One Step at a Time: Analysis of Neural Responses During Multi-State Tasks

By

Talora Bryn Grey
Bachelor of Science, University of Victoria, 2007

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in Interdisciplinary Studies

We acknowledge with respect the Lekwungen peoples on whose traditional territory the university stands and the Songhees, Esquimalt and W̱SÁNEĆ peoples whose historical relationships with the land continue to this day.

Supervisory Committee

One Step at a Time: Analysis of Neural Responses During Multi-State Tasks

By

Talora Bryn Grey
Bachelor of Science, University of Victoria, 2007

Supervisory Committee

Dr. Olave E. Krigolson, Supervisor
The School of Exercise Science, Physical and Health Education

Dr. Alona Fyshe, Co-Supervisor
Department of Psychology

Abstract

Substantial research has been done on the electroencephalogram (EEG) neural signals generated by feedback within a simple choice task, and there is much evidence for the existence of a reward prediction error signal generated in the anterior cingulate cortex of the brain when the outcome of this type of choice does not match expectations. However, less research has been done to date on the neural responses to intermediate outcomes in a multi-step choice task. Here, I investigated the neural signals generated by a complex, non-deterministic task that involved multiple choices before final win/loss feedback in order to see if the observed signals correspond to predictions made by reinforcement learning theory. In Experiment One, I conducted an EEG experiment to record neural signals while participants performed a computerized task designed to elicit the reward positivity, an event-related brain potential (ERP) component thought to be a biological reward prediction error signal. EEG results revealed a difference in amplitude of the reward positivity ERP component between experimental conditions comparing unexpected to expected feedback, as well as an interaction between valence and expectancy of the feedback. Additionally, results of an ERP analysis of the amplitude of the P300 component also showed an interaction between valence and expectancy. In Experiment Two, I used machine learning to classify epoched EEG data from Experiment One into experimental conditions to determine if individual states within the task could be differentiated based solely on the EEG data. My results showed that individual states could be differentiated with above-chance accuracy. I conclude by discussing how these results fit with the predictions made by reinforcement learning theory about the type of task investigated herein, and implications of those findings on our understanding of learning and decision-making in humans.

Table of Contents

List of Tables

List of Figures

List of Equations

Acknowledgements

I would like to thank the NeuroEducation Network for, in part, making this work possible. I would like to thank my supervisors, Dr. Olav Krigolson and Dr. Alona Fyshe, for their invaluable guidance and support during the entirety of this degree. I very much appreciate the opportunity they gave me to complete this work, despite the challenges faced along the way, as well as the environment they provided to grow as a researcher and as a person. I want to thank Cameron Hassall for his ready assistance with experimental questions, advice on academic matters, and talking through the theory behind Wizards' Duel. I'd also like to thank the members of the Krigolson Lab over the past few years for their camaraderie, questions, and answers. Finally, I want to thank my partner, Adam, for his constant and unwavering belief in me throughout this process.

Chapter 1: A Review of Reinforcement Learning Theory, Electroencephalograms, and Machine Classification

1.1 Introduction

Learning to complete a complex sequence of states in order to achieve a goal is common in our everyday lives. For example, each time we need to navigate through a city to reach a new destination, we choose a route that is composed of many decision points: each intersection presents a variety of options, some of which get us closer to our destination and others that take us further away. Barring the use of a GPS navigation aid, we make a choice at each intersection, and then learn at the completion of the trip whether each choice along the way was a "good" decision: did it help or hinder us in reaching our destination? Could one of the other choices have led to a quicker and easier route? For instance, was there construction on part of the route that caused slowdowns, and consequently the decision to go that way was a poor choice? Working backward from the outcome of the trip, we will, over time, decide which is the best route to that destination.

In learning and decision-making research, these questions are often examined within the framework of reinforcement learning (RL) theory. Reinforcement learning has its origins in Thorndike's (1898) work on learning theory that led to his Law of Effect, which stated that behaviours by an organism which produced a desirable outcome would increase the rate of the behaviour that led to the desirable outcome, and that behaviours which produced an undesirable outcome would decrease the rate of the behaviour that caused the undesirable outcome. In the early and mid-20th century, B. F. Skinner built upon Thorndike's work as well as Pavlov's theories of learning through classical conditioning, and formalized the theory of operant conditioning (Pavlov, 1897; Skinner, 1938, 1953). Operant conditioning, as part of the

behaviourism theory in psychology, extended Thorndike's Law of Effect to include positive and negative reinforcers and punishers as mechanisms of learning. In behaviourism, positive is used in the additive sense, while negative is used in the subtractive sense (Skinner, 1938, 1953). So, learning occurs by the addition or removal of reinforcing or punishing outcomes for an organism's behaviours.

