Asymmetric Learning Rates for Positive and Negative Feedback:

A Formal Model Comparison

Kelsey McDonald

Princeton University

This thesis was submitted to Princeton University in partial fulfillment of the requirements for the Degree of Bachelor of Arts in Psychology & Certificate in Neuroscience

Adviser: Yael Niv

April 13, 2015

Honor Pledge

I pledge my honor that this paper represents my own work in accordance with University regulations.

Kelsey McDonald

Acknowledgements

The process of writing this thesis has been a humbling one. I am so grateful to the many scientists, mentors, friends, and family members who have contributed to this work. I am incredibly thankful to have had Professor Yael Niv as an advisor, mentor, and teacher. She has provided me with an immense amount of guidance throughout my independent work at Princeton and has inspired me to pursue a career in psychology and neuroscience research. I would also like to thank Nicolas Schuck for being such an amazing and dedicated mentor. I have learned so much from both Nicolas and Yael, and I hope one day I can mentor other aspiring scientists with the amount of dedication, knowledge, and patience in which they have mentored me. I would also like to thank everyone in the Niv Lab, especially Eran Eldar, for providing so much help, advice, and support. Thank you Niv Lab! Also, a special thank you to Ben Eppinger for authorizing the release of the original dataset from Eppinger et al., (2013).

I would not have been able to write this thesis without the amazing support of my family, friends, and Terrace F. Club (FOOD = LOVE). I especially received an immense amount of support and love from my amazing boyfriend Edward “Eddie Boy” Paul and our beautiful Husky puppy Khaleesi. Finally, I am thankful for the financial support provided by both Princeton’s Department of Psychology and the Office of the Dean of the College who graciously granted two senior thesis grants through the Horton Elmer 1942/1992 Fund which has made this thesis possible. So many people have offered their support, mentorship, guidance, and love throughout the writing of this thesis, and I am forever grateful.

Abstract

Reinforcement learning studies how learning systems interact with their environments in order to maximize a numerical reward signal. One of the central concepts in reinforcement learning is the reward "prediction error", which is the numerical value of the reward received minus the expected reward value. The classic Rescorla-Wagner model posits that the learning agent updates their original reward estimate by stepwise error-correction: multiplying the prediction error by a learning rate parameter. The main limitation with the Rescorla-Wagner model is the implied valence symmetry with which feedback updates an action's value estimate. This contradicts evidence that learning from positive and negative feedback has different effects on behavior and decision-making. In this thesis, I conduct a formal model comparison of the Rescorla-Wagner model with an alternative class of asymmetric learning models which discriminate based on valence. Our analyses show that behavioral choice data in a probabilistic learning experiment is more accurately described by an asymmetric learning algorithm rather than a symmetric learning rule which does not discriminate based on valence.

TABLE OF CONTENTS

Chapter 1 Introduction: Learning How to Maximize Reward

One of the most fundamental goals in behavioral neuroscience is understanding the decision-making processes that humans and animals use to learn how to select actions that will maximize reward or minimize punishment. There are many instances in the real world in which this learning occurs via trial-and-error, such as a gambler in Las Vegas choosing which arm on a slot machine will yield the most points or a busy commuter deciding which route will minimize travel time. One methodological approach to these learning problems is to associate each potential action, given the current state, with an estimated scalar value representing the expected reward of choosing that action. Recent efforts by computational neuroscientists have attempted to formally describe learning rules that the agent uses to estimate and update an action's expected reward value.

The goal of this thesis is to compare candidate learning rules to determine which model more accurately describes participant learning behavior in two reinforcement learning tasks. We hypothesize that an asymmetric learning model that discriminates learning and value updating based on the valence of feedback will better describe human learning and decision making than a symmetric learning model which does not discriminate based on feedback valence. In this chapter, I will first introduce the formal reinforcement learning framework, followed by a review of evidence connecting the principles of reinforcement learning to neural correlates in the brain. I will then explore the biological and psychological research supporting asymmetric valence learning.

1.1. Introduction to Reinforcement Learning

Reinforcement learning studies how biological and artificial learning systems interact with their environments so as to learn how to maximize a numerical reward signal, either in the form of obtaining rewards or avoiding punishments (Sutton & Barto, 1998). Rather than being explicitly told which actions are best in each situation, as in many forms of machine learning and supervised learning, in reinforcement learning the agent learns which actions yield the most reward through trial and error. In behavioral psychology, reinforcement learning has been investigated through the lenses of both classical conditioning and instrumental, or operant, conditioning. Derived from machine learning, formal reinforcement learning theory seeks to provide algorithms that serve as a mathematical and systematic solution to the reward optimization problem faced by learning agents (Sutton & Barto, 1998).

The learning agent's decision-making environment consists of states, actions, and outcomes (Dayan & Niv, 2008; Sutton & Barto, 1998). States are internal or external situations, such as positions in a game or locations in a maze. Given which state a learning agent is in, a choice will be selected from a set of available actions that will move the agent to another state and potentially provide it with a non-zero numerical outcome, either a reward (positive outcome) or a punishment (negative outcome). These outcomes can change the learner's behavior and future actions.

An important concept in reinforcement learning is error-driven learning. The value of an action or stimulus is the prediction of future reward expected from choosing that action or stimulus. The scalar difference value between estimation and realization is termed the reward "prediction error" (Niv & Schoenbaum, 2008), defined mathematically as:

$$\delta_t = R_t - V_t \quad \text{(Equation 1.1)}$$

where δ_t is the prediction error in trial t, R_t is the reward received, and V_t is the expected value of the state before action selection. The concept of a reward prediction error is foundational to how agents learn to forecast outcomes of future events based on past and current information (Behrens et al., 2007; Niv & Schoenbaum, 2008; Pearce & Hall, 1980; Rescorla and Wagner, 1972). A reward prediction error is positive when an outcome is better than expected, and negative when worse than expected. When a learning agent's value estimate of a particular action is not equal to the reward received by selecting that action, the agent updates its estimate in proportion to the prediction error. The goal for many reinforcement learning algorithms, including those presented in this thesis, is to maximize reward by learning to generate correct value predictions.

In summary, reinforcement learning is the study of how learning agents maximize reward via trial and error. Each of the learning rules analyzed in this thesis are based on error-driven learning. In the next section, I will discuss important lines of research connecting the principles and findings of reinforcement learning to the function of dopaminergic neurons in the midbrain as well as to human neuroimaging studies.

1.2. Reinforcement Learning in the Brain

The reward prediction error hypothesis of dopamine is a theoretical framework describing the close correspondence between phasic dopaminergic firing patterns and the characteristics of a reward prediction error (Montague et al., 1996; Schultz et al., 1997). Many electrophysiological studies using single-cell recordings suggest that the phasic activity of midbrain dopaminergic cells encode a scalar prediction error signal in the brain (Niv et al., 2005;

Schultz et al., 1997). In line with computational learning models, dopaminergic neurons respond to reward early in training when the reward was unexpected, but not later in learning after the reward was predicted (Schultz et al., 1997). As shown in Figure 1.1, phasic dopamine signals reliably transfer during learning from the onset of an unpredicted reward to the onset of a conditioned stimulus learned to predict the reward (Niv et al. 2005; Schultz et al., 1997). Further evidence shows that when dopamine responses were analyzed on a trial-by-trial basis, firing patterns on each trial coincided with the prediction error computed by reinforcement learning models (Bayer & Glimcher, 2005).

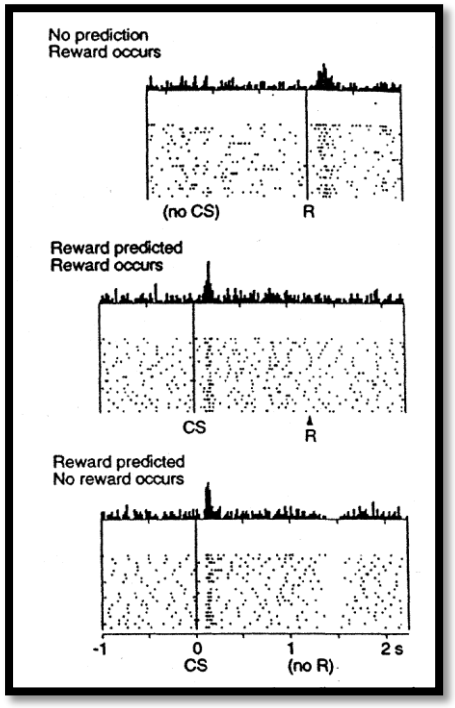

Figure 1.1. Firing patterns of dopaminergic neurons in monkeys performing an instrumental conditioning task. Each histogram shows action potentials with different rows representing differential trials aligned on the time of the reward or cue. Before learning (top), presentation of a reward generates a positive reward prediction error. After learning (middle), the conditioned stimulus predicts reward and no prediction error is generated. After learning (bottom), if a conditioned stimulus is presented, signaling a reward, but the reward is not presented, the activity of the dopamine neuron is depressed exactly at the time when the reward would have occurred. Figure from Schultz et al., 1997.

Reinforcement learning computational models reliably describe phasic dopaminergic activity when a positive prediction error is generated; however, there is less of a consensus regarding the relationship between dopaminergic activity and negative reward prediction error signaling. Positive and negative prediction errors are not similarly represented, physiologically, due to dopamine's low basal firing rate. An unexpected reward presentation (positive prediction error) may increase firing rates to 20 or 30 Hz. However, dopamine neurons have a 3- to 5-Hz baseline, so omitting the same reward will decrease the firing rate temporarily to 0 Hz, a decrease of only 3-5 Hz in total. This means that a positive prediction error can be represented by a large range (approximately 270% above baseline), whereas a negative prediction error can only be represented by a decrease of approximately ~55% below baseline (Fiorillo et al., 2003). This physiological asymmetry in the neural coding of prediction errors lends support for a learning model that updates value estimates differently depending on the valence of the prediction error.

Because of the invasiveness of single-cell recordings, functional magnetic resonance imaging (fMRI) has been a methodology widely used by researchers to investigate neural processes associated with reinforcement learning in the human brain. fMRI measures the blood-oxygen-level-dependent (BOLD) signal, which is thought to be a correlate of brain activity. Neural correlates of reward prediction error signals have been found in various regions of the brain, including midbrain dopaminergic neurons, striatum, amygdala, and prefrontal cortex (Chakravarthy et al., 2010; Niv & Schoenbaum, 2008). fMRI studies have implicated the nucleus accumbens and the orbitofrontal cortex (OFC) as brain areas responsible for coding prediction errors, since these areas have showed neural activation to be modulated by the predictability of reward delivery (Berns et al. 2001). fMRI regression analyses have revealed a significant

correlation between prediction error signaling and activity in the ventral striatum (a subcortical area that receives strong projections from midbrain dopaminergic neurons) and the orbitofrontal cortex (O'Doherty et al. 2003).

Researchers have also found supporting evidence for a topographical difference in the representation of aversive and appetitive reward signals using fMRI. Seymour et al. (2007) conducted an fMRI study showing functional segregation within the striatum, in which the more anterior regions showed selectivity for rewards and the more posterior regions for losses. A similar topographical valence-specific gradient was found in the OFC, in which the lateral OFC is activated following a punishing outcome, and the medial OFC is activated following a rewarding outcome (O'Doherty et al., 2001). A comprehensive meta-analysis of 142 neuroimaging studies show that the nucleus accumbens (NAcc) is often activated by both positive and negative outcomes during anticipation, outcome, and evaluation of reward processing (Liu et al., 2011). The medial OFC and posterior cingulate cortex, in the meta-analysis, preferentially responded to positive rewards, whereas the anterior cingulate cortex, bilateral anterior insula, and lateral prefrontal cortex preferentially responded to negative rewards. Another meta-analysis noted that reward and aversive prediction errors were coded throughout the brain by broadly segregated neural networks with minimal integration between positive and negative prediction errors in the ventral striatum and antero-medial cingulate cortex (Garrison et al., 2013).

