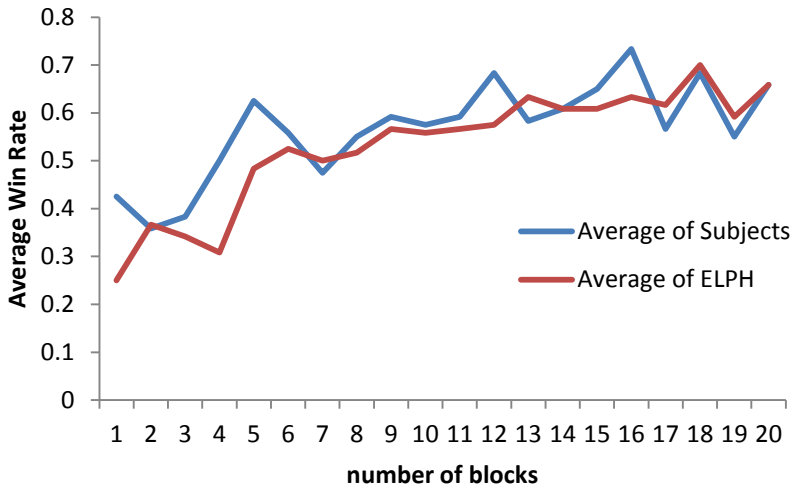

Figure 18. Average win rate for RELPH and that of HCs. The red line in the figure represents the average win rate of RELPH and the blue line shows the average win rate of all the participants in HC group.

Figure 18 shows the average win rate for both RELPH and HCs. RELPH is successful at capturing the behavior of our HCs both in terms of max win rate ($F(1, 22) = 0.039$, $p > 0.5$, $\eta^2 = .020$) and learning rate ($F(1, 22) = 1.228$, $p > 0.1$, $\eta^2 = 0.53$). Best matches parameters are also represented in Table 4 (H_{thr}: Mean= 2.2, SD=0.8; n: Mean=2.1, SD = 0.67; α: Mean = .27, SD= .22; d: Mean= .23, SD = 0.16)

Table 4. Optimal parameters of RELPH for each participant in HC group calculated using ML estimation approach.

Participant Number	Best H_{thr}	Best n	Best α	Best d
1	2.8	2	0.10	0.20
2	3.0	3	0.13	0.33
3	1.8	2	0.50	0.30
4	0.0	1	0.15	0.60
5	2.3	3	0.75	0.00
6	2.0	2	0.28	0.05
7	2.6	2	0.03	0.40
8	2.3	3	0.20	0.10
9	2.6	2	0.05	0.20
10	2.8	1	0.50	0.20
11	2.1	2	0.40	0.17
12	3.0	2	0.10	0.20

The average win rate for RELPH and RBDs are also shown in Figure 19 for the parameters represented in Table 5 (H_{thr}: Mean=1.72, SD=0.9; n: Mean= 2, SD = 0.7; α: Mean= 0.3, SD=0.23; d: Mean = 0.18, SD=0.18). A repeated-measures ANOVA with within subject factors of block is used to compare the max win rate and learning rate of RELPH and RBDs. The result shows no significant difference between these two groups for max win rate ($F(1, 26) = 0.679$, $p > 0.1$, $\eta^2 = .025$) or learning rate ($F(1, 26) = 0.096$, $p > 0.1$, $\eta^2 = .004$). Although learning rate cannot be really defined for RBDs, as it is shown in chapter 1 that they didn't really learn the computer strategy.

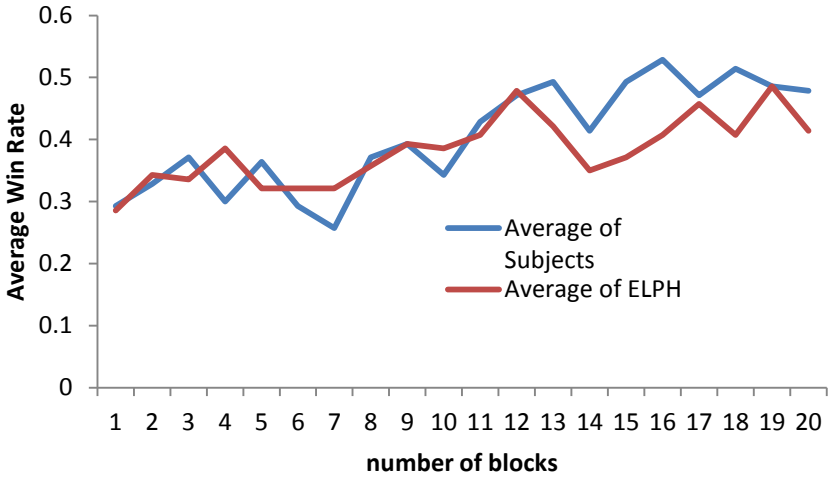

Figure 19. Average win rate for RELPH and that of RBDs. The red line in the figure represents the average win rate of RELPH and the blue line shows the average win rate of all the participants in RBD group.

This result is promising since it shows the acceptable performance of our model in replicating the result of both RBDs and HCs for different sets of parameters. Comparing the values of these parameters for these two different groups, however, doesn't show any significant difference in any of these parameters for two groups. A simple one-way ANOVA analysis with dependent variable of group (which is a nominal variable with two different values; either "HC" or "RBD") is employed to compare the values of each parameters between two groups (H_{thr}: $F(1, 25) = 2.797$, $p > 0.1$; n: $F(1, 25) = .099$, $p > 0.5$; α: $F(1, 25) = 0.077$, $p > 0.5$; d: $F(1, 25) = 0.471$, $p > 0.5$). This analysis

reveals that not any one of these parameters is significantly different from one group to another one. The rest of this section is an effort to explain these results.

Table 5. Optimal parameters of RELPH for each participant in RBD group calculated using ML estimation approach.

Participant Number	Best H_{thr}	Best n	Best α	Best d
1	2.2	1	0.51	0.00
2	2.7	3	0.40	0.15
3	1.6	2	0.10	0.00
4	2.2	2	0.40	0.10
5	0.4	3	0.46	0.64
6	2.2	3	0.05	0.25
7	0.4	2	0.10	0.10
8	0.5	2	0.75	0.40
9	2.2	2	0.07	0.10
10	2.2	2	0.10	0.10
11	2.6	1	0.30	0.00
12	2.8	2	0.00	0.10
13	0.6	2	0.58	0.30
14	1.4	1	0.26	0.31

The values for H_{thr} and α for each group separately are plotted in Figure 20. Although no significant difference is detected, this figure suggests different patterns for the values of this parameter for each group (this difference between groups is less significant for two other parameters [data not shown]). HCs tend to have large H_{thr} and lower α compared to RBDs (in the horizontal axis 1 represents HC and 2 RBD group). Larger H_{thr} means keeping more random hypotheses and lower α values means having more trust on the current model (equation (13)). Putting these together suggests that although HCs are open to new hypotheses, they are not willing to change their belief about the value of each hypothesis. They keep those not-predictive hypotheses in their HS but they don't trust them so soon. This observation can be significant since it suggests that there might be a time lapse between the time that HCs starts generating new hypotheses and the time they start using these new generated ones. If this is the case, these results suggest that there is a time period between when HCs

detect the change-points (the point that world is changed) and when they switch to these new hypotheses generated for this new world.