Operant conditioning was formalized in reinforcement learning theory, which states that learning occurs from behaviours that affect the environment and the resulting outcomes from those behaviours (Rescorla & Wagner, 1972). Thus, in reinforcement learning theory, learning is achieved by trial-and-error experimentation rather than by expert instruction—as in supervised learning theory, although the feedback in RL can come from an expert—or working from correct/incorrect examples. RL theory defines six elements that make up a system: an agent, actions, the environment, a policy, a reward signal, and a value function. In the following section, I will detail the purpose of each of these elements, and discuss how they interact to enable learning. Following that, I will discuss evidence to date that the human brain implements reinforcement learning, how electroencephalogram studies have contributed to our understanding of reinforcement learning in humans, and the gaps in that body of research that this current work attempts to fill. This is necessary background material for the electroencephalogram experiment detailed in Chapter 2. The last two subsections of this first chapter will explain the background necessary for the second experiment, discussed in Chapter 3 of this thesis, which concerns machine classification analysis of the electroencephalographic data collected in the first experiment. The method of machine classification used herein is a type of supervised machine learning; specifically, it is a support vector machine that categorizes new data samples based on a

training set of labeled samples. Finally, in Chapter 4, I discuss the implications of this body of work, its limitations, and future directions for investigation of the examined research questions.

1.2 Reinforcement Learning

In RL theory, the agent (human, animal, or computer learner) can take actions within the environment that may have a beneficial, neutral, or aversive effect on the agent. The agent is attempting to learn a task in order to achieve a goal; an example is a rat learning to navigate a maze in order to eat the food placed at the end. Actions are immediate choices that cause the agent's state to change within its environment; in our example, the rat can, at each intersection in the maze, choose a direction in which to proceed. An environment is comprised of one or more states; e.g. each position in the rat maze is a state, with the maze being the environment, and states being positions within the environment specified by a combination of sensory inputs: a rat learning to navigate the maze can look, smell, and hear, and therefore know its location based on these inputs. In order to learn, the agent, upon execution of an action, receives a signal from the environment that can indicate correct or incorrect performance; this signal is commonly thought of as having a numerical value which is an indicator of whether the chosen action leads the agent closer to or further away from its goal. Overall, the goal of the agent is always to maximize its value function, which is the total amount of reward the agent can expect to receive from this state plus all possible subsequent states, and can be thought of as the "long-term desirability of states" (R. S. Sutton & Barto, 1998). Therefore, the agent uses the value function to choose which action to take in each possible state in the environment. In our rat example, the value function could be the number of food rewards it can find in a set amount of time, and the rat’s goal is to find as many as possible. With practice, the rat will learn which positions in the maze are closer to the food rewards, and thus those positions will gain value and be desirable in and of themselves.

It is important to differentiate between the reward signal and the value function: the reward signal informs the agent how beneficial the action it just executed was, and is evaluated in the context of the value function. The chosen action is the one that will lead to the highest overall reward, signified by the amplitude of the reward value received following the action. Thus, the action selection method may mean foregoing an action that will result in a higher short-term reward signal in favour of an action that results in higher long-term cumulative rewards. In an unfamiliar environment, the agent does not know which actions are most beneficial in each state, and must learn by trial and error which states are associated with a higher reward signal, and which actions maximize the value function, which is done by computing the reward prediction error for each action taken. The reward prediction error is the difference between the expected reward for that action and the actual reward value received. With enough trials, the approximate value for each state is learned (if there is no noise in the reward amount, then the true value will be learned for that action), and the reward prediction error decreases to near zero (if the true value is known for that action, the prediction error will be zero), thus stabilizing the expected reward values over time. Importantly, reward prediction errors cause prior states' values to be updated to reflect the expected reward of subsequent states. For example, if moving to state *n* gives a large positive reward prediction error, the value of the previous state *n - 1* will be increased by some amount to reflect that state *n - 1* leads to states with higher rewards, and therefore will increase the likelihood that the agent will choose to move to state *n - 1* and then state *n* in the future. Thus, in a multi-step process, value slowly propagates back from the goal state to prior states over repeated trials (Rescorla & Wagner, 1972; R. S. Sutton & Barto, 1998).