In summary, the prediction error hypothesis of dopamine has provided a link between principles described in computational reinforcement learning models and neural substrates, especially in regards to phasic dopaminergic signals in the mammalian midbrain. Despite this progress, there is still a debate as to how signed prediction errors are represented in the brain.

Due to the low basal firing rate of dopaminergic neurons, the physiological representation of positive and negative prediction errors are asymmetric. Neural correlates of reinforcement learning both provide evidence for the validity of reinforcement learning algorithms, as well as support for our hypothesis that positive and negative feedback may be represented differentially in the brain. In the next section, we will investigate further lines of research supporting asymmetric valence learning models.

1.3. Evidence for Asymmetric Learning Models

Many reinforcement learning algorithms generally update value estimates of an action by multiplying the generated prediction error following action selection by a single learning rate, implicitly indiscriminate of feedback valence. However, behavioral, neural, and physiological data indicate that outcomes that are either better or worse than expected usually do not have symmetric impacts on human learning and decision-making. This asymmetry has been proposed in explaining age-related changes in learning (Eppinger et al., 2013), deficits or biases in learning due to psychiatric disorders (Frank et al., 2007; Frank et al., 2004), as well as systematic deviations from "rational" decision-making (Niv 2009; Niv & Montague, 2008).

Asymmetric learning models are in line with physiological findings mentioned previously regarding the low basal firing rates of dopaminergic neurons. The ventral tegmental area, one of the primary regions of dopamine synthesis in the human brain, has been found to contain separate populations of neurons that are stimulated and inhibited by appetitive and aversive stimuli (Matsumoto & Hikosaka, 2009). Further, striatal D1 and D2 dopamine receptors have been shown to have different response patterns to positive and negative rewards (Frank et al., 2007). Bayer & Glimcher (2005) examined activity of single dopamine neurons during a reinforcement learning task and concluded that these neurons encoded the difference between the

current reward and a weighted average of previous rewards, a reward prediction error, but only for outcomes that were better than expected. Thus, the average firing rate of dopamine neurons in the post-reward interval accurately carried information about positive reward prediction errors but not about negative reward prediction errors (Bayer & Glimcher, 2005). Clinical data show that patients with Parkinson's disease, who have low striatal dopamine levels, who are either on or off medication have different deficits in reinforcement learning tasks and asymmetrically biased learning from positive or negative outcomes (Frank et al. 2004). Meaning, different concentrations of striatal dopamine manifest in different reinforcement learning deficits which are mediated by feedback valence.

Asymmetric responses to positive and negative feedback also manifest behaviorally. Prospect theory, proposed by Kahneman & Tversky, is an influential behavioral economic theory of decision making under uncertainty and risk. A behavioral phenomenon particularly well-described by Prospect theory is that of loss aversion, in which people have a tendency to prefer avoiding losses rather than accruing gains of the same magnitude (Kahneman & Tversky, 1979). Empirical data show that the loss-to-gain sensitivity ratio is usually 2:1. For example, a loss of $100 is more aversive than a gain of $100 is rewarding. This is in stark contrast with rational economic theory in which gaining or losing a fixed magnitude should have a similar, but opposite, effect on a decision-maker's utility (Kahneman & Tversky, 1983).

Tom et al. (2007) investigated the neural correlates of loss aversion in an fMRI study by observing brain activity while participants accepted or rejected mixed gambles. As potential gains increased, a host of brain areas including midbrain dopaminergic areas and their targets increased in activity. Many of these same brain regions, such as the striatum, ventromedial prefrontal cortex, and medial OFC, also showed decreased activity as potential losses increased.

Interestingly, the behavioral and neural indexes of loss aversion were significantly correlated (r = 0.85) in regions such as the ventral striatum and prefrontal cortex. Mediation analysis suggested that individual differences in behavioral loss aversion were driven primarily by individual differences in neural loss aversion in these areas (Tom et al., 2007).

Other fMRI studies have provided supporting evidence that the sign of the prediction error matters and is represented differentially in the brain. O'Doherty et al. (2003) separately regressed two learning models that represented the positive and negative components of a temporal difference prediction error at the time of presentation of the unconditioned stimulus. This allowed the authors to detect brain areas in which a negative prediction error yielded a negative BOLD response (differentially-signed response) and, additionally, negative prediction errors which yielded a positive BOLD response (absolute-value response). The results demonstrated that the signed prediction error model showed significant effects, compared to the absolute value prediction error model (O'Doherty et al. 2003).

Recent computational modeling studies have also provided evidence for the superiority of asymmetric valence models in a learning task. Cazé & van der Meer showed that simulated learning agents using asymmetric learning rates can be adaptive in the sense that they earn more reward than agents using symmetric learning rate models (Cazé & van der Meer, 2013). Gershman (2015) conducted a computational modeling experiment which fit several different reinforcement learning models to human behavioral data from a two-choice learning task that manipulated the average learning reward rate across blocks. This study found strong support for a learning model with fixed, separate learning rates for positive and negative prediction errors (Gershman, 2015).

The research explored in this chapter provide support for our hypothesis that asymmetric-valence learning models may be better descriptive learning rules for biological agents than symmetric-valence learning rules. In the next chapter, I describe the candidate learning algorithms compared in this thesis as well as the computational modeling design used to analyze the behavioral learning processes in Experiments 1 and 2.

Chapter 2 General Methods: Trial-by-Trial Modeling of Behavior

This chapter explains the computational modeling procedures that were conducted for Experiments 1 and 2 of this thesis. Both behavioral experiments consisted of a reinforcement learning task in which the participant's goal was to maximize reward by either choosing actions or stimuli that resulted in monetary gain or avoiding actions or stimuli that resulted in monetary loss. The reward and choice data from these studies were then analyzed with four different learning models (described below): the Rescorla-Wagner rule, the Asymmetric Prediction Error rule, the Asymmetric Outcome rule (only in Experiment 2), and the Split-Half rule. The only difference between these learning models is how each rule updates the existing value estimate of an action or stimulus in the presence of a prediction error. A formal model comparison is conducted for both behavioral tasks to determine which of the four models fit the behavioral data the best in each task.

Each model is a variation of the proposal that learning is updated and modulated on a trial-by-trial basis in response to feedback. We use computational models to analyze the behavioral data from both experiments on a trial-by-trial basis, as opposed to activity averaged across many trials. These computational models allow us to formalize qualitative assumptions as quantitative hypotheses that make predictions about behavior on each trial based on the outcomes of preceding choices (Daw 2009). We measure each model's goodness of fit to the data on a trial-by-trial basis, rather than their goodness of fit to averaged data. Each computational model constitutes and proposes a different hypothesis about how agents learn and make decisions in order to maximize reward. Another reason why modeling reinforcement learning theories of behavior is so useful is because these theories allow researchers to quantify variables that otherwise would have been subjective. In other words, on a trial-by-trial basis, we

can quantify and track "hidden variables" in reinforcement learning such as the expected value of a given choice or stimulus, or the prediction error generated after receiving an outcome.

In order to determine the optimal likelihood for each model, we use two built-in MATLAB functions called *fmincon* and *fminsearch*. Given a function to compute the likelihood of data given a model, *fminsearch* and *fmincon* search over all possible values to find the parameters that explain a participant's behavior best. In other words, the function searches for parameters that minimize the likelihood function, which is the probability of obtaining the dataset given a set of parameters. Both functions find the minimum of a given multivariable function; the difference between the two is *fmincon's* ability to constrain the parameter search. We use *fminsearch* to model the data from Experiment 1 and *fmincon* to model the data from Experiment 2 because constraining the learning rate parameter search in the latter experiment resulted in a better parameter optimization, and vice versa for *fminsearch* in Experiment 1. Because *fmincon* and *fminsearch* find the minimum of a function, we use negative log likelihoods as inputs. These functions in conjunction with our model functions call multiple candidate parameters from different starting points, and declare the best answer to be the parameter that results in the lowest negative log likelihood (Daw 2009).

Both experiments seek to model behavioral data from a "bandit" task, a standard decision experiment in reinforcement learning. A bandit task is a commonly used behavioral paradigm in which a learner is faced repeatedly with a choice among *n* different options or actions (Sutton & Barto, 1998). A subject repeatedly chooses between multiple options and receives rewards or punishments according to his or her choice. The *n-armed bandit problem*, so named by analogy to a slot machine or "one-armed bandit", refers to multi-choice reinforcement learning tasks, such as those conducted in Experiments 1 and 2.

2.1. Rescorla-Wagner Model

The Rescorla-Wagner learning rule is an influential model which proposes that learning only occurs when there is a mismatch between the expected outcome and the actual outcome (Rescorla & Wagner, 1972; Sutton & Barto, 1990). Quantitatively, the Rescorla-Wagner model is described as:

$$V_{t+1}(c_t) = V_t(c_t) + \alpha * \delta \qquad \text{(Equation 2.1)}$$

where $V_{t+1}(c_t)$ is the expected value of selecting choice *t*, $V_t(c_t)$ is the estimated value of choice *t* before action selection, α is a learning rate parameter that determines the extent to which the expected value of the current state is updated on each trial, and δ is the prediction error, defined in Equation 1.1. On trial *t*, a participant makes a binary choice c_t between the left or right option, and receives a reward R_t. At the beginning of each trial, the subject assigns an expected value to each option or stimulus. After a reward is delivered, the value for the chosen stimulus or option is updated by multiplying the scalar difference between the reward received and the expected value of a choice by a single learning rate, indiscriminate of valence. In this value update learning rule, there is one free learning rate parameter, $0 \leq \alpha \leq 1$.

Rescorla-Wagner learning is driven by the differences between estimated and actual outcomes—i.e. whenever there is a "surprise" in the reward received. Predictions from different stimuli within one conditioning trial are summed together to create the total predicted value for a given trial. This iterative equation constructs a weighted sum of previous rewards of a given state or action. The learning rate determines the rate at which value updating takes place. High learning rates emphasize the outcomes of more recent learning episodes at the expense of past events while low learning rates more heavily weight reward history compared to the current

outcome, and thus require many learning trials to significantly change an event's value estimate (Behrens et al. 2007; Niv & Schoenbaum, 2008).

The Rescorla-Wagner model explains many learning phenomena such as blocking, overshadowing, and inhibitory conditioning; however, this model has a few limitations (Niv 2009; Sutton & Barto, 1990). First, the Rescorla-Wagner model treats learning as a series of discrete conditioning trials, in which learning is updated at the end of the sometimes arbitrarily-set trial (Niv & Montague, 2008). For purposes of this thesis, however, the main limitation with the Rescorla-Wagner model is based on the implied symmetry with which a prediction error, regardless of valence, updates an action's value estimate. For example, regardless of whether an action generated a reward of five points above (positive prediction error) or below (negative prediction error) an agent's previous estimate, the prediction error will be multiplied by the same learning rate and updated in the same proportion. This is at odds with the physiological data from dopaminergic neurons described in chapter 1 suggesting asymmetric reactions to positive and negative prediction errors, respectively (Fiorillo et al., 2003; Niv et al. 2005; Niv & Schoenbaum, 2008; Schultz et al. 1997). The following two proposed learning models seek to solve this valence symmetry problem.

2.2. Asymmetric Prediction Error Model

The Asymmetric Prediction Error model, see Equation 2.2, is an extension of the Rescorla-Wagner learning rule that allows for two learning rates, α^+ and α^-.

$$V_{t+1}(c_t) = V_t(c_t) + \begin{cases} \alpha^+ * \delta, & if\ \delta \geq 0 \\ \alpha^- * \delta, & if\ \delta < 0 \end{cases} \quad \text{(Equation 2.2)}$$

Every trial calculates the difference between the reward actually received and the model's estimated value, according to equation 1. This prediction error is then used to update the chosen action's value estimate, using α^+ if the prediction error was positive (i.e. outcome was better than expected), or α^- if the prediction error was negative (i.e. outcome was worse than expected). The positive and negative learning rates can be unequal in magnitude. In this learning model, there are two free learning rate parameters, $0 \leq \alpha^+ \leq 1$ and $0 \leq \alpha^- \leq 1$.