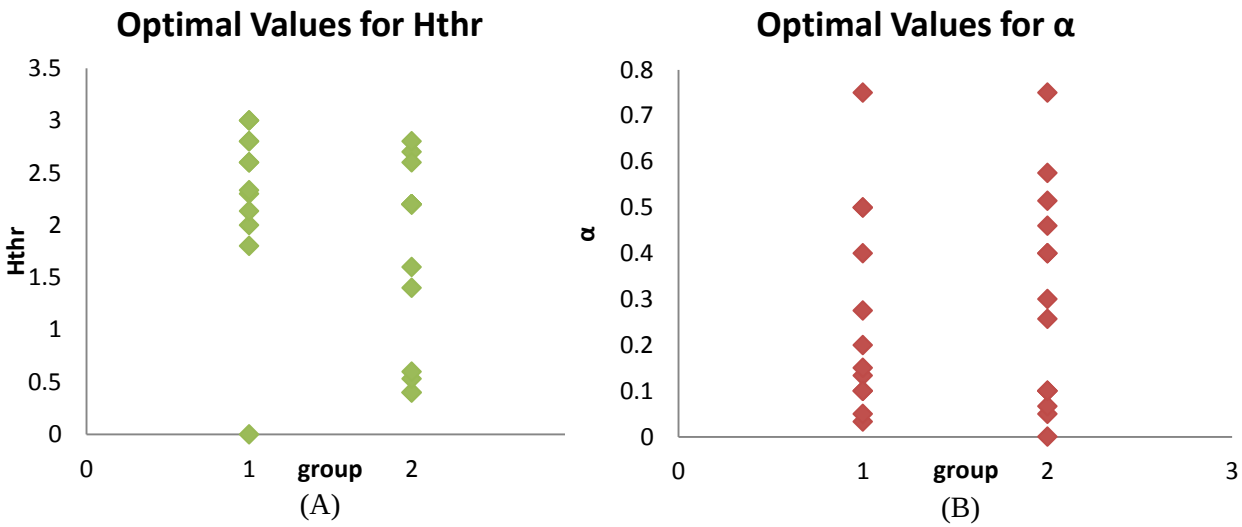

Figure 20. The values of (A) H_{thr} and (B) α for the participants in both HC and RBD group.1 on the x axis represents the HC group and 2 RBD group. Also each dot in this plot represents one participant.

On the other hand the values of both H_{thr} and n for RBD participants seem more uniformly random. These parameters have totally different values for participants in RBD group which shows no similarity between these patients in this sense. There might be two reasons behind this observation: (1) these results can be due to the potential heterogeneity existing in RBDs' behavior, (2) The difference between HCs and RBDs are not due to the significant difference in one of these parameters. To examine the first idea, a scatter plot of the total number of wins versus the values of these two parameters of RBD patients and HCs is plotted in Figure 21. Firstly, this plot demonstrates the heterogeneity in the result of RBDs. Although most of them do not reach a high total win rate; there are some that are as good as some of the HCs. Secondly, no relationship between winning and the value of these parameters in RBD group can be observed in this plot. Thus although RBD patients behaved differently in our task, there is no relation between the values of these parameters and the win rate.

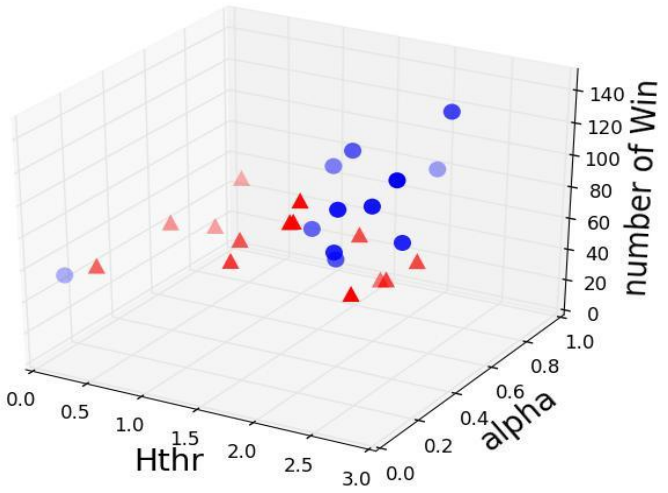

Figure 21. The total number of win rate versus H_{thr} and α. Each dot represents one participant. Red triangles are RBD patients and blue circles are HC participants.

A second possible explanation is that the impairment in RBDs is not due to the failure of a single compartment of the model. Regardless of these results we can demonstrate that our computational model can generate different patterns of performance with different parameter sets. Stated another way, dysfunction can occur from an abnormality at multiple levels of the model.

This result can be interpreted in two different ways. First my outcome measurement, win rate, was not rich enough to fully reflect the characteristics of our patients' behavior, and therefore our model cannot simply account for patient performance. If this is the case, we need to be more specific in specifying the nature of impairment, and we will need to select other aspects of RBDs' behaviors as outcome measures. Second there may not be a single reason for RBDs to fail the task. In other words, exactly as observed in RELPH, the reason behind the failure of each RBD patient is different from another. This explanation might also be plausible due to the complexity of the tasks and variety of brain areas involved in brain damage patients. This also is discussed more in chapter 6.

It is important to note that the number of our participants in each group is not large and this might lead to small statistical power for my analysis in this section. Adding more healthy controls and brain damaged patients might strengthen this result. However recruiting brain damaged patients is not easy as there is not many patients that are in the appropriate health for our lengthy behavioral task.

At the end, it is important to make sure that this model still is able to replicate the result of LBDs by simply removing the soft-max component. To examine this idea, I played RELPH against the computer, with the same parameter sets as healthy controls. The only change here is the soft-max component being removed. This means at each trial, the hypothesis with the highest value and the action with highest associated reward are always selected to predict the next action. I refer to this version of RELPH as "greedy-RELPH". Figure 22 shows the average win rate for both LBDs and greedy-RELPH. Same as the previous model, greedy-RELPH max win rate is still same as LBDs' (F (1, 18) $= .335$, $p > 0.5$, $\eta^2 = 0.018$). In addition, in contrast to ELPH, there is no significant difference between the learning rate of greedy-RELPH and LBDs ($F(1, 18) = 1.890$, $p > 0.1$, $\eta^2 = 0.095$).