It is important to note that this is not hierarchical reinforcement learning, in that the states are not subtasks, but rather sequential steps to completion of a task. In other words, the steps examined herein are tasks at the same level of abstraction for a given multi-step activity. As an illustration of the difference, contrast these two descriptions of a goal-directed activity: making coffee. The sequential steps include boiling the kettle, getting the French press out of the cupboard, putting the correct amount of coffee grounds into it, filling the press with boiling water, waiting the correct amount of time for the coffee to brew, pushing the plunger down, and pouring the coffee into a mug. These steps are all roughly equal in complexity and abstraction. Contrast that with the following description of part of the same activity: to boil water, reach for the handle of the kettle, grasp it in an effective manner, lift the kettle, turn the body forty-five degrees to the right, reach with the other hand and grasp the kitchen faucet, lift the lever to turn on the water—and so on. In a hierarchical model, the "boil water" step in the former scenario includes many subtasks, as would all following steps in the sequence to make coffee. For this current work, only the top-level steps, such as those listed in the first description, will be examined, and thus we will use non-hierarchical reinforcement learning theory to interrogate the neural responses generated during a multi-step task.

The implementation details of RL depends on the specific algorithm used. There are many ways of implementing reinforcement learning; a multitude of models have been developed over the years to explain various phenomena in ethology and psychology, and solve various problems in artificial intelligence; these models include Rescorla-Wagner, Temporal-Difference, Q-learning, SARSA, Actor-Critic, and others (Barto, 1995; Barto, Sutton, & Anderson, 1983; Rescorla & Wagner, 1972; Rummery & Niranjan, 1994; R. S. Sutton, 1988; R. S. Sutton & Barto, 1998; Watkins, 1989).

1.3 Computational Models of RL

Turning now to the mathematical/computational theory underlying reinforcement learning, the main RL algorithms that have been used in modeling neural responses in animals including humans are Rescorla-Wagner, Temporal Difference, and Actor-Critic. These algorithms are detailed below.

An RL model for Pavlovian (or classical) conditioning was published by Rescorla and Wagner in 1972. Their model, referred to here as the Rescorla-Wagner model, stated, informally, that learning will only occur when expectations about events are violated (Rescorla & Wagner, 1972). The Rescorla-Wagner model was an attempt to explain many phenomena seen in behavioural experiments exploring Pavlovian conditioning. Formally, the learning rule in the Rescorla-Wagner model is as follows:

$$\Delta V = \alpha\beta(\lambda - \Sigma V) \tag{1}$$

Equation 1. Rescorla-Wagner Update Rule. V is the associative strength between a conditional[1] stimulus (CS) and an unconditional stimulus (US), Δ indicates the change in strength for this update, α is a measure of the salience of the conditional stimulus, β is the associability between the CS and the US, λ is the maximum associative strength that is possible between the conditional and unconditional stimuli, and ΣV is the current total amount of associative strength for all the conditional stimuli present.

Equation 1 describes how the associative strength of a conditional stimulus will change on each trial. On each trial, the value V for a given CS will be increased or decreased by the difference between the value of the actual reward (the US) and the sum of the values of all conditional stimuli present in this trial, multiplied by the rate of learning (represented by the term $\alpha\beta$). Thus, the sum of the associative strength for all conditional stimuli present will converge on the value

[1] In this document, I use the terms conditional and unconditional stimuli, as suggested in R. S. Sutton and Barto (1998).

of the unconditional stimulus; each individual CS contributes a larger or smaller amount to this sum depending on how salient each CS is to the animal.

There were two innovative contributions that Rescorla and Wagner's model made: first, it formalized the idea that learning occurs when events are surprising; that is, when predictions about the outcome are incorrect (Rescorla & Wagner, 1972). Second, it stipulated that the total prediction for a trial is the result of summing the predictions of each available stimulus. These two points allowed the Rescorla-Wagner model to explain several puzzling aspect of animal behaviour, including blocking, overshadowing, and over-expectation (Rescorla & Wagner, 1972). Blocking occurs when a previously trained conditional stimulus *A* is then simultaneously paired in time with a second, neutral stimulus *B*, which both precede the unconditional stimulus. In this case, the associative strength between *B* and the unconditional stimulus will be much weaker than if *A* had not been pre-trained. Overshadowing is similar, except that neither conditional stimulus (*A* or *B*) is pre-trained, and they are presented simultaneously; the result is that one of the conditional stimuli will, with learning, have a stronger association with the unconditional stimulus than the other, depending on which of *A* or *B* the learner finds more salient. Lastly, the phenomenon of over-expectation involves separately conditioning two stimuli *A* and *B* with the unconditional stimulus. When both *A* and *B* are presented together, the response to the unconditional stimulus is greater than to either conditional stimulus alone.