2.3. Asymmetric Outcome Model

The Asymmetric Outcome model, see Equation 2.3, is identical to the Asymmetric Prediction Error model described in chapter 2.2, except for the conditions in which the different learning rate parameters are used.

$$V_{t+1}(c_t) = V_t(c_t) + \begin{cases} \alpha^+ * \delta, & if\ R \geq 0 \\ \alpha^- * \delta, & if\ R < 0 \end{cases} \quad \text{(Equation 2.3)}$$

In Equation 2.3, the positive learning parameter is used to update the value estimate when a positive reward was received (R > 0), and the negative learning parameter is used to update the value when a negative reward, or loss, was received (R < 0).

The question of whether a learning agent asymmetrically updates its value estimate based on the valence of the prediction error or the outcome is of theoretical interest. This difference is nontrivial because it is possible for an agent to encounter a scenario in which a loss is delivered (negative outcome) but the prediction error is actually positive (i.e. one loses less than expected). The reward structure in Experiment 1 will not be able to tease these two concepts apart, but the design in Experiment 2 will. Thus, the Asymmetric Outcome rule will not be included in the formal model comparison in Chapter 3, but will be included in Chapter 4.

2.4. Split-Half Model

We include a "Split-Half" model, see below, as a form of control for the asymmetric learning rules. We propose that an asymmetric learning rule more accurately describes human choice data specifically because the accommodation of two discriminating learning rates allows us to model the decision making of agents that have individual valence biases. We posit that these individual valence biases are not random, but in fact reflect systematic differences in how a learning agent evaluates feedback valence and how it uses this biased information to make decisions. Thus, we include a Split-Half model, which generates random asymmetric variation, as a control to show that individual differences in decision making generated from our two asymmetric test models are not random. The Split-Half model is defined as:

$$V_{t+1}(c_t) = V_t(c_t) + \begin{cases} \alpha^+ * \delta, & if \ \mathrm{mod}(t, 2) = 0 \\ \alpha^- * \delta, & if \ mod(t, 2) = 1 \end{cases} \qquad \text{(Equation 2.4)}$$

where α^+ is used to update the value estimate if the current trial number is even and α^- is used to update the value estimate when the current trial number is odd. It is reasonable to assume that the trial number, itself, has no impact on which learning rate is used in this asymmetric model. The Split-Half learning rule allows us to model a two-parameter learning rule with random variation.

One of the hypotheses we will test in both Experiments 1 and 2 is whether an asymmetric learning rule that discriminates based on feedback valence fits human choice data better than a model with the same number of parameters but with random variation. If so, we can conclude that the mechanism responsible for the superiority of an asymmetric learning rule is because the additional variation accounted for by the class of asymmetric learning rules is systematic, not random. This model protects our analysis against overfitting. In other words, a more complicated model with a higher number of free parameters will usually fit better than a simpler model with fewer parameters. A learning rule that generates random variation with the same number of fitted parameters as our test asymmetric models lets us confirm that our models are superior to the symmetric learning rate model specifically because they capture the systematic variation of biased value-updating based on valence, as opposed to random variation.

2.5. Choice Probabilities

To explain a participant's choices C_t in terms of the values V_t of the available actions, we translate subjects' value estimates into choices probabilistically, according to a softmax distribution (Daw 2009; Sutton & Barto, 1998):

$$P(C_t = L \mid V_t(L), V_t(R)) = \frac{\exp(\beta * Vt(L)}{\exp(\beta * Vt(R)) + \exp(\beta * Vt(L)} \quad \text{(Equation 2.5)}$$

Here, given a choice between left or right, the probability of a participant choosing left (given the estimated value of the left and right options) is modeled by a softmax distribution, where β is the inverse temperature parameter. Low inverse temperatures cause the chances of selecting an action among all available actions to be (nearly) equiprobable, whereas high inverse temperatures cause a greater difference in selection probability for actions that differ in their value estimates. The value of beta is important not only for capturing decision noise (i.e. a

participant knows one stimulus has a higher expected value, but still chooses the other option due to mistakes or memory errors) but also for the classic explore/exploit problem in reinforcement learning, in which a learning agent must decide between choosing the stimulus/action with the highest expected reward value (exploit), or learn more information about its environment by choosing a stimulus/action with greater uncertainty (explore) (Sutton & Barto, 1998). In every learning model analyzed in this thesis, the value of beta was conventionally set at $\beta = 3$.

2.6. Model Comparison

The four proposed learning models are identical except for 1) the number of fitted parameters, and in the case of the asymmetric models 2) the condition in which either α^+ or α^- are used for updating. Each of these models makes different assumptions about a learning agent's choice behavior, and can be considered quantitative and testable hypotheses about learning and decision-making processes. By comparing each model's goodness of fit to the participant choice data in Experiments 1 and 2, we can determine how the participants evaluate and update an action's estimated value or reward depending on the valence of feedback.

Given a particular model, the probability of the entire dataset D (i.e. the entire sequence of choices $c = c_1 \ldots T$ and rewards $r = r_1 \ldots T$ for a given participant) is the product of each trial's choice probability from the softmax equation, described in chapter 2.5 (Daw 2009):

$$\Pi_t P(c_t = L \mid V_t(L), V_t(R)) \qquad \text{(Equation 2.6)}$$

Our parameter estimation is based on maximum likelihood. Meaning, our final estimate of a participant's learning rate is the point estimate (θ_M) by which the likelihood function (the probability of our data D, given model M), $P(D|M, \theta_M)$, is maximized. The Rescorla-Wagner rule estimates the single learning rate parameter α, for each participant. The other two asymmetric

learning models and the split-half model fit both α^+ and α^-, again for each participant. The free learning rate parameter estimates of α and its variations are constrained to be $0 \leq \alpha \leq 1$. Our model loops over the data computing the probability of the participant's choices on each trial given the learning model, and returns the aggregate likelihood of the data. Due to the fact that the product of probabilities from the softmax equation is a very small number, we compute the sum over trials of the log of the choice probability for each trial in order to ensure numerical stability (Daw 2009).

Comparing these four proposed learning models on the basis of the likelihood they assign to the given data has a potential pitfall in that it might bias our analyses toward declaring an asymmetric model superior due to the increased number of fitted parameters. This is because the likelihood is computed at parameters chosen to optimize the likelihood function, thus a model with a higher number of free parameters will naturally "fit" the data better than a model with fewer optimized parameters. The Rescorla-Wagner model fits one learning rate, whereas the two asymmetric valence models and the split-half model all fit two learning rates. For our model analysis, the likelihood assigned to the data, given each proposed model, is a flawed method of model comparison since the likelihood is computed with parameters chosen to optimize it (Daw, 2009). The likelihood measure must be corrected for overfitting to allow a fair comparison between the Rescorla-Wagner (with one fitted learning rate) and the other three asymmetric models with two fitted learning rates.

To determine model superiority for purposes of this thesis, we will use the Bayesian Information Criterion (BIC) as a metric of model superiority as another way to prevent overfitting the data (in addition to including the Split-Half model in our analysis). BIC approximates the posterior probability of a model, given behavioral data. The formula for BIC,

see Equation 2.7, is particularly useful for our analysis because it contains a penalty term for models with a higher number of parameters.

$$\log(P(D|M)) \approx \log(P(D|M, \theta_M)) - \frac{n}{2} \log m \qquad \text{(Equation 2.7)}$$

where *m* is the number of data points (in this case, choices) in the dataset, D is the dataset, and M is the model. $P(D|M, \theta_M)$ is the probability of the data, given model M and optimal parameters θ_M. Importantly, the second term is a penalty term that penalizes the likelihood score according to the number of free parameters, *n*. The fit of a free parameter is only penalized to the extent it actually contributes to explaining the data, meaning a parameter that has no effect on the observable data is irrelevant (Daw 2009). In this thesis, we will use the BIC as our metric to assess candidate models of learning and choice behavior, since it allows models with different numbers of fitted parameters to be compared to one another.

2.7. Overview of Experiments

In this thesis, we test the superiority of the four proposed learning models by running a preliminary model comparison on data from a previously run study (Experiment 1, Eppinger et al., 2013) as well as on data collected from a behavioral learning task on Amazon Mechanical Turk (Experiment 2). In both experiments, participants engaged a learning task in which they chose between pairs of stimuli via trial-and-error in the attempt to learn the "correct" image that yielded more points or lost fewer points, on average. Their choice and reward data was then analyzed using our four computational learning models: the Rescorla-Wagner model, the Asymmetric Prediction Error model, the Asymmetric Outcome model, and the "control" Split-Half model. Each model differed only in how it dealt with feedback, via step-wise error

updating. This analysis allows us to draw conclusions regarding how humans learn from positive and negative feedback and how they use that feedback to make decisions.

In the following chapters, we present the results from our model-fitting analysis of the Eppinger (2013) study as well as the behavioral and model results from our Amazon Mechanical Turk experiment. The Eppinger (2013) study was originally run to examine age differences in learning from appetitive and aversive outcomes. The Mechanical Turk Learning Game was conducted in order to determine if asymmetric updating was occurring based on the valence of the prediction error or the valence of the outcome. The results and analysis from the Eppinger study are discussed in Chapter 3, and those from the Mechanical Turk Learning Game are discussed in Chapter 4. Each experiment chapter contains a description of the task as well as the behavioral analysis and model comparison results.

Chapter 3 Experiment 1: The Eppinger (2013) Study

In order to gain empirical support for our hypothesis that an asymmetric learning model more accurately describes and predicts human choice data, we conducted a pilot model comparison on behavioral data originally collected from Eppinger et al. (2013). Eppinger et al.'s study examined learning differences in older adults and younger adults. We tested our hypothesis on this set of data, in particular, because it consisted of a learning task that had the goal of maximizing monetary reward by: 1) learning to choose actions that lead to monetary reward, and 2) avoiding actions that lead to monetary losses. For purposes of our analysis, we will fit our learning models exclusively on the younger participants' data. In this chapter, I describe the general experimental design used by Eppinger et al. and the computational models we used to analyze the pre-existing results.

Despite the fact that this thesis also seeks to examine whether an asymmetrical prediction error model or an asymmetric outcome model fits behavioral learning data better than the Rescorla-Wagner model, our model comparison for the Eppinger data will only consist of comparing the Rescorla-Wagner model and the Asymmetric Prediction Error learning model (and the split-half model, as a reference). As will be discussed in chapter 3.1.3, the Asymmetric Outcome model will be excluded from the model comparison in this results chapter due to the sign of the prediction error and the sign of the reward outcome always being identical. Our decision to model the prediction error rule, instead of the outcome rule, is somewhat arbitrary, since the results are identical in both cases. Based on the reinforcement learning literature's focus on prediction errors as a signal of learning, we chose to model the Asymmetric Prediction Error rule in the Eppinger model analysis. All four learning rules will be analyzed in the Mechanical Turk Learning Game.

3.1. Eppinger (2013) Bandit Task

Eppinger et al. (2013) examined whether older adults differ from younger adults in how they learn from rewarding and aversive outcomes. All participants were asked to learn to choose actions that either lead to monetary reward or to avoid actions that lead to monetary losses.

3.1.1. Participants

Our analysis of the Eppinger (2013) study included an extended sample of the young adult participant pool reported in the official article. In total, we examined pre-collected data from 27 young adult participants. Subjects were from the Princeton University community and were all right-handed. Participants gave written informed consent to participate in the study, which was approved by the Institutional Review Board of Princeton University (Eppinger et al., 2013).