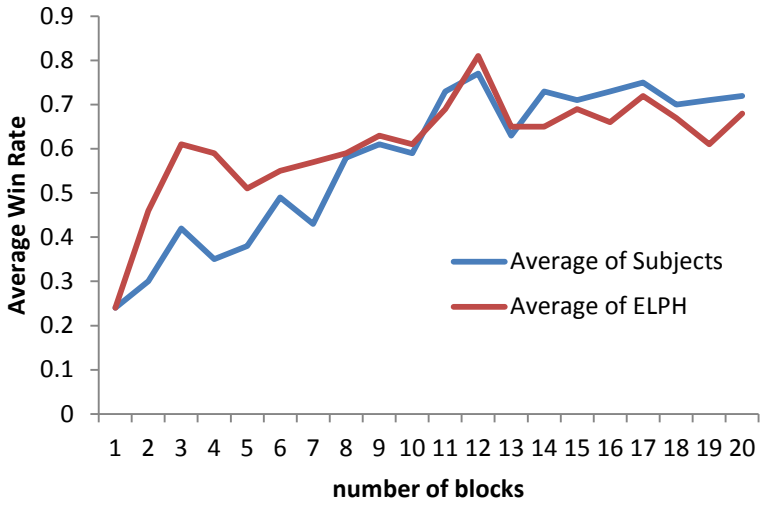

Figure 22. Average win rate for greedy-RELPH and that of LBDs. The red line in the figure represents the average win rate of greedy-RELPH and the blue line shows the average win rate of all the participants in LBD group.

Discussion

My effort to improve ELPH in chapter 3 indicates that the reason behind ELPH outperformance is more than a missing component. ELPH learns faster than our healthy participants and my investigations in the previous chapter demonstrate that this difference is not due to memory constraints. While ELPH's method consists of two main functions, learning and pruning, the only

remaining explanation is that one of these functions is not consistent with the one that our participants employed in this task. Since the difference is in the learning phase and reflects a steeper learning slope, referred to as learning rate, I deemed an alternative learning mechanism to be the most useful modification. Among all the three main learning mechanisms in the brain, reinforcement learning seems to be the best alternative considering both the nature of our task and the brain area damaged in RBDs. Supervised learning is still a conceivable alternative as after each trial the computer's choice is revealed to the participants.

A combination of RL and ELPH named RELPH is suggested in this chapter as a model of human decision making. In this new version of ELPH, learning happens based on one of the simplest updating rule in RL called the delta rule and pruning happens based on the entropy of the outcomes. Thus, in this model, there is dissociation between learning and pruning functions compared to ELPH since two totally different mechanisms are used for each. As will be discussed in the chapter 6, this assumption is neurally plausible since different brain areas are known to be involved in these two mechanisms.

This new model is successful at replicating the result of our three different groups in this task. LBDs result can be replicated by removing the soft-max compartment which suggests that LBDs are impaired in the exploration phase and more specifically in developing an action-selection rule. This model is also capable of reproducing the result of both RBD and HC groups, though I could not find any one particular parameter that consistently accounted for group differences. My result so far suggests that the updating failure in RBD is not due to a single specific damaged compartment in RELPH. Deficits in any compartment of RELPH potentially can lead to the same poor performance as RBDs. As is discussed in the previous section, this conclusion is only based on the preliminary results represented in that section and definitely more investigation is required to make an accurate and concrete conclusion.

Chapter 5: General Discussion

Comparing the failures of RBD patients in several tasks pointed us to an updating impairment in these patients. Updating is defined as the process of building mental models and adapting them to the changes observed in the environment. Since the environment is continually changing, to make reasonable decisions one needs to evaluate one's models constantly and adapt them to those changes. Adapting mental models can be as challenging as building them. To be able to adapt successfully, two main sub-processes are necessary: (1) detecting the change happening in the environment and (2) modifying the current model according to the detected change. These two steps together describe the updating process.

Changes in the environment, also known as "change-points" (Wilson & Niv, 2011; Adams & MacKay, 2007) must be detected as the first step to update the current model. If you have a mental model which is able to predict the next event accurately enough, it should be able to identify these change-points rapidly. When the environment is changed, this model is no longer an accurate model of the external world and therefore its prediction is not consistent with observations, and this in turns causes errors in prediction. Thus, to detect change-points, one needs to detect the mismatches between the prediction of one's model and real observations. The next step is to modify the model in such a way that it again fits the new circumstance. This step requires online learning in which the learner learns the optimal actions while she is taking them. After taking any action, the model is improved based on the feedback it receives from the environment. This adjustment continues to the point that the learner believes that this new model again is able to predict accurately. As is clear from the definition, the updating process requires an initial predictive model. Without it, there would be no difference between the concepts of learning and updating. Indeed, one of the main differences between these two processes is this need to detect the unpredictability of the previous model.

To have a better understanding of this process, an RPS task was conducted in our lab. In order to capture both steps of the updating process, this experiment was designed in such a way that the computer changed its strategy twice. It started with a random condition, and then switched to a moderate bias of rock and finished with a strong bias of paper. Thus two change-points exist in this task; one from a "random" to a "lite" condition and one from a "lite" to a "strong" bias. Designing this experiment, we expected that the second change-point would be more important for our purpose, since our participants had already developed a model about the task from what they learned in lite

condition. In fact, we expected our participants to learn the slight bias of rock in the lite condition and therefore when they were exposed to the strong condition they would notice the failure of their model and the need for updating. This change-point could help us to learn more about both steps of updating. Our participants, however, failed to learn the lite condition. As is discussed in chapter 3, our investigations demonstrate that contrary to what we had assumed about the task, participants didn't seem to develop an efficient model in lite condition. In fact, it turns out that the frequency strategy is one of the hardest for humans to learn since they generally expect temporal dependency in their daily observations. Regardless of the potential reasons behind this failure, what is important is that this result limits us to only studying the second step of updating, which is learning the best model. The fact that our participants didn't show any sign of detecting the computer strategy in the lite condition, prevents us from elaborating on the first step. Nevertheless, this result leads us to follow up studies which are described in the next chapter.

The result of our RPS experiment shows the impairment of RBDs in learning the strong bias in the computer play while healthy participants clearly learned the bias. In order to learn more about the potential difference between our groups, I employed a computational modeling approach. The primary goal was to have a model which was able to replicate the result of HCs. If this model is a plausible model for human decision making, it should be able to generate the result of RBDs after some modifications. Modifications can include changing some parameters or functions in the model. Knowing what has been changed in our "brain-damaged" model can help us to better understand the updating process and its impairment.

My first model, ELPH, was a model offered for temporal sequence learning in a non-stationary environment. The model consists of two main processes, learning and pruning. ELPH generates hypotheses about the computer strategy and learns the best one. To be able to keep track of hypotheses, it removes the inaccurate ones from its memory. The initial result indicates that this model outperforms both our HCs and RBDs. It learns faster (better learning rate) and achieves higher results (reaching larger max win rate). To improve the performance of ELPH in replicating the behavior of our participants (especially HCs), I examined several possible reasons behind this outperformance, which is summarized in the following paragraphs.