3.1.2. Materials

Stimuli consisted of 20 computerized, colored images of objects. The feedback stimuli indicated either a loss of 50 cents (negative feedback), a gain of 50 cents (positive feedback), or a neutral outcome of 0 cents. Additionally, participants received feedback of "too slow" if the response deadline was missed. Participants indicated a choice between pairs of stimuli by pressing either a left or right button. In the original study, data was collected during a 3 tesla fMRI scan at Princeton University. The stimuli were projected on a screen mounted at the rear of the scanner bore, which participants viewed through a series of mirrors. EPrime software was used to present the stimuli (Eppinger et al., 2013).

3.1.3 Procedures

The bandit task in the Eppinger (2013) study required participants to make a two-choice decision (either the left or right response button) upon presentation of a stimulus. Subjects learned these stimulus-response pairings via trial-and-error based on deterministic feedback and were instructed to maximize their wins and minimize their losses (see Figure 3.1).

The experiment design consisted of five learning blocks with 48 trials per block (24 per learning condition, randomly intermixed). Trials were divided into either the positive learning condition or the negative learning condition. The positive learning condition consisted of a choice between two stimuli which delivered either nothing or a monetary reward, whereas the negative learning condition consisted of choosing between two stimuli which either delivered nothing or a monetary loss. Each block consisted of four new stimuli, two per positive learning condition and two per negative learning condition, that were presented 12 times in random order. Both the stimulus period and outcome period had a mean interstimulus interval (ISI) of 5 seconds.

Stimulus A and B of each block were assigned to the positive learning condition. When presented with stimulus A, participants won 50 cents when they pressed the left button and received the neutral outcome, 0 cents, when they pressed the right button; this ordering was reversed for stimulus B. Stimuli C and D of each block were assigned to the negative learning condition. Participants received a neutral outcome when they responded with a right button press to C, and lost 50 cents if they responded with the left button (vice versa for D). These stimulus-response assignments were counterbalanced across learning blocks and subjects. Feedback was deterministic, meaning the probability of receiving an assigned outcome given a stimulus-

response was 100%. In the context of this experiment, there was a "right" and "wrong" button assigned to each stimulus, and these assignments never changed for a given stimulus.

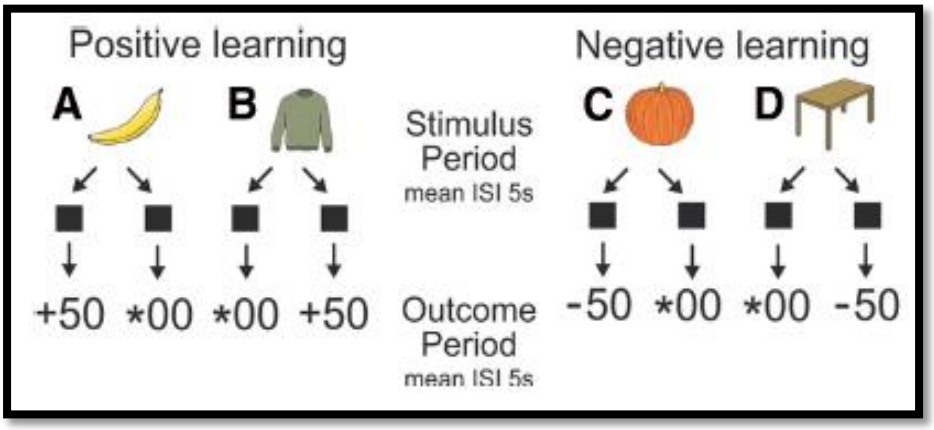

Figure 3.1. Schematic of the task from Eppinger et al., (2013). The experimental design consisted of 5 learning blocks with 48 trials each (24 per learning condition, randomly intermixed). Each block involved a new set of four stimuli (two per learning condition). Feedback was deterministic. Figure from Eppinger et al., (2013).

Because of the deterministic nature of the feedback as well as the compartmentalization of valence (i.e. in the positive condition you only gain reward, whereas in the negative condition you only lose reward), the prediction error will drive the estimated value of the correct choice in one direction--up in the positive learning condition, and down in the negative condition. The reward structure of the Eppinger (2013) study ensures a perfect correlation between the sign of the prediction error and the sign of the outcome, which means the results from modeling our Asymmetric Prediction Error Rule and our Asymmetric Outcome rule will be exactly the same. The Asymmetric Outcome rule will be excluded from our model comparison of the Eppinger study.

3.2. Analysis and Results

3.2.1. Behavioral Analysis

All analyses were carried out using MATLAB (MathWorks) software. Unless otherwise stated, the significant threshold was set to alpha = 0.05. For our behavioral analysis, we analyzed reward and choice data from Eppinger's (2013) study. Individual participant task performance was evaluated by calculating the percentage of correct choices made on each trial across all five blocks. A "correct choice" was defined in the positive learning condition as choosing the button press that resulted in gaining a positive reward and in the negative learning condition as choosing the button that did not lose any points (-0). This data was then averaged across participants to generate a separate accuracy learning curve for the positive and negative learning conditions (see Figure 2). Chance choice behavior was defined as 50% accuracy (1 in 2 chance of choosing the correct button press). Average performance on the first five trials across blocks was found to not be significantly different from chance ($t(26) = 0.30$, $p = 0.76$), whereas the average performance of the last five trials across blocks was found to be significantly different from chance ($t(26) = 24.46$, $p < 0.001$). A within-subject t-test showed that participants' performance for the first five trials were significantly different than their performance for the last five trials ($t(26) = -20.08$, $p < 0.001$). These statistical tests show that participants sufficiently learned the correct stimulus-response assignments throughout the duration of the experiment. Further, accuracy among the first 10 trials of the positive condition were significantly different within-subjects than the first 10 trials of the negative condition ($t(26) = 2.68$, $p = 0.012$), demonstrating a difference in performance moderated by valence.

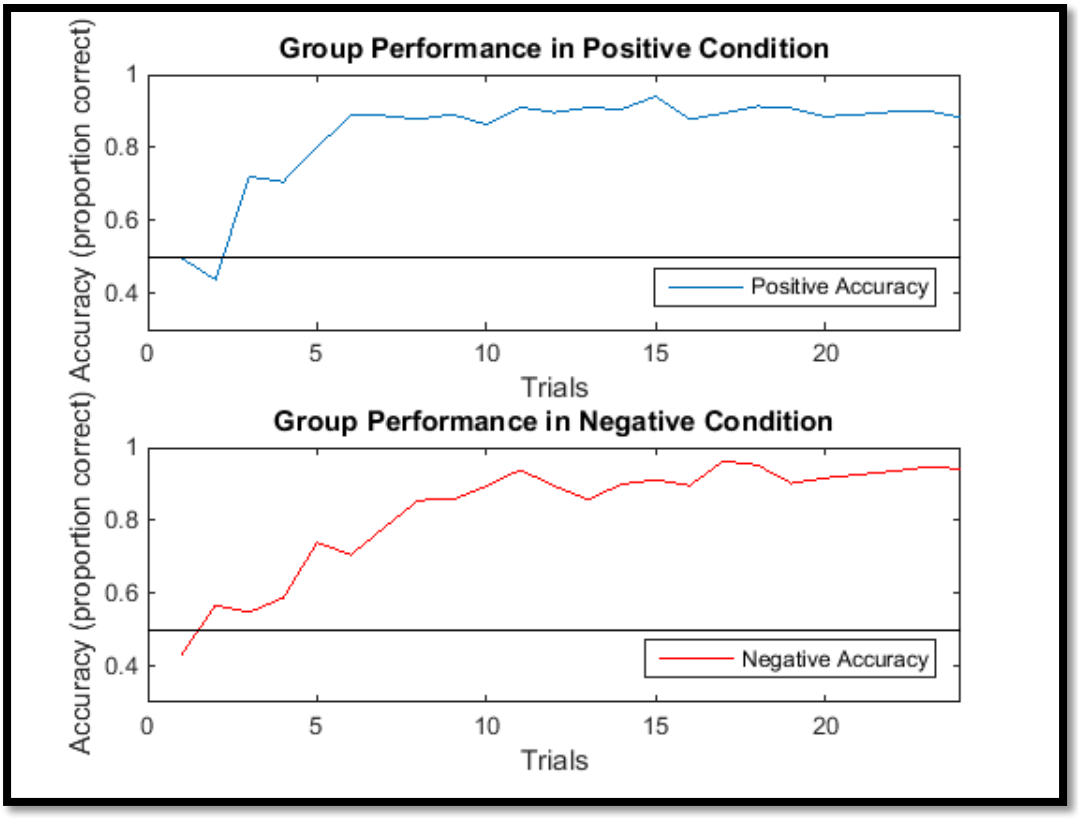

Figure 3.2. Accuracy in proportion of correct choices selected (y-axis). Correct choice data averaged across blocks and participants, separated by learning condition. Accuracy in the positive learning condition is plotted in blue, and the accuracy in the negative learning condition is plotted in red. The black reference line indicates chance level (50%).

3.2.2. Modeling Results

Behavioral data from the Eppinger (2013) study were analyzed using the Rescorla-Wagner model, the Asymmetric Prediction Error model, and the Split-Half Model (described in depth in chapter 2). Value estimates for all stimuli were initialized at zero. Table 3.1 summarizes the average best-fit parameter values, the average log likelihood, and average likelihood per trial for all three Eppinger (2013) models. The average likelihood per trial can be interpreted as the average likelihood of each choice given a particular model. Figure 3.3 plots the average likelihood per trial of each model. All models had a performance that was significantly different from chance, defined as 50% accuracy (Rescorla-Wagner: $t(26) = 15.12$, $p < 0.001$; Asymmetric

Prediction Error: $t(26) = 18.69$, $p < 0.001$; Split-Half: $t(26) = 15.99$, $p < 0.001$). The Asymmetric Prediction Error model provided the best fit to our behavioral choice data with an average likelihood per trial of .733. The Rescorla-Wagner model performed the worst, with an average likelihood per trial of .719, and the Split-Half model's average likelihood per trial was in the middle with .725. A paired t-test demonstrated that the Asymmetric Prediction Error model performed significantly better than both the Rescorla-Wagner model ($t(26) = 4.00$, $p < 0.001$) and the Split-Half model ($t(26) = 2.12$, $p = 0.04$). According to our BIC analysis (see Table 2), the Asymmetric Prediction Error model was the model that best fit the Eppinger dataset. The preferred model is the model with the lowest BIC approximation.

Model	**Average Parameter Fit(s)**	**Average Log Likelihood**	**Average Likelihood per Trial**
Rescorla-Wagner	$\alpha = .22$	76.65	0.719
Asymmetric Prediction Error	$\alpha^{+} = .15$ $\alpha^{-} = .28$	72.10	0.733
Split-Half	$\alpha^{even} = .21$ $\alpha^{odd} = .25$	74.78	0.725

Table 3.1. Summary statistics of best-fit estimates from the Rescorla-Wagner, Asymmetric Prediction Error, and Split-Half models on the Eppinger (2013) data.

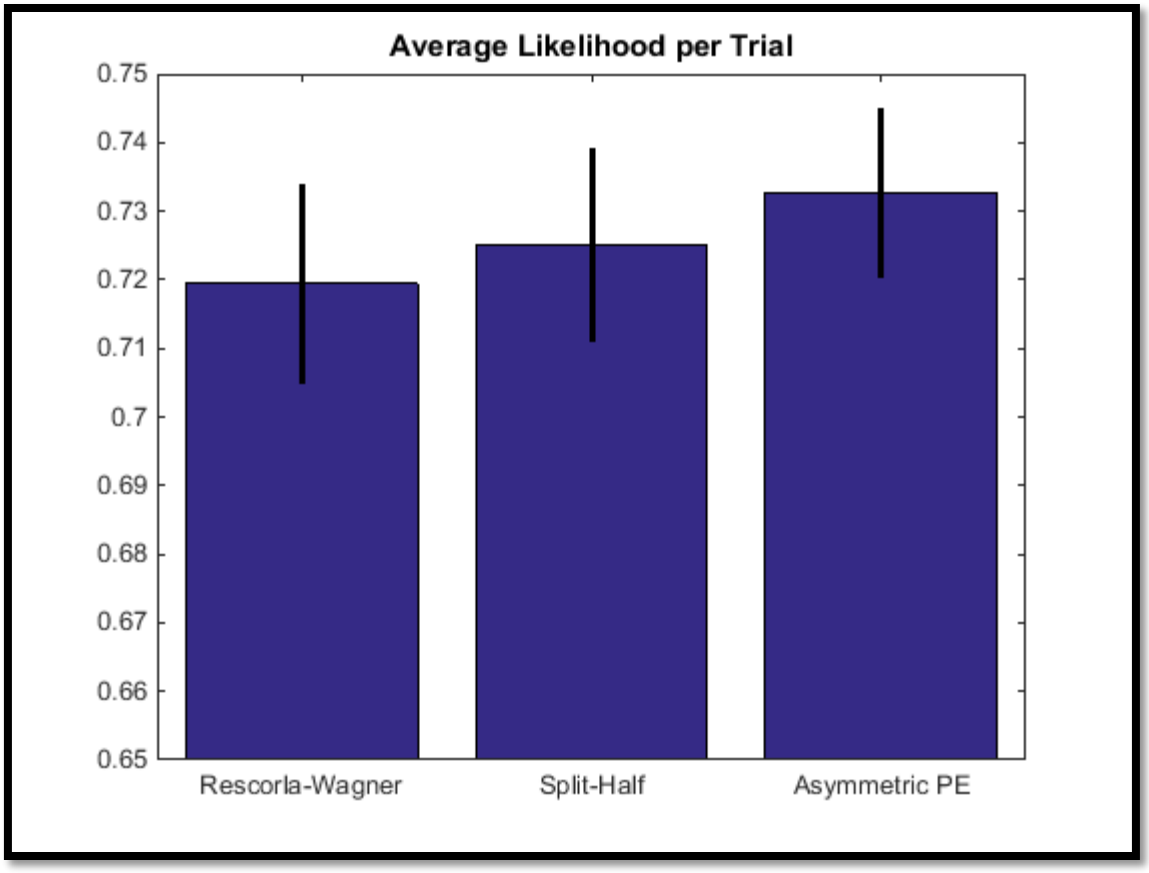

Figure 3.3. Average Likelihood per Trial for the Rescorla-Wagner model, the Asymmetric Prediction Error model, and the Split-Half model. All models performed significantly better than chance (50%). The Asymmetric Prediction Error model performed significantly better than both the Split Half model and Rescorla-Wagner model.

Model	BIC Approximation
Rescorla-Wagner	155.3
Asymmetric Prediction Error	148.2
Split-Half	153.56

Table 3.2. BIC approximation of each model in the Eppinger analysis. The model with the lowest BIC is the preferred model. The Asymmetric Prediction Error is the preferred model, followed by the Split-Half model, and the Rescorla-Wagner model.

An important proposition to our model analysis is that the Asymmetric Prediction Error model is superior to the Rescorla-Wagner model because it allows individual valence biases to be accounted for in the model. That is, the variation in the asymmetric model is systematic rather than random. Figure 3.4 validates this hypothesis by plotting the difference between the α^+ and α^- estimates from the Asymmetric Prediction Error model and the Split-Half model. If a single system is responsible for updating value estimates regardless of whether the prediction error was positive or negative, we would expect the average difference between the positive and negative alpha fits from our asymmetric models to be zero. If the differences are found to be non-zero in the asymmetric model, then there is a significant difference between the two alpha parameters, demonstrating that learning rates update value estimates differently and independently, depending on valence of the feedback.

Taking the difference between the two fitted alpha parameters in both the Asymmetric Prediction Error model and the Split Half model shows that the mean difference in the Split Half model is -0.049, whereas the mean difference in the Prediction Error parameters are -0.13. A t-test showed that the difference in fitted alpha parameters between the two models were significantly different from each other ($t(26) = -2.67$, $p = .013$). Therefore, not only have we concluded that the Asymmetric Prediction Error model significantly outperforms both the Rescorla-Wagner model and the Split-Half model in explaining the Eppinger data, but also we have found that the asymmetric model is superior specifically because it captures systematic variation that is not accounted for in either the Split-Half model or the Rescorla-Wagner model.

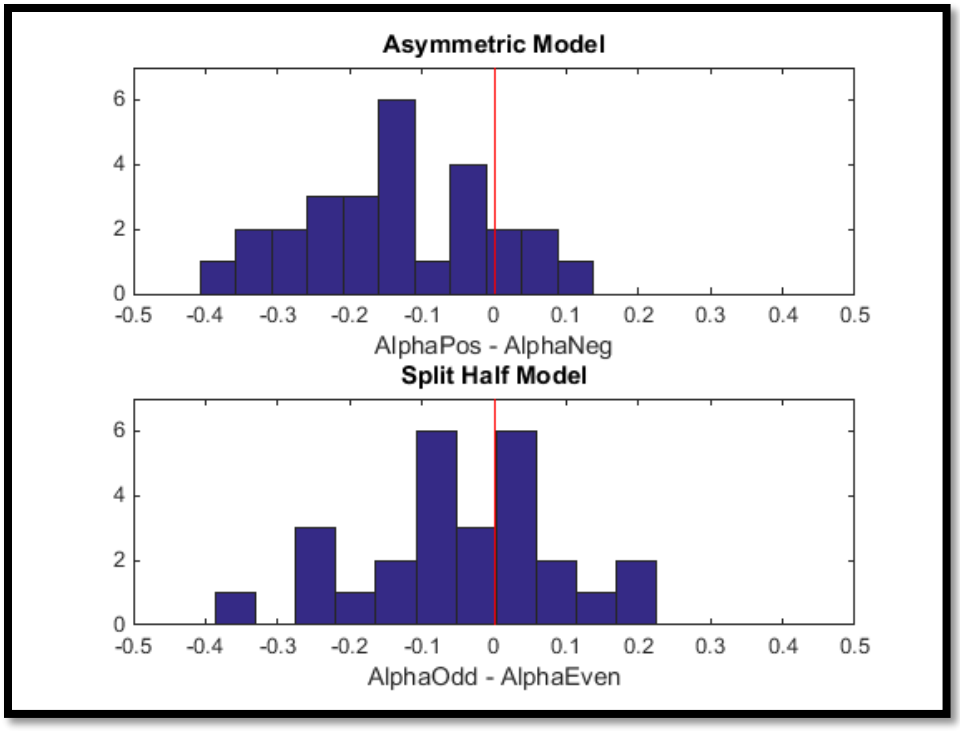

Figure 3.4. Histogram of the differences between α^+ and α^- fits from the Asymmetric Prediction Error model (top) and the Split-Half model (bottom). The red vertical line marks '0', meaning no difference between the two parameter fits. The Split-Half model is more centered at zero (random variation) than the Asymmetric model (systematic variation).

3.3. Discussion

Experiment 1 focused on analyzing behavioral and model data from the Eppinger et al., (2013) study, in which subjects participated in a learning game with the goal of learning to choose the stimuli that earned the most reward in the positive learning condition as well as learning to choose the stimuli that lost the least amount of reward in the negative learning condition. Examining participants' behavioral choice data indicated they were successful in learning which stimuli were the "correct" choices in the pursuit of earning the most rewards.

We conducted a formal model comparison on behavioral data from the Eppinger (2013) study as a way to test qualitative assumptions of learning and decision making as quantitative

hypotheses. Specifically, we challenged an assumption implicitly built into the classic Rescorla-Wagner model: that learning agents do not discriminate or bias their value estimates based specifically on the valence of feedback. In this chapter, we proposed an alternative learning model, the Asymmetric Prediction Error model, which posits that a "positive learning rate" would be used to update value estimates following positive prediction errors and a "negative learning rate" to update those following negative prediction errors. Our model analysis showed that the Asymmetric Prediction Error model fit subjects' behavior significantly better than both the Rescorla-Wagner model and the Split-Half control model. This confirmation that subjects discriminately updated value estimates based on valence is consistent with other experimental findings using similar modeling paradigms (Cazè & van der Meer, 2013; Gershman 2015). Following the Asymmetric Prediction Error model, the Split-Half rule fit the behavioral data better than the Rescorla-Wagner rule. We also showed that the difference in learning rate parameter fits in the Asymmetric model was skewed away from zero compared to that of the Split-half rule. We concluded that the Asymmetric model was superior to the Split-Half model specifically because it accounted for systematic, rather than random, variation in value updating.

In summary, modeling data from Eppinger (2013) provided support for our hypothesis that biological learning agents discriminate their value-estimate updating based on the valence of feedback they receive. However, we were not able to deduce whether learning agents are specifically discriminating based on the valence of the prediction error (i.e. better or worse than was expected) or the valence of the reward itself (i.e. gain or loss). In the next chapter, we analyze behavioral choice data from a bandit task with a different reward structure to directly investigate whether learning agents are biasing their value updating specifically based on valence the prediction error or the valence of the outcome.

Chapter 4 Experiment 2: Amazon Mechanical Turk Bandit Task

Analysis of data from the Eppinger (2013) study in Chapter 3 provided quantitative support for our hypothesis that human learning and decision making processes are more accurately described by an asymmetric learning rule which discriminates value updating based on the valence of feedback than by the Rescorla-Wagner rule. Yet, a crucial question is whether participants are updating their value estimates based on the valence of prediction errors or on outcome valence. In order to answer this question, we ran a probabilistic learning experiment on Amazon Mechanical Turk that investigated whether learning agents differentiate feedback based on the valence of the prediction error or the valence of the outcome.

Amazon Mechanical Turk (AMT) is an online crowdsourcing service where anonymous participants are financially compensated for completing web-based tasks. Recently, AMT has attracted attention from behavioral scientists interested in gathering human subject data more efficiently and automatically (Crump et al., 2013). Specifically, AMT provides a way for data to be collected from a large sample of people in a short amount of time. See Crump et al. (2013) for evidence that psychological experiments, including cognitive behavioral experiments, can be run effectively on AMT.

4.1. Mechanical Turk Learning Game

4.1.1. Participants

The effective sample consisted of 52 participants recruited through the AMT web service (ages 18-63 years, mean age = 34.1 years). This study took approximately 10-15 minutes in length, for which subjects were paid a flat rate of $0.50. Any AMT participant 18 years of age or older in the United States whose completed tasks (Human Intelligence Tasks, or HITs) had an

approval rate of 95% or higher was eligible to participate in this experiment. As has been suggested for behavioral experiments conducted on AMT, we defined an a priori exclusion criteria to ensure data quality (Crump et al., 2013). Participants were excluded (but still compensated) if they did not perform at chance accuracy (defined as 50% correct choices) on one or both blocks of the study. Using this exclusion criteria, 52 participants were excluded from the 104 total recruited participants, leaving 52 subjects in the effective sample. Informed consent was obtained from each participant before he or she accepted the HIT. This study was approved by the Princeton University Institutional Review Board.

4.1.2. Materials

Stimuli consisted of eight abstract symbols adapted from Omniglot, an online website that contains symbols from various writing systems and languages, see Figure 4.1. Eight symbols from Omniglot were vectorized with Adobe Illustrator. Feedback stimuli for positive learning trials were either "+1" or "+3", signaling to the participant that they gained either one or three points. For negative learning trials, feedback stimuli were either "-1" or "-3", indicating the participant lost one point, or three points, respectively. Participants received feedback of "too slow" if the response deadline was missed. Participants chose either the left stimulus by pressing the left arrow key on their computer keyboard or the right stimulus by pressing the right arrow key.