ELPH reaches a better max win rate. The fact that both our HCs and ELPH clearly play better than chance reflects that they both learned that the computer plays paper more often than the two other options. Yet some of the healthy controls didn't give up playing other options even long after they

realized this bias. I believe this reflects different action-selection rules between HCs and ELPH. Humans keep looking for a temporal dependency which does not exist in the computer play. The fact that this exploration behavior is not observed in LBDs confirms this idea. As Danckert and colleagues argued (Danckert et al., 2012), these results are in accordance with the assumption that the right hemisphere is responsible for processing the statistical properties of uncertain environments and for developing representations for decision making (Gazzaniga, 1995; M. B. Miller, Valsangkar-Smyth, Newman, Dumont, & Wolford, 2005; Vickery & Jiang, 2009; Roser, Fiser, Aslin, & Gazzaniga, 2011; Wolford, Miller, & Gazzaniga, 2000; Wolford et al., 2004) while the left hemisphere is an "interpreter" for those statistics (Gazzaniga, 1995).

ELPH also learns faster. Intuitively, there must be some limitation in human brains which has not been applied to our model. Not surprisingly, the most obvious limitation is memory capacity limit. Forgetting might slow the learning process. To examine this idea, I applied three forgetting mechanisms to our model; limiting the length of STM (n), limiting the capacity of HS and memory decay. The first one represents the working memory limits, the second one signifies short term memory limit and the last one models the memory decline over time. Instead, implementing these forgetting mechanisms failed to slow learning, and in some cases, even increased the speed of learning.

Limiting the variation range of n did not help significantly since the optimal value for this parameter in our task is 1. STM with the length of n means that the hypothetical patterns between n+1 consecutive observations are investigated by ELPH. In our task there is no absolute temporal dependency between successive items in computer play and thus any value larger than 1 for n does not help learning. All three strategies played by the computer are from frequency type strategies which are those in which one option is chosen more than others. In other words, the computer's choices were made independent of prior trials (statistically, the observations are independent and identically distributed; i.i.d). Therefore, there is no reason to consider STM capacities beyond size 1.

Limiting the capacity of STM and memory decay also failed to slow learning. These mechanisms cause one to forget what has been observed in the past to make space for new observations. This process can decrease the performance if it leads to forgetting the relevant information, but if the information stored in memory is no longer relevant, removing the information can actually facilitate the learning process. Therefore, in a nonstationary environment (represented in our task by the

transition from lite to strong conditions) it is not always helpful to remember all past events, since some of them might no longer hold true, and this can stall the learning process.

If the improved performance of ELPH in learning is not due to some limitation omitted in the model, it must be due to the learning mechanism per se. ELPH can learn faster because it uses a faster learning mechanism compared to HCs. To learn the computer strategy, ELPH makes hypotheses and learns the best one by keeping track of the opponent's play. Indeed, to learn the strategy there is no need for ELPH to play against the opponent, as watching the sequence of choices is sufficient. This is not surprising since ELPH is originally a pattern detection method. Among all three types of learning discussed in chapter 4, ELPH's learning mechanism is closest to supervised learning. After each play, the computer's play was revealed to the participant and these observations are used in ELPH to update the generated hypotheses and thus improve the subsequent prediction. Since the amount of guidance received from the environment is greater than any other sort of learning, it is reasonable to believe that this type of learning is faster than any other. This can explain the faster learning phase observed in ELPH; our participants might recruit another type of learning.

My investigation to improve the model demonstrates that our participants most likely employed reinforcement learning instead of any sort of supervised learning. The result in chapter 4 also confirms this claim. Our task has an implicit reward motivation, since it is a competitive game ending in a win, lose or tie. The comparison between the brain area involved in reinforcement learning and the brain area damaged in RBDs also favors reinforcement learning. Although supervised learning is still a conceivable alternative, slower learning rate, overlapping brain areas, and a reward motivation being compatible with our tasks are three main reasons that persuade us that reinforcement learning is the best candidate for our task. The success of this new model, RELPH, in replicating the result of our three groups of participant also confirms our idea of reward-based learning.

While reinforcement learning seems to be a good framework for our task, there were two main concerns with using the currents methods in RL as a model of our participants' RPS performance. The first consideration, discussed in chapter 1, is that typical reinforcement learning methods are not developed for nonstationary environments and are known to be slow in adapting to change (Poupart et al., 2006; Doya, Samejima, Katagiri, & Kawato, 2002; Basso & Engel, 2009). The whole idea of updating however is to capture the ability of humans to adjust to changes quickly and this is strength of ELPH. The second problem is finding the proper state representation for an RL method. State representation consists of all the information about the environment that is required to learn the

optimal decision. Depending on the task, finding these influential factors might not be straightforward especially for the type of decision making tasks which humans face in their daily life. For classic RL methods however state representation is not a big challenge since most of the time the proper representation is given to the agent by the designer. The idea of generating hypotheses, as ELPH does, can be a solution for the state-representation. This idea however is not very different from the concept of state representation; it seems more abstract and easier to tackle especially for human decision tasks. For these reasons, we believe that ELPH with modifications in the learning process, called RELPH, is a reasonable model of human decision-making.

Although this is not the main focus of this thesis, this result favors reinforcement learning methods against optimal Bayesian Learning as human learning processes, which is currently under debate in the computational neuroscience field of study (Payzan-LeNestour, 2010; Payzan-LeNestour & Bossaerts, 2011; Gläscher, Daw, Dayan, & O'Doherty, 2010). It is not clear yet if humans' learning process happens mainly based on reward learning or probability learning. Our results suggest that at least in our task, if the participants had employed the probability approach as in ELPH, they would have been faster in learning and adapting. On the other hand, reinforcement learning approach, with the help of pruning function (which is probability driven) fits the data. Therefore based on our model, reward has a main contribution in learning even for tasks such as our task in which there is no explicit reward involved.

Yet the reason that participants limited themselves to the reward signal, when it was possible for them to use more information from outside, is still not clear. One reason might be due to their limited attentional resources available in the context of a complex task, such as ours. Humans are social creatures which are forced to deal with this complicated world every day. To be able to cope with this complexity, they learned to consider several possibilities which sometimes may not be necessarily relevant to the task at hand. As an example, RELPH is not flexible in terms of generating hypotheses about the opponent's possible strategy. It doesn't investigate any other possibly related features than the opponents' sequence of plays. The participants however consider other possibilities such as the chance of the computer plays being dependent on their own play.

Chapter 6: Future Work

As is discussed in the previous chapter, the main focus of this thesis is on the second step of updating, the learning process, as our task failed to capture the first step of updating. It is important to note that even though participants did not learn the computer's strategy in the lite condition, they may still have built a model about the computer's strategy; however, any model developed was inefficient. Since we do not have access to their mental models in the lite condition, it is almost impossible for us to investigate whether they noticed the change-point. As noted in the previous chapter, the first step of updating requires a model of the external world that was initially accurate. For this reason, the focus of this thesis was mainly on the learning function. In this chapter, the first step of updating is discussed in more detail, with a view towards future work in this area.