4.1.3. Procedures

The AMT Learning Game is a computerized bandit task in which participants are instructed to choose between abstract symbols with the goal of earning as many points as possible. The bandit task consisted of two blocks, with 40 trials in each block for a total of 80

trials in the entire experiment. On each trial, participants were shown two stimuli and asked to choose between them. In total, there were four pairs of stimuli. Stimuli pairs were fixed, so that the two stimuli belonging in a pair always occurred together. The order in which the two stimuli in a pair were featured on either the left or the right were counterbalanced.

The bandit task contained both positive and negative trials; on the positive trials, participants chose between images that yielded a point gain of either +1 or +3, whereas choices on the negative trials resulted in a point loss of either -1 or -3. Each block featured a pair of stimuli belonging to the positive learning condition and another pair belonging to the negative learning condition. Within each block, participants saw twenty repetitions each of one positive pair and one negative pair.

For purposes of our analysis, trials were labeled "correct" when the participant chose the stimulus that had the higher probability of earning them +3 points in the positive learning condition or -1 points in the negative learning condition. Feedback was probabilistic, meaning that participants received a stimulus's assigned reward 75% of the time, and the alternative reward 25% of the time. Specifically, in the positive condition, one stimulus yielded a reward of '+3' 75% of the time and '+1' 25% of the time, whereas the other possible choice yielded '+1' with a 75% probability and '+3' otherwise. In the negative condition, each pair of stimuli contained one choice that had a 75% chance of leading to a -1 point loss, whereas the other choice lead to the reverse outcome pattern (see Figure 4.1). The probabilistic structure with two possible positive and two possible negative outcomes ensured that the valence of the reward and the valence of the prediction error are not necessarily the same on every trial.

Upon acceptance of an HIT, participants were randomly assigned to be in an interleaved or blocked version of the experiment. The interleaved version presented alternating positive and

negative trials throughout both blocks, whereas the blocked version presented all of the positive trials in one block and all of the negative trials in the other block (the order was counterbalanced across participants).

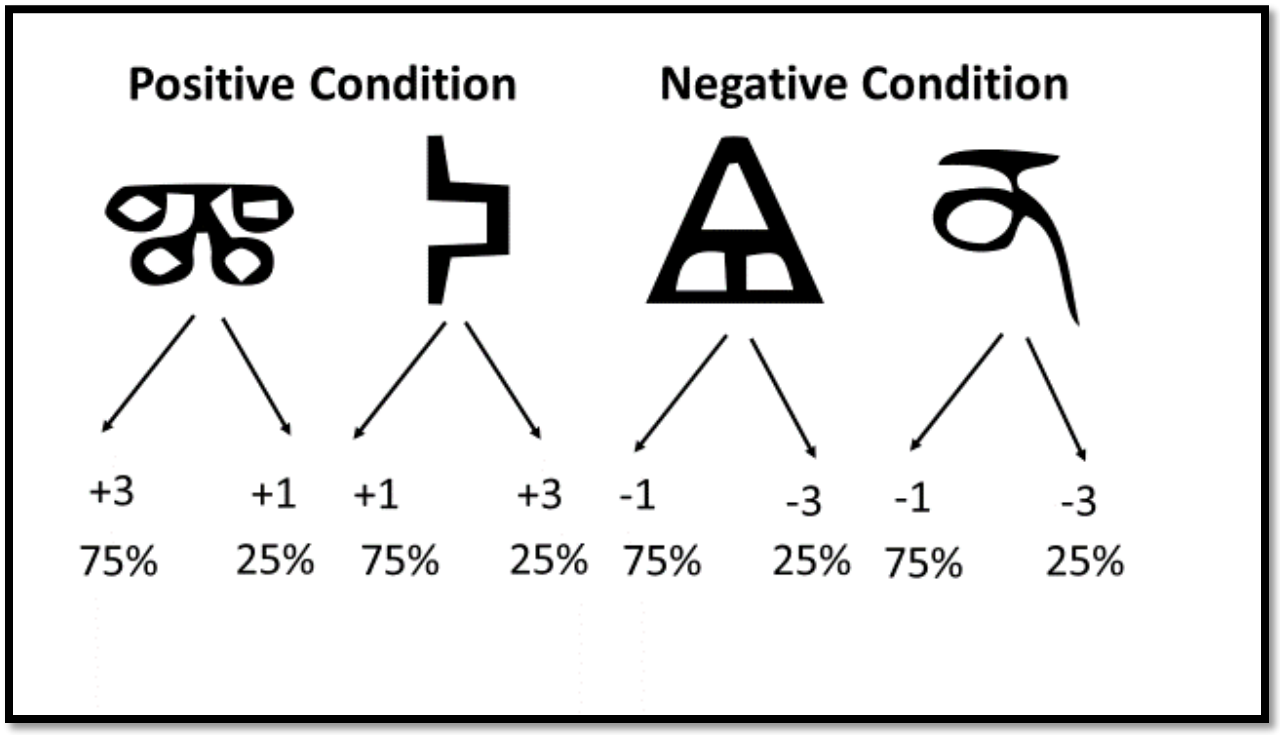

Figure 4.1. Schematic of the bandit task. In each block, four stimuli were assigned to each valence condition. Feedback was probabilistic. Each stimulus presented a particular reward (depending on learning condition) 75% of the time, and presented the other reward 25% of the time.

4.2. Analysis and Results

4.2.1. Behavioral Analysis

Task performance was evaluated by calculating the percentage of correct choices made throughout each block. As previously stated, a "correct" choice was when a participant chose the stimulus that yielded +3 points in the majority of trials in the positive learning condition or the stimulus that resulted in -1 points in the majority of trials in the negative learning condition. Thus, trials in which the correct choice was made but led to the uncommon (worse) reward were

still counted as correct trials. In other words, whether a stimulus was categorized as correct or incorrect was based on its expected value rather than on the reward itself. All statistical tests were carried out in MATLAB (Mathworks) with a significance threshold of alpha = 0.05. Learning curves for the positive and negative learning conditions were created by averaging the accuracy percentage across each learning condition, see Figure 4.2.

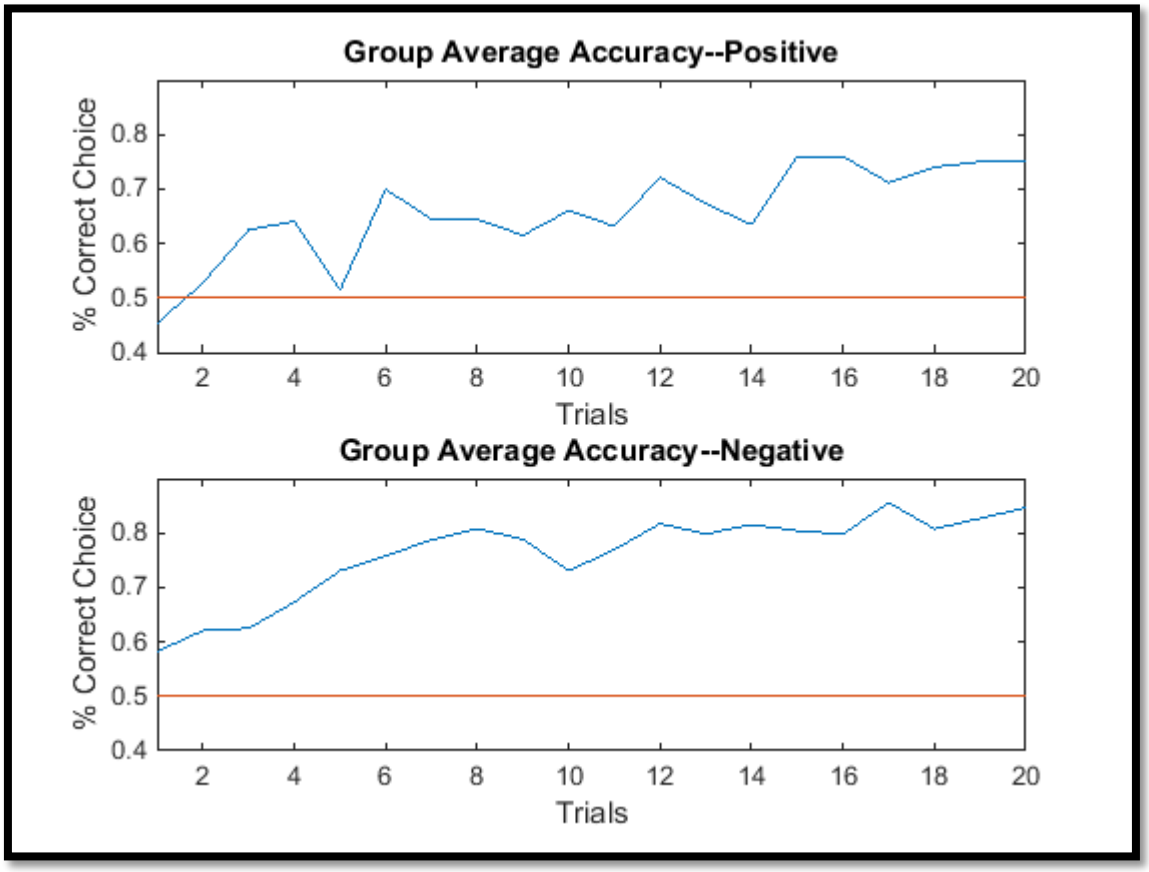

Figure 4.2. Group average accuracy across trials, separately for positive (top) and negative (bottom) learning conditions. Accuracy is plotted in blue and chance performance (50%) is plotted in orange.

Participants were able to learn which stimulus in each trial type pair earned the most reward. A one-sample t-test showed that the first three trials of averaged accuracy across participants and blocks was not significantly different from chance accuracy defined as 50% ($t(51) = 0.66$, $p = 0.51$), whereas the last three trials of averaged accuracy was significantly

different from chance ($t(51) = 12.20$, $p < 0.001$). The average accuracy of the first three trials was also found to be significantly from the average accuracy of the last three trials ($t(51) = -6.66$, $p < 0.001$). Further, an unpaired (two-sample) t-test showed that the averaged accuracy of the last five trials among the blocked participants was not significantly different from the last five trials among the interleaved participants ($t(50) = -0.41$, $p = 0.68$). For the rest of this results chapter, we will pool data from the blocked and interleaved conditions.

4.2.2. Modeling Results

Choice and reward data from the AMT study were analyzed using the Rescorla-Wagner model, the Asymmetric Prediction-Error model, the Asymmetric Outcome model, and the Split-Half model (see chapter 2.1-2.4 for detailed descriptions). For all four models, initial value estimates for the stimuli in the positive condition were set to +2 and stimuli in the negative condition were set to -2. Initial value estimates are, in essence, a set of parameters that must be picked by the user and provides an easy way to supply prior knowledge about what level of rewards can be expected (Sutton & Barto, 1998). Initializing value estimates in the bandit task to zero, in this case, would cause the model to get "stuck" in a few special cases. For example, in the negative learning condition, if a participant chooses a stimulus and receives -1, the model will not be able to recognize that this was the correct choice when compared to the other stimulus's current initialized value of zero. We initialize value estimates of the stimuli in the positive and negative learning condition to +2 and -2, respectively, as a way of encouraging exploration in the model (Sutton & Barto, 1998).

Table 4.1 summarizes the average best-fit parameter values, the average log likelihood, and average likelihood per trial for all four AMT learning models. Figure 4.3 plots the average likelihood per trial for each model. All models performed significantly better than chance, as

defined as each models' likelihood per trial (Rescorla-Wagner: $t(51) = 10.37$, $p < 0.001$; Asymmetric Prediction Error: $t(51) = 11.07$, $p < 0.001$; Asymmetric Outcome: $t(51) = 11.95$, $p < 0.001$; Split-Half: $t(51) = 10.73$, $p < 0.001$). The Asymmetric Outcome model and the Asymmetric Prediction Error model both provided the best fit to our AMT choice data with an average likelihood per trial of 0.59. The Rescorla-Wagner model followed with an average likelihood per trial of 0.57, and the Split-Half model performed the worst with an average likelihood per trial of 0.56.