To update one's model, one needs to be convinced that one's current model is no longer an accurate one, a realization that requires us to see errors in our model's predictions. But as is discussed in chapter 1, detecting the mismatches is not always straightforward due to the non-deterministic nature of the environment. Since the environment is in flux and our sensory inputs are noisy, it is not always our model that is the source of those errors. Rather, errors can be the consequence of the probabilistic nature of the task. This introduces an uncertainty about the source of errors, and to make reasonable decisions, a person must estimate this uncertainty. A brief literature review is presented in this section to help us to formulate future projects.

Two main types of uncertainty are known to be associated with the decision making process; expected uncertainty and unexpected uncertainty. The term "expected uncertainty" refers to the type of uncertainty still existing even after ideal learning (Payzan-LeNestour & Bossaerts, 2011) as it stems from the stochastic nature of the environment. Remember the example in chapter 1; the reward probability of taking a certain action in a particular circumstance leads to winning 80% of the time and losing the rest. In this case, even learning the rule does not prevent us from making errors. Still there is a 20% chance of observing an outcome different from what is expected; but this error does not reflect the inaccuracy of our model. In contrast, "unexpected uncertainty" refers to a fundamental and dramatic change in the environment, when the errors in the predictions actually reflect the changes happening in the external world. In our example, if the rule is changed so that taking that particular action in the same environment leads to 30% chance of winning, then our errors are due to

using an outdated model. Clearly based on this classification, mismatches in our definition of updating are those errors that are interpreted as unexpected uncertainty.

While learning mechanisms in the brain have been widely studied mathematically, computational models for uncertainty are few, especially ones that differentiate unexpected and expected uncertainty. Two general approaches have been taken concerning modeling uncertainty: the Bayesian approach and the reinforcement learning approach, which are based on similar premises. For the sake of simplicity, I will omit mathematical formulas, and offer a general introduction. Similar to reinforcement learning, Bayesian leaning also has a parameter called "learning rate" (denoted by α in equation (20)). In both methods, this parameter determines the degree by which value functions should be updated (Behrens, Woolrich, Walton, & Rushworth, 2007; Nassar, Wilson, Heasly, & Gold, 2010). A large learning rate shows the importance of incoming data compared to the history of observations. A small learning rate, in contrast, shows the reliance on previous experiences over new observations. It is clear that the value of α should be high when the level of uncertainty is high; either the observations are not yet adequate for prediction, or the environment is so fast-changing and the rules change so rapidly that past experiences expire quickly. In contrast, the small value of this parameter is appropriate when the historical information is salient (Krugel, Biele, Mohr, Li, & Heekeren, 2009; d'Acremont et al., 2013).

In a recent study, Payzan-LeNestour and Bossaerts show that in Bayesian learning adjusting the learning rate based on the level of uncertainty can capture the unexpected uncertainty or so called "probability of a jump" (Payzan-LeNestour & Bossaerts, 2011). They suggest that in order to estimate the expected uncertainty, the entropy measure can be employed; equation (1). As I explained in the unsupervised learning section, the entropy of a distribution shows the amount of randomness of that function. Therefore, the entropy of outcome probabilities is a measure of the randomness associated with predicting the outcome. Knowing the amount of the predictability of the outcome is equivalent to estimating the expected uncertainty. In their model, computed expected uncertainty influences the choice selection rule (same as the soft-max rule in our model).

Thus expected vs. unexpected uncertainty can be highly connected to the exploration versus exploitation dilemma in decision making (Cohen et al., 2007). In reinforcement learning, exploration is required to search for all possible options to make sure what is determined to be the best action is actually the optimal one. Exploitation, on the other hand, is employing what has been learned so far to act optimally in the current situation. Although exploration is necessary, especially at the beginning

of learning, selecting alternative options decreases the performance overall. This exploration-exploitation tradeoff can be regulated by uncertainty (Bland & Schaefer, 2012). Exploitation is persisting in our current model even in the presence of possible mistakes in prediction. This means believing in whatever error we get is due to expected uncertainty. Unexpected uncertainty, however, can lead us to switch from the exploitation phase to the exploration phase again; in essence, this is akin to our definition of updating.

Neuroimaging studies reveal the contribution of the anterior insular cortex (Bossaerts, 2010), anterior cingulate cortex (ACC) (Stern, Gonzalez, Welsh, & Taylor, 2010), orbito-frontal cortex (OFC) (Tobler, O'Doherty, Dolan, & Schultz, 2007), posterior parietal cortex (PPC), dorsolateral prefrontal cortex (DLPFC) (Huettel, Song, & McCarthy, 2005) and amygdala (Hsu, Bhatt, Adolphs, Tranel, & Camerer, 2005) in the processing of uncertainty (Bland & Schaefer, 2012). Most of the literature in the neural mechanism of uncertainty is concentrated on expected uncertainty. Almost all of these experiments induce uncertainty by varying the level of probabilities. Their results suggest that expected uncertainty is coded mostly in the cortical areas mentioned above. These various areas are known to have different roles in estimating the uncertainty. For example, Tanaka et al. suggest the involvement of OFC in predicting the immediate reward. DLPFC and inferior parietal cortex in PPC, however, is suggested to be involved in future reward (Tanaka et al., 2004).

Unfortunately, there are few studies which investigate the neural mechanisms involved in unexpected uncertainty (Daw et al., 2006). One reason is that researchers realized only recently that there might be a distinction between these types of uncertainty in the brain. For the first time in 2005, Yu and Dayan suggested that distinct neurotransmitters might be important for different types of uncertainty. They showed that acetylcholine is specialized for expected uncertainty and norepinephrine for unexpected uncertainty (Yu & Dayan, 2005). LeNestour and Bossaerts also predict that the amygdale-hippocampus complex, precuneus and anterior cingulate cortex might be involved in unexpected uncertainty (Payzan-LeNestour & Bossaerts, 2011).

Going forward, we hope to modify our model according to what has been suggested about estimating uncertainty in the brain. This short review suggests that modifying the action selection rule (soft-max rule) and learning rate parameters might be a reasonable starting point. We also believe that we need a task that enables us to be certain that participants learn a first model, in order to investigate the first step in the updating process. This is necessary, because, as the literature reviewed above suggests, the insula and the ACC are involved in this first step of updating. Moreover, the insula is

commonly damaged in our RBD patients. This result persuades us to believe that this step of updating might also be impaired in neglect patients. Secondly, this new task must make it easy to interpret what model participants pick even if they pick an incorrect model; RPS, unfortunately, is too complex for this purpose. To study more about the possible reasons behind a failure it is always a great help to know what alternatives have been picked instead of the right one. Nevertheless, in our task, more investigation about the possible differences between brain damaged patients is not simple since there are several incorrect alternative models imaginable for our RBD patients. It is almost impossible to keep track of our participants' hypotheses only by observing their sequence of plays. A proper task with a limited amount of alternatives gives us this opportunity to study more about their impairment by investigating the wrong models they select.

Bibliography

Adams, R.P., & MacKay, D.J. (2007) Bayesian online changepoint detection (University of Cambridge Technical Report, Cambridge, UK).

Albert, M. L. (1973). A simple test of visual neglect. *Neurology, 23*(6), 658-664.

Allen, G., Buxton, R. B., Wong, E. C., & Courchesne, E. (1997). Attentional activation of the cerebellum independent of motor involvement. *Science, 275*(5308), 1940-1943.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science, 15*(2), 106-111.

Anderson, J. J. R., & Lebiere, C. J. (1998). *The atomic components of thought* Psychology Press.

Anderson, J. R., & Matessa, M. (1997). A production system theory of serial memory. *Psychological Review, 104*(4), 728.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*(6), 396-408.

Anderson, R. B., & Tweney, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition, 25*(5), 724-730.

Aslin, R. N., & Newport, E. L. (2012). Statistical learning from acquiring specific items to forming general rules. *Current Directions in Psychological Science, 21*(3), 170-176.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*(4), 321-324.

Bach, D. R., & Dolan, R. J. (2012). Knowing how much you don't know: A neural organization of uncertainty estimates. *Nature Reviews Neuroscience, 13*(8), 572-586.

Baddeley, A. (1986). Working memory (vol. 11).

Basso, E. W., & Engel, P. M. (2009). Reinforcement learning in non-stationary continuous time and space scenarios. *Artificial Intelligence National Meeting (Enia),7,* 1-8.

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

Berns, G. S., & Sejnowski, T. J. (1998). A computational model of how the basal ganglia produce sequences. *Journal of Cognitive Neuroscience, 10*(1), 108-121.

Bisiach, E., Capitani, E., Nichelli, P., & Spinnler, H. (1976). Recognition of overlapping patterns and focal hemisphere damage. *Neuropsychologia, 14*(3), 375-379.

Bland, A. R., & Schaefer, A. (2012). Different varieties of uncertainty in human decision-making. *Frontiers in Neuroscience, 6*, 85-96.

Bossaerts, P. (2010). Risk and risk prediction error signals in anterior insula. *Brain Structure and Function, 214*(5-6), 645-653.

Bostan, A. C., Dum, R. P., & Strick, P. L. (2013). Cerebellar networks with the cerebral cortex and basal ganglia. *Trends in Cognitive Sciences, 17*(5), 241–254.

Brownell, H. H., Potter, H. H., Bihrle, A. M., & Gardner, H. (1986). Inference deficits in right brain-damaged patients. *Brain and Language, 27*(2), 310-321.

Callicott, J. H., Mattay, V. S., Bertolino, A., Finn, K., Coppola, R., Frank, J. A., et al. (1999). Physiological characteristics of capacity constraints in working memory as revealed by functional MRI. *Cerebral Cortex, 9*(1), 20-26.

Chen, S., Cowan, C., & Grant, P. (1991). Orthogonal least squares learning algorithm for radial basis function networks. *Neural Networks, IEEE Transactions on, 2*(2), 302-309.

Cohen, J. D., McClure, S. M., & Angela, J. Y. (2007). Should I stay or should I go? how the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1481), 933-942.

Conway, C. M., Bauernschmidt, A., Huang, S. S., & Pisoni, D. B. (2010). Implicit statistical learning in language processing: Word predictability is the key. *Cognition, 114*(3), 356-371.

Cover, T. M., Thomas, J. A., & Kieffer, J. (1994). Elements of information theory. *SIAM Review, 36*(3), 509-510.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research, 169*, 323-338.

d'Acremont, M., Fornari, E., & Bossaerts, P. (2013). Activity in inferior parietal and medial prefrontal cortex signals the accumulation of evidence in a probability learning task. *PLoS Computational Biology, 9*(1), e1002895.

Danckert, J., Ferber, S., Doherty, T., Steinmetz, H., Nicolle, D., & Goodale, M. A. (2002). Selective, non-lateralized impairment of motor imagery following right parietal damage. *Neurocase, 8*(2), 194-204.

Danckert, J., Ferber, S., Pun, C., Broderick, C., Striemer, C., Rock, S., et al. (2007). Neglected time: Impaired temporal perception of multisecond intervals in unilateral neglect. *Journal of Cognitive Neuroscience, 19*(10), 1706-1720.

Danckert, J., Stöttinger, E., Quehl, N., & Anderson, B. (2012). Right hemisphere brain damage impairs strategy updating. *Cerebral Cortex, 22*(12), 2745-2760.

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

Dayan, P. (1999). Unsupervised learning. *The MIT Encyclopedia of the Cognitive Sciences,*

Decety, J., Jeannerod, M., & Prablanc, C. (1989). The timing of mentally represented actions. *Behavioural Brain Research, 34*(1), 35-42.

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

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences, 11*(8), 333-341.

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

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology, 10*(6), 732-739.

Doya, K. (2007). Reinforcement learning: Computational theory and biological mechanisms, *Hfsp j 1*(1), 30-40.

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

Driver, J., & Mattingley, J. B. (1998). Parietal neglect and visual awareness. *Nature Neuroscience, 1*(1), 17-22.

Duhame, J., Goldberg, M. E., Fitzgibbon, E. J., Sirigu, A., & Grafman, J. (1992). Saccadic dysmetria in a patient with a right frontoparietal lesion the importance of corollary discharge for accurate spatial behaviour. *Brain, 115*(5), 1387-1402.

Ebbinghaus, H. (1885). 1964. *Memory: A Contribution to Experimental Psychology,*

Fellows, L. K. (2004). The cognitive neuroscience of human decision making: A review and conceptual framework. *Behavioral and Cognitive Neuroscience Reviews, 3*(3), 159-172.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12*(6), 499-504.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology, 47*(6), 381.

Fletcher, P., Anderson, J., Shanks, D., Honey, R., Carpenter, T., Donovan, T., et al. (2001). Responses of human frontal cortex to surprising events are predicted by formal associative learning theory. *Nature Neuroscience, 4*(10), 1043-1048.

Fuster, J. (2008). *The prefrontal cortex* Access Online via Elsevier.

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition, 109*(3), 416-422.

Gauvain, J., & Lee, C. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *Speech and Audio Processing, Ieee Transactions on, 2*(2), 291-298.

Gazzaniga, M. S. (1995). Principles of human brain organization derived from split-brain studies. *Neuron, 14*(2), 217-228.

Ghahramani, Z. (2004). Unsupervised learning. *Advanced lectures on machine learning* (pp. 72-112) Springer.

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.

Glickstein, M. (1992). The cerebellum and motor learning. *Current Opinion in Neurobiology, 2*(6), 802-806.

Goldman-Rakic, P. S. (1996). Regional and cellular fractionation of working memory. *Proceedings of the National Academy of Sciences, 93*(24), 13473-13480.

Gray, J. R., & Braver, T. S. (2002). Trait emotion predicts cognitive activation in caudal anterior cingulate cortex. *Cognitive, Affective & Behavioral Neuroscience*,

Green, C., Benson, C., Kersten, D., & Schrater, P. (2010). Alterations in choice behavior by manipulations of world model. *Proceedings of the National Academy of Sciences, 107*(37), 16401-16406.