According to our BIC analysis (see Table 4.2), the Asymmetric Outcome model was the model that best fit our AMT dataset, followed by the Asymmetric Prediction Error model, the Rescorla-Wagner model, and the Split-Half model. Significance tests showed that the BIC of the Asymmetric Prediction Error model was significantly different than the BIC of both the Rescorla-Wagner and the Split-Half models (Rescorla-Wagner: $t(51) = -5.73$, $p < 0.001$; Split-Half: $t(51) = -6.43$, $p < 0.001$) and the Asymmetric Outcome model's BIC was also significantly different from both that of the Rescorla-Wagner model and the Split-Half model (Rescorla-Wagner: $t(51) = -6.78$, $p < 0.001$; Split-Half: $t(51) = -7.29$, $p < 0.001$). The Asymmetric Outcome model's BIC and the Asymmetric Prediction Error model's BIC were not significantly different from each other ($t(51) = -0.18$, $p = 0.86$).

Model	Average Parameter Fit(s)	Average Log Likelihood	Average Likelihood per Trial
Rescorla-Wagner	$\alpha = 0.11$	44.51	0.57
Asymmetric Prediction Error	$\alpha^{+} = 0.24$ $\alpha^{-} = 0.11$	42.72	0.59
Asymmetric Outcome	$\alpha^{+} = 0.17$ $\alpha^{-} = 0.11$	42.65	0.59
Split-Half	$\alpha^{even} = 0.11$ $\alpha^{odd} = 0.05$	45.67	0.56

Table 4.1. Summary statistics of best-fit estimates from the Rescorla-Wagner, Asymmetric Prediction Error, Asymmetric Outcome, and Split-Half models on AMT data.

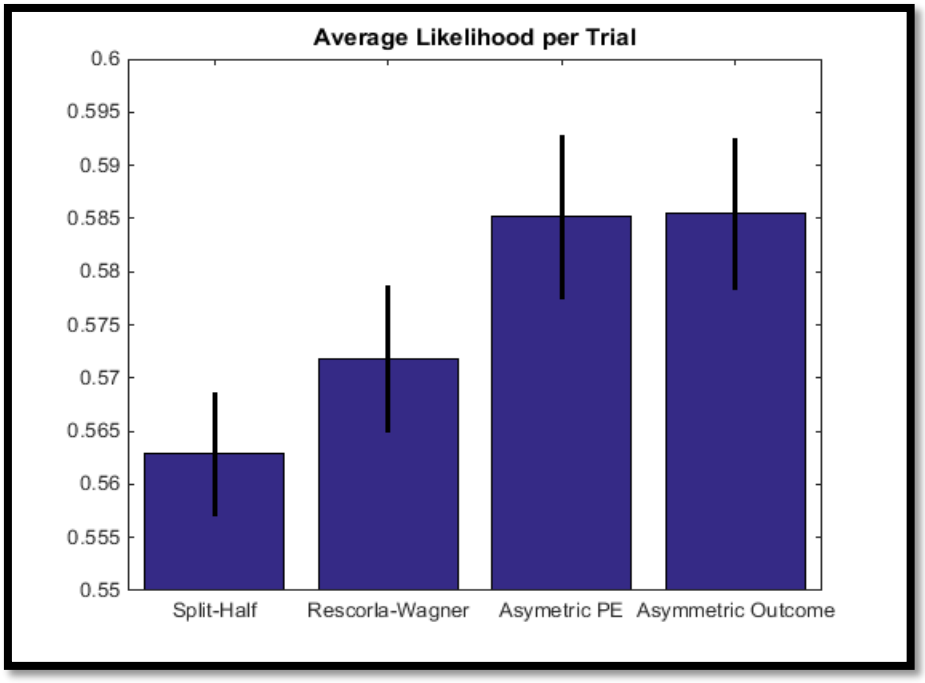

Figure 4.3. Average Likelihood per Trial for the Rescorla-Wagner model, both Asymmetric models, and the Split-Half model. All models performed significantly better than chance (50%).

Model	BIC
Rescorla-Wagner	91.01
Asymmetric Prediction Error	89.44
Asymmetric Outcome	89.29
Split-Half	95.35

Table 4.2. BIC approximation of each model in the AMT analysis. The model with the lowest BIC is the preferred model. The Asymmetric Outcome model is the preferred model, followed by the Asymmetric Prediction Error model, the Rescorla-Wagner model, and the Split-Half model.

4.3. Discussion

In Experiment 2, we used a probabilistic bandit learning task that allowed us to disentangle the valence of the prediction error from the valence of the outcome. Participants engaged in a learning task in which their objective was to maximize their overall points earned. In the positive learning condition, their goal was to maximize their gains; in the negative learning condition, their goal was to minimize their losses. Participants' reward and choice data were fit with four proposed learning models and compared for superiority of fit using similar techniques used in chapter 3 and described in depth in chapter 2. Learning curves indicated that in both the positive and negative conditions, participants were able to learn which symbol in each block was the on-average best choice.

A model comparison as measured via BIC approximation demonstrated that both the Asymmetric Prediction Error model and the Asymmetric Outcome model performed significantly better in fitting the data than either the Rescorla-Wagner or Split-Half rule, yet were not significantly different from each other. This validated our hypothesis that an asymmetric

learning rule more accurately describes human choice data specifically because it includes valence discrimination in the decision maker's value estimates and actions. However, our main focus in Experiment 2 was to separate whether the valence discrimination was occurring based on how the learner compares a received outcome to their prior expectations or whether it was occurring simply based on the valence of the reward itself. The model comparison was not able to differentiate this nuance of decision-making; more research must be done to see whether there is indeed a difference between discrimination based on the valence of the prediction error or the valence of the outcome, as well as determining in which contexts these discriminations manifest.

An interesting difference between the fitted learning rate parameters of the two asymmetric learning models is that the Asymmetric Outcome model returned three participants with a negative learning rate parameter of zero, whereas the Asymmetric Prediction Error model returned 15 participants' negative learning rate parameter as zero. This is compared to no zero fits with the Rescorla-Wagner model, and 22 zero fitted parameters for the Split-Half model (four positive parameters, 18 negative parameters). We infer that this means the model could not fit or predict participants' choice data when an outcome was worse than expected. This might indicate that, in this particular experiment with this particular reward structure, the Asymmetric Outcome model can fit more participants than the Asymmetric Prediction Error model. The Asymmetric Outcome model might also be superior to the Asymmetric Prediction Error model in the AMT bandit task because the outcomes were categorical. Meaning, there were only four outcomes (two negative, two positive), so participants may have performed the task via "corrective search" where they searched for the on-average correct stimulus rather than attempting to estimate every stimuli's expected reward value. Indeed, one of the presumptions of the prediction error model is that a value estimate is being maintained regarding each stimulus,

whereas the Asymmetric Outcome model switches learning rate parameters simply based on the valence of the outcome. An interesting future experiment would be a similar bandit task with rewards stochastically delivered based on a Gaussian distribution. Replacing the AMT task's categorical reward structure with that of a stochastic Gaussian could place more emphasis on estimating a stimulus's expected value as opposed to finding the on-average "correct" stimulus.

Another potential reason why some participants had a model-fitted asymmetric learning rate of zero could be because some decision makers discriminate value updating based on outcome valence while others discriminate based on the sign of a trial's prediction error. Indeed, of the fifteen participants who had a fitted negative learning rate of zero with the Asymmetric Prediction Error model, only one participant also had a fitted negative learning rate of zero with the Asymmetric Outcome model. Our proven hypothesis that human decision makers differentiate learning based on the valence of feedback could be true when aggregated across a large population of decision makers, while individual differences regarding whether the differentiation occurs based on the prediction error or the outcome could still be in play.

Chapter 5 General Discussion

This thesis sought to prove computationally that biological learners and decision makers discriminate learning and value updating based on the valence of feedback received. Asymmetric valence algorithms in reinforcement learning contexts have been proposed in the past (Cazè & van der Meer 2013; Daw et al., 2002; Gershman 2015; Niv et al., 2012); however, this thesis is novel in its attempt to distinguish asymmetric learning models based on whether learners discriminate value-updating specifically based on the valence of the prediction error or the valence of the outcome. The contribution of this thesis is model-fitted, empirical evidence that an asymmetric learning rule outperforms the Rescorla-Wagner rule in two reinforcement learning experiments.

5.1. Asymmetric Learning Models

The results obtained in Experiments 1 and 2 provide evidence for the superiority of reinforcement learning models with separate learning rate parameters for positive and negative feedback, as opposed to reinforcement models symmetrically updating both positive and negative feedback with the same learning rate parameter. We analyzed human participants' reward and choice data from the Eppinger (2013) study and from our AMT bandit task and fit these data with four candidate learning models. We compared the Rescorla-Wagner learning rule with an Asymmetric Outcome rule, an Asymmetric Prediction Error rule, and a Split-Half rule. Both asymmetric learning rules were our test models; the Split-Half model was included as a form of control to see whether our test models performed better because they included additional, important information, rather than them performing better than the Rescorla-Wagner rule simply because of the inclusion of an additional model-fitted parameter. In other words, the

Split-Half rule protected our hypothesis against biasing towards the asymmetric models due to overfitting the data.

We have concluded that an asymmetric learning model with two systematic learning rate parameters for positive and negative feedback fit learning choice data better than a traditional learning rule with only one learning rate parameter used for both positive and negative feedback. In Experiment 1, the asymmetric learning rule performed significantly better than both the Rescorla-Wagner rule and the Split-Half rule. In Experiment 2, both of our candidate asymmetric learning models performed significantly better than both the Rescorla-Wagner learning rule and the Split-Half rule; however, there was no significant difference between the Asymmetric Outcome and Asymmetric Prediction Error rules. This was surprising, since the AMT task's reward structure created a learning environment in which the valence of a participant's calculated prediction error was independent of whether the participant received a loss or gain. Another surprising finding was the parameter fits of 'zero' for some participants' fitted negative learning rate (and some positive learning rates in the case of the Split-Half model). We speculate that this might be due to individual differences in how humans discriminate valence in reward-based learning tasks. In other words, some participants' choice data could be best characterized by the Asymmetric Prediction Error model, while other participants' data could be best described by the Asymmetric Outcome rule. These possible individual differences in how people react to and learn from positive and negative feedback could potentially be related to individual differences in other psychological phenomena regarding gains and losses, such as loss aversion (Kahneman & Tversky, 1979; Tom et al., 2007). More research needs to be done regarding possible individual differences between candidate asymmetric learning models.

Reinforcement learning models are *normative frameworks*, in the sense that they provide quantitative solutions by which optimal value prediction and action selection can be achieved (Niv 2009; Niv & Montague, 2008). While the premise of a learning rule with separate learning rates for probabilistic positive and negative feedback, by definition, biases value estimates and would be expected to lead to suboptimal performance of maximizing reward, other researchers have showed that separate learning rates enables a better separation of learned reward probabilities, and thus can be adaptive (Cazè & van der Meer 2013). In Chapter 3, we show that the asymmetric learning rule outperforms the Split-Half rule specifically because the former model includes individual, systematic reliance of one learning rate parameter over the other which cannot be captured in the latter rule nor in the Rescorla-Wagner model. Thus, we have provided evidence that an asymmetric learning rule which includes an individual learner's systematic valence biases provides a better framework for adaptive reinforcement learning in that it more accurately describes and prescribes learning behavior as opposed to a symmetric learning rate model.