Griffin, R., Friedman, O., Ween, J., Winner, E., Happé, F., & Brownell, H. (2006). Theory of mind and the right cerebral hemisphere: Refining the scope of impairment. *Laterality, 11*(03), 195-225.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17*(9), 767-773.

Happé, F., Brownell, H., & Winner, E. (1999). Acquiredtheory of mind'impairments following stroke. *Cognition, 70*(3), 211-240.

Heide, W., Blankenburg, M., Zimmermann, E., & Kömpf, D. (1995). Cortical control of double-step saccades: Implications for spatial orientation. *Annals of Neurology, 38*(5), 739-748.

Heilman, K. M., Bowers, D., Valenstein, E., & Watson, R. T. (1987). Hemispace and hemispatial neglect. *Advances in Psychology, 45*, 115-150.

Houk, J. C., & Wise, S. P. (1995). Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: their role in planning and controlling action. *Cerebral Cortex, 5*(2), 95-110.

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., & Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. *Science, 310*(5754), 1680-1683.

Huettel, S. A., Song, A. W., & McCarthy, G. (2005). Decisions under uncertainty: Probabilistic context influences activation of prefrontal and parietal cortices. *The Journal of Neuroscience, 25*(13), 3304-3311.

Ito, M. (2000). Mechanisms of motor learning in the cerebellum. *Brain Research, 886*(1), 237-245.

Ivry, R. B., & Baldo, J. V. (1992). Is the cerebellum involved in learning and cognition? *Current Opinion in Neurobiology, 2*(2), 212-216.

Jeannerod, M. (1997). *The cognitive neuroscience of action.* Blackwell Publishing.

Jenkins, P., Earle-Richardson, G., Slingerland, D. T., & May, J. (2002). Time dependent memory decay. *American Journal of Industrial Medicine, 41*(2), 98-101.

Jensen, S., Boley, D., Gini, M., & Schrater, P. (2005). Non-stationary policy learning in 2-player zero sum games. *Proceedings of the National Conference on Artificial Intelligence, , 20.* (2) pp. 789.

Jonides, J., Schumacher, E. H., Smith, E. E., Lauber, E. J., Awh, E., Minoshima, S., et al. (1997). Verbal working memory load affects regional brain activation as measured by PET. *Journal of Cognitive Neuroscience, 9*(4), 462-475.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99,* 122-149.

Kangas, B. D., Berry, M. S., Cassidy, R. N., Dallery, J., Vaidya, M., & Hackenberg, T. D. (2009). Concurrent performance in a three-alternative choice situation: Response allocation in a Rock/Paper/Scissors game. *Behavioural Processes, 82*(2), 164-172.

Karnath, H., Berger, M. F., Küker, W., & Rorden, C. (2004). The anatomy of spatial neglect based on voxelwise statistical analysis: A study of 140 patients. *Cerebral Cortex, 14*(10), 1164-1172.

Karnath, H., Ferber, S., & Himmelbach, M. (2001). Spatial awareness is a function of the temporal not the posterior parietal lobe. *Nature, 411*(6840), 950-953.

Karnath, H., Himmelbach, M., & Rorden, C. (2002). The subcortical anatomy of human spatial neglect: Putamen, caudate nucleus and pulvinar. *Brain, 125*(2), 350-360.

Kim, S., Ugurbil, K., & Strick, P. (1994). Activation of a cerebellar output nucleus during cognitive processing. *Science, 265*(5174), 949-951.

King B. R., Fogel, S. M., Albouy, G., & Doyon, J. (2013).Neural correlates of the age-related changes in motor sequence learning and motor adaptation in older adults, *Frontiers in human neuroscience,* 7:142.

Koehler, D. J., & James, G. (2009). Probability matching in choice under uncertainty: Intuition versus deliberation. *Cognition, 113*(1), 123-127.

Kristjansson, A. (2008). I know what you did on the last trial—a selective review of research on priming in visual search. *Frontiers in Bioscience, 13,* 1171-1181.

Kristjánsson, Á., & Campana, G. (2010). Where perception meets memory: A review of repetition priming in visual search tasks. *Attention, Perception, & Psychophysics, 72*(1), 5-18.

Krugel, L. K., Biele, G., Mohr, P. N., Li, S., & Heekeren, H. R. (2009). Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proceedings of the National Academy of Sciences, 106*(42), 17951-17956.

Lee, D., Conroy, M. L., McGreevy, B. P., & Barraclough, D. J. (2004). Reinforcement learning and decision making in monkeys during a competitive game. *Cognitive Brain Research, 22*(1), 45-58.

Lee, D., McGreevy, B. P., & Barraclough, D. J. (2005). Learning and decision making in monkeys during a rock–paper–scissors game. *Cognitive Brain Research, 25*(2), 416-430.

Lee, D., & Seo, H. (2007). Mechanisms of reinforcement learning and decision making in the primate dorsolateral prefrontal cortex. *Annals of the New York Academy of Sciences, 1104*(1), 108-122.

Losier, B.J.W., & Klein, R.M. (2001). A review of the evidence for a disengage deficit following parietal lobe damage. *Neuroscience & Biobehavioral Reviews, 25*(1), 1-13.

Maljkovic, V., & Nakayama, K. (1994). Priming of pop-out: I. role of features. *Memory & Cognition, 22*(6), 657-672.

Maljkovic, V., & Nakayama, K. (1996). Priming of pop-out: II. the role of position. *Perception & Psychophysics, 58*(7), 977-991.

Marois, R., & Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in Cognitive Sciences, 9*(6), 296-305.

McElree, B. (2001). Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(3), 817.

Merrifield, C., Hurwitz, M., & Danckert, J. (2010). Multimodal temporal perception deficits in a patient with left spatial neglect. *Cognitive Neuroscience, 1*(4), 244-253.

Mesulam, M. (1985). Attention, confusional states, and neglect. *Principles of Behavioral Neurology, 3*, 125-168.

Meyer, T., & Olson, C. R. (2011). Statistical learning of visual transitions in monkey inferotemporal cortex. *Proceedings of the National Academy of Sciences, 108*(48), 19401-19406.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). Plans and the structure of behavior. holt, rinehart and winston. *Inc., New York,*

Miller, M. B., Valsangkar-Smyth, M., Newman, S., Dumont, H., & Wolford, G. (2005). Brain activations associated with probability matching. *Neuropsychologia, 43*(11), 1598-1608.

Mirman, D., Magnuson, J. S., Estes, K. G., & Dixon, J. A. (2008). The link between statistical segmentation and word learning in adults. *Cognition, 108*(1), 271-280.

Moghaddam, B., & Pentland, A. (1997). Probabilistic visual learning for object representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 19*(7), 696-710.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47*(1), 90-100.

Nassar, M. R., Wilson, R. C., Heasly, B., & Gold, J. I. (2010). An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience, 30*(37), 12366-12378.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. *Basic Processes in Reading: Visual Word Recognition, 11*

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. *Cognitive Skills and their Acquisition,* , 1-55.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. *Human Brain Mapping, 25*(1), 46-59.

Paulin, M. G. (1993). The role of the cerebellum in motor control and perception. *Brain, Behavior and Evolution, 41*(1), 39-50.

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology, 7*(1), e1001048.

Payzan-LeNestour, E. (2010). Bayesian learning in unstable settings: Experimental evidence based on the bandit problem. *Swiss Finance Institute Research Paper, 10*(28), 1–41.

Posner, M. I., and Cohen, Y. (1984). Components of visual orienting. *In: Bouma H, Bowhuis D, editors. Attention and Performance X. Hillsdale, NJ: Erlboum*, 531-556.

Poupart, P., Vlassis, N., Hoey, J., & Regan, K. (2006). An analytic solution to discrete bayesian reinforcement learning. *Proceedings of the 23rd International Conference on Machine Learning*, pp. 697-704.

Rabiner, L., & Juang, B. (1986). An introduction to hidden markov models. *ASSP Magazine, IEEE, 3*(1), 4-16.

Ramnani, N. (2006). The primate cortico-cerebellar system: Anatomy and function. *Nature Reviews Neuroscience, 7*(7), 511-522.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature Neuroscience, 3*, 1199-1204.

Rose, S. A., Feldman, J. F., & Jankowski, J. J. (2004). Infant visual recognition memory. *Developmental Review, 24*(1), 74-100.

Roser, M. E., Fiser, J., Aslin, R. N., & Gazzaniga, M. S. (2011). Right hemisphere dominance in visual statistical learning. *Journal of Cognitive Neuroscience, 23*(5), 1088-1099.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review; Psychological Review, 103*(4), 734.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926-1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*(1), 27-52.

Sato, Y., Akiyama, E., & Farmer, J. D. (2002). Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences, 99*(7), 4748-4751.

Schenkenberg, T., Bradford, D., & Ajax, E. (1980). Line bisection and unilateral visual neglect in patients with neurologic impairment. *Neurology, 30*(5), 509-509.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology, 80*(1), 1-27.

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *The Journal of Neuroscience, 13*(3), 900-913.

Shallice, T. (1988). *From neuropsychology to mental structure* Cambridge University Press.

Shamay-Tsoory, S. G., Tomer, R., Berger, B., & Aharon-Peretz, J. (2003). Characterization of empathy deficits following prefrontal brain damage: The role of the right ventromedial prefrontal cortex. *Journal of Cognitive Neuroscience, 15*(3), 324-337.

Shaqiri, A., & Anderson, B. (2012). Spatial probability cuing and right hemisphere damage. *Brain and Cognition, 80*(3), 352-360.

Sirigu, A., Duhamel, J., Cohen, L., Pillon, B., Dubois, B., & Agid, Y. (1996). The mental representation of hand movements after parietal cortex damage. *Science-New York then Washington-, ,* 1564-1568.

Stein, J., & Glickstein, M. (1992). Role of the cerebellum in visual guidance of movement. *Physiological Reviews, 72*(4), 967-1017.

Stern, E. R., Gonzalez, R., Welsh, R. C., & Taylor, S. F. (2010). Updating beliefs for a decision: Neural correlates of uncertainty and underconfidence. *The Journal of Neuroscience, 30*(23), 8032-8041.

Striemer, C., & Danckert, J. (2007). Prism adaptation reduces the disengage deficit in right brain damage patients. *Neuroreport, 18*(1), 99-103.

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

Tanaka, S. C., Doya, K., Okada, G., Ueda, K., Okamoto, Y., & Yamawaki, S. (2004). Prediction of immediate and future rewards differentially recruits cortico-basal ganglia loops. *Nature Neuroscience, 7*(8), 887-893.

Thrun, S. B. (1992). The role of exploration in learning control. *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches.Van Nostrand Reinhold, New York,*

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2007). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *Journal of Neurophysiology, 97*(2), 1621-1632.

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature, 428*(6984), 751-754.

Trimmer, P. C., Houston, A. I., Marshall, J. A., Mendl, M. T., Paul, E. S., & McNamara, J. M. (2011). Decision-making under uncertainty: Biases and bayesians. *Animal Cognition, 14*(4), 465-476.

Tsumoto, T., & Suda, K. (1979). Cross-depression: An electrophysiological manifestation of binocular competition in the developing visual cortex. *Brain Research, 168*(1), 190-194.

Turk-Browne, N. B., Isola, P. J., Scholl, B. J., & Treat, T. A. (2008). Multidimensional visual statistical learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*(2), 399.

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition, 92*(1), 231-270.

Unturbe, J., & Corominas, J. (2007). Probability matching involves rule-generating ability: A neuropsychological mechanism dealing with probabilities. *Neuropsychology, 21*(5), 621.

Vallar, G., & Perani, D. (1986). The anatomy of unilateral neglect after right-hemisphere stroke lesions. A clinical/CT-scan correlation study in man. *Neuropsychologia, 24*(5), 609-622.

Vickery, T. J., & Jiang, Y. V. (2009). Inferior parietal lobule supports decision making under uncertainty in humans. *Cerebral Cortex, 19*(4), 916-925.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys, 14*(1), 101-118.

Walthew, C., & Gilchrist, I. D. (2006). Target location probability effects in visual search: An effect of sequential dependencies. *Journal of Experimental Psychology: Human Perception and Performance, 32*(5), 1294.

Wig, G. S. (2012). Repetition suppression and repetition priming are processing outcomes. *Cognitive Neuroscience, 3*(3-4), 247-248.

Wilson, B., Cockburn, J., & Halligan, P. (1987). Development of a behavioral test of visuospatial neglect. *Archives of Physical Medicine and Rehabilitation, 68*(2), 98-102.

Wilson, R.C., & Niv, Y. (2011). Inferring relevance in a changing world. *Frontiers in human neuroscience, 5*: 189.

Winner, E., Brownell, H., Happé, F., Blum, A., & Pincus, D. (1998). Distinguishing lies from jokes: Theory of mind deficits and discourse interpretation in right hemisphere brain-damaged patients. *Brain and Language, 62*(1), 89-106.

Wise, R. A., Spindler, J., & Gerberg, G. (1978). Neuroleptic-induced" anhedonia" in rats: Pimozide blocks reward quality of food. *Science, 201*(4352), 262-264.

Wise, R. A., Spindler, J., & Legault, L. (1978). Major attenuation of food reward with performance-sparing doses of pimozide in the rat. *Canadian Journal of Psychology/Revue Canadienne De Psychologie, 32*(2), 77.

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science, 2*(6), 409-415.

Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition, 25*(5), 731-739.

Wolford, G., Miller, M. B., & Gazzaniga, M. (2000). The left hemisphere's role in hypothesis formation. *The Journal of Neuroscience,*

Wolford, G., Newman, S. E., Miller, M. B., & Wig, G. S. (2004). Searching for patterns in random sequences. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expé Rimentale, 58*(4), 221.

Wolpert, D. M., Ghahramani, Z., & Flanagan, J. R. (2001). Perspectives and problems in motor learning. *Trends in Cognitive Sciences, 5*(11), 487-494.

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron, 46*(4), 681-692.