5.2. Directions for Future Research

The AMT task suffered from a high percentage of participant exclusion. A total of 104 AMT subjects participated in Experiment 2, of which 52 subjects were excluded due to at or below chance performance in at least one block. This high rate of exclusion could have been an artifact of running the study on AMT, as opposed to an in-lab study in which participants generally have a higher motivation to perform well on the task. This also could be an explanation for the "zero" model fits for some of the models in Experiment 2. Replicating this study in a laboratory setting could increase the quality of the behavioral data. Further, as was suggested in chapter 4.3, it is possible that instead of maintaining a value estimate for each stimulus,

participants could have viewed stimuli as “correct” or “incorrect”. An interesting adjustment to Experiment 2 would be to replace the categorical, probabilistic reward structure with a structure in which each trial’s reward was probabilistically selected from a Gaussian distribution in which the expected value of a particular stimulus would be equal to the mean of the distribution. A more stochastic reward structure with an increased number of possible reward outcomes could provide another learning environment to determine whether learning agents asymmetrically update their value estimates based on the valence of a trial’s outcome or prediction error.

In the AMT bandit task, we pooled data from the blocked and interleaved versions of the task after demonstrating average accuracy among the last five trials in each bandit version were not significantly different. However, there was an interesting difference in the number of remaining participants in each condition after removing subjects who performed at or below chance. Of the 52 participants analyzed, 18 were in the blocked version and 34 were in the interleaved version. This was a surprising finding, since we had originally hypothesized that, if there was found to be a difference in learning accuracy, blocked participants would most likely perform better than the interleaved participants since they would have better compartmentalization of learning the different types of signed stimuli. Future research should investigate whether presenting stimuli that deliver gains separately with stimuli that deliver losses or presenting both types of stimuli together would impact learning accuracy.

Another promising avenue of research would be to run a model-based fMRI study in which signals derived from our candidate computational learning models would be correlated against fMRI data collected from participants performing the AMT task. This would allow us to determine brain regions showing a response profile consistent with our asymmetric learning models (for a review of model-based fMRI techniques, see O’ Doherty et al., 2007). It would be

particularly interesting, given our AMT modeling results, whether some brain regions' activity are better described using the Asymmetric Outcome learning rule while others are better described by the Asymmetric Prediction Error learning rule.

A more thorough understanding of how humans learn across the gains-losses domain have major implications for society at large. Continued research of asymmetric learning models have the potential to provide insight regarding age-related changes in learning (Eppinger et al., 2013; Samanez-Larkin et al., 2014) as well as learning deficits due to psychiatric disorders impacting midbrain dopaminergic regions, such as Parkinson's disease (Frank et al., 2007; Frank et al., 2004). Finally, further investigation of computational learning models of asymmetric learning has implications for a better understanding of machine learning and artificial intelligence, two fields which are growingly rapidly and have major promise in the future.

References

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

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. http://doi.org/10.1038/nn1954

Berns, G. S., McClure, S. M., Pagnoni, G., & Montague, P. R. (2001). Predictability modulates human brain response to reward. *The Journal of Neuroscience*, *21*(8), 2793–2798.

Cazé, R. D., & van der Meer, M. A. (2013). Adaptive properties of differential learning rates for positive and negative outcomes. *Biological Cybernetics*, *107*(6), 711–719.

Chakravarthy, V. S., Joseph, D., & Bapi, R. S. (2010). What do the basal ganglia do? A modeling perspective. *Biological Cybernetics*, *103*(3), 237–253. http://doi.org/10.1007/s00422-010-0401-y

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410.

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

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*(4), 603–616.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196. http://doi.org/10.1016/j.conb.2008.08.003

Den Ouden, H. E. M., Kok, P., & de Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology*, *3*, 548. http://doi.org/10.3389/fpsyg.2012.00548

Eppinger, B., Schuck, N. W., Nystrom, L. E., & Cohen, J. D. (2013). Reduced striatal responses to reward prediction errors in older compared with younger adults. *The Journal of Neuroscience*, *33*(24), 9905–9912.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*(5614), 1898–1902.

Frank, M. J., Moustafa, A. A., Haughey, H. M., Curran, T., & Hutchison, K. E. (2007). Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning. Proceedings of the National Academy of Sciences, 104(41), 16311-16316.

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

Garrison, J., Erdeniz, B., & Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience and Biobehavioral Reviews*, *37*(7), 1297–1310. http://doi.org/10.1016/j.neubiorev.2013.03.023

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review*, 1–8.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *108*(Suppl 3), 15647–15654. http://doi.org/10.1073/pnas.1014269108

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

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. American psychologist, 39(4), 341.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, 263–291.

Liu, X., Hairston, J., Schrier, M., & Fan, J. (2011). Common and distinct networks underlying reward valence and processing stages: A meta-analysis of functional neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, *35*(5), 1219–1236. http://doi.org/10.1016/j.neubiorev.2010.12.012

Lohrenz, T., & Montague, P. R. (2009). Prediction Errors in Neural Processing: Imaging in Humans. In L. R. Squire (Ed.), *Encyclopedia of Neuroscience* (pp. 885–893). Oxford: Academic Press. Retrieved from http://www.sciencedirect.com/science/article/pii/B9780080450469015564

Maia, T. V. (2009). Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cognitive, Affective & Behavioral Neuroscience*, *9*(4), 343–364. http://doi.org/10.3758/CABN.9.4.343

Matsumoto, M., & Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. Nature, 459(7248), 837-841.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of Neuroscience*, *16*(5), 1936–1947.

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

Niv, Y. (2013). Neuroscience: Dopamine ramps up. *Nature*, *500*(7464), 533–535. http://doi.org/10.1038/500533a

Niv, Y., Duff, M. O., & Dayan, P. (2005). Dopamine, uncertainty and TD learning. *Behavioral and Brain Functions*, *1*(6), 1–9.

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *The Journal of Neuroscience, 32*(2), 551–562.

Niv, Y., & Montague, P. R. (2008). Theoretical and empirical studies of learning. *Neuroeconomics: Decision Making and the Brain*, 329–50.

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences, 12*(7), 265–272. http://doi.org/10.1016/j.tics.2008.03.006

O'Doherty, J., Kringelbach, M. L., Rolls, E. T., Hornak, J., & Andrews, C. (2001). Abstract reward and punishment representations in the human orbitofrontal cortex. *Nature Neuroscience, 4*(1), 95–102.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337.

O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-Based fMRI and Its Application to Reward Learning and Decision Making. *Annals of the New York Academy of Sciences, 1104*(1), 35–53.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical conditioning: current research and theory*. Retrieved from http://www.researchgate.net/publication/233820243_A_theory_of_Pavlovian_conditioning_Variations_in_the_effectiveness_of_reinforcement_and_nonreinforcement

Samanez-Larkin, G. R., Worthy, D. A., Mata, R., McClure, S. M., & Knutson, B. (2014). Adult age differences in frontostriatal representation of prediction error but not reward outcome. Cognitive, Affective, & Behavioral Neuroscience, 14(2), 672-682.

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

Seymour, B., Daw, N., Dayan, P., Singer, T., & Dolan, R. (2007). Differential encoding of losses and gains in the human striatum. *The Journal of Neuroscience*, *27*(18), 4826–4831.

Sutton, R. S. (1998). *Reinforcement learning an introduction / Richard S. Sutton and Andrew G. Barto.* Cambridge, Mass: MIT Press.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of pavlovian reinforcement. Retrieved from http://doi.apa.org/psycinfo/1991-97439-012

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The Neural Basis of Loss Aversion in Decision-Making Under Risk. *Science*, *315*(5811), 515–518. http://doi.org/10.1126/science.1134239

Form for Collaboration in Senior Thesis Work

Please use this form to indicate the relationship between previous work and your senior thesis and to indicate whether your thesis involved collaboration with others.

Indicate below whether there is any overlap between your senior thesis and earlier work that you did for junior reports, junior papers, or papers for various courses.

Overlap _____

No Overlap _____X_____

If you checked the box indicating that there is overlap between your senior thesis and previous work, please describe the overlap on a **separate page**, and include it within the thesis after this form.

Readers of your thesis may, if they choose, ask to see earlier papers that you indicate have some overlap with your senior thesis.

Indicate below whether all or part of your thesis resulted from work done collaboratively with one or more other people.

Collaboration _____X_____

No Collaboration _____

If you checked the box indicating that your thesis work was done entirely, or in part, in collaboration with other people, describe the nature of the collaboration and what resulted from it on a separate page, and include it within the thesis after this form.

Nature of Collaboration

Experiment 1 of this thesis was a modeling analysis of an earlier study run by Ben Eppinger, Nicolas Schuck, Leigh Nystrom, and Jonathan Cohen. The results of the original work were reported in Eppinger et al., 2013. During the summer of 2014, I worked in the Niv Lab with Nicolas Schuck who provided me with the data from the original 2013 study, authorized by Ben Eppinger. Nicolas and I both wrote a significant amount of code which contributed to the results reported in chapter 3 of this thesis.

The modeling results collected in chapter 4 of this thesis were significantly written with the help of code and tutorials I wrote with the help of Yael Niv, Nathaniel Daw, and Reka Daniel-Weiner at the SRNDNA computational modeling workshop following the Decision Neuroscience and Aging 2015 conference. The final code used to analyze the results in chapter 4 as well as a significant amount of the theory and mathematical reasoning behind the methods in chapter 2 were a product of this conference. Following the conference, Nicolas and I further altered and updated the modeling code, collaboratively.

In regards to the AMT study, itself, I collaborated with Nicolas Schuck, Eran Eldar, and Angela Radulescu in coding the experiment and uploading it to Amazon's Mechanical Turk web service. Eran Eldar provided code for the bandit task which Nicolas and I both adapted for purposes of our study. Eran also graciously provided us with computer tools he had coded that allowed us to upload the experiment onto AMT. Also, the Omniglot vectorized images used in the AMT bandit task in chapter 4 were provided by Angela Radulescu.

Approval Form for Undergraduate Research Involving Experimental Animals

All research involving experimental animals at Princeton University must receive the prior approval from the Institutional Animal Care and Use Committee (IACUC). The IACUC bases decision about approval on the NRC Guide for the Care and Use of Laboratory Animals. All students conducting research involving animals as part of their junior independent work or senior thesis must receive approval from the IACUC prior to beginning their research. Students should consult first with their advisers about whether the procedures they intend to use are already covered by previously approved submission to the IACUC. The IACUC meets only once a month and it is common for new submissions to require revision before receiving approval so students are strongly encouraged to attend to IACUC issues early in their planning.

Did your Senior Thesis research involve the use of experimental animals?

Yes _____ No ____X__

If you answered "Yes" to the above, you *must* also include a statement at the beginning of your methods section that verifies the work you have done with animals was approved by the Princeton University IACUC.

Lastly, please include this form at the back of your thesis (even if you answered "No"). **If you answered "Yes," please record the IACUC protocol number and date, below.**

IACUC # _____ Approval Date _____

Approval Form for Undergraduate Research Involving Human Subjects

The Institutional Review Board for Human Subjects (IRB) is charged by the University Research Board with the task of protecting the interests and rights of human subjects involved in Princeton research. The IRB's responsibility includes the oversight of research conducted by undergraduate as part of their junior independent work and senior thesis work as well as that conducted in fulfillment of course requirements. All students conducting research involving human subjects as part of their junior independent work or senior thesis must receive approval from the IRB prior to beginning their research. Obviously, the sooner students submit their requests to IRB the sooner they will receive this approval. Students should be encouraged to submit their materials to the IRB as soon as possible in the semester. The IRB meets only once a month and it is common for student submissions to require revisions, primarily because of the incompleteness of the original submission, before receiving approval.

Did your Senior Thesis involve research with human subjects?

Yes ___X___ No _____

If your Senior Thesis DID involve research with human subjects, please indicate your IRB Case Number below.

Case Number: ____#4452____ Approval Date: _October 27, 2014_