However, although the Rescorla-Wagner model neatly and precisely explains the above three aspects of behaviour, it does not explain second-order conditioning, where a neutral stimulus takes on some of the rewarding properties of an unconditional stimulus and becomes a predictor of a reward, nor does it take into account temporal effects on conditioning and stimuli, including when the conditional stimulus is presented relative to the unconditional stimulus (Niv,

2009). Motivated by these shortcomings in the Rescorla-Wagner algorithm, Sutton developed the temporal-difference (TD) method of reinforcement learning (R. S. Sutton, 1988). The TD method attempts to solve the above issues with the Rescorla-Wagner model by accounting for second-order conditioning effects and the effect of relative timing of stimuli on Pavlovian learning. In Rescorla-Wagner, the goal of the learning agent is to maximize the immediate reward, whereas in TD, the goal is to maximize the cumulative long-term value of rewards received (R. S. Sutton, 1988; R. S. Sutton & Barto, 1998). Thus, the learning system must account for the passage of time over which to integrate all rewards received. TD does this by explicitly encoding the passage of time in the algorithm; each time point has a corresponding state and thus an associated value. Additionally, TD treats both conditional and unconditional stimuli the same, unlike in Rescorla-Wagner, instead of just CSs and USs at the trial level. TD estimates the probability of receiving rewards in all future states that are possible from the current state onward in time, which results in the value for a state reflecting the estimate of reward across all future states. The TD algorithm handles these concepts by modifying the difference (or "surprise") term in the Rescorla-Wagner equation. Whereas Rescorla-Wagner calculates the difference term as in Eq. 1, TD calculates it as follows:

$$\delta_t = r_t + \gamma V(S_{t+1}) - V(S_t) \quad (2)$$

Equation 2. The Temporal-Difference Update Rule, where t is the temporal-difference error, r_t is the reward at time t in state S_t, γ is the discount rate parameter, and S_{t+1} is the observed state at time $t+1$ (R. S. Sutton & Barto, 1998).

Thus, the discount rate reduces the value of each state the further one gets from the reward state.

Another RL algorithm, Actor-Critic, is a variation of TD learning that postulates two separate mechanisms within the agent: an Actor and a Critic (Barto et al., 1983). The Actor

keeps track of the association between states (S_i; see Figure 1) and actions given the action probabilities, $\pi(a|S)$, and chooses the actions taken, while the Critic receives the reward and calculates an update to the weights/association strengths of states S_i and values V, which it uses as input to the temporal difference equation to calculate a prediction error signal that is then used by both the Critic to update the State-Value mappings and the Actor to update the State-Action probabilities.

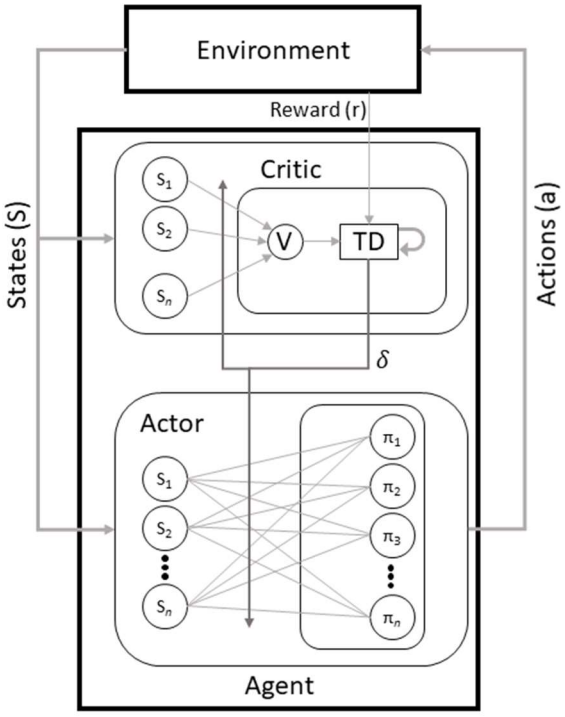

Figure 1. Actor-Critic Architecture. Adapted from Takahashi et al., 2008.

Substantial evidence has been collected that animal and human brains use RL for learning and decision-making (Dayan & Niv, 2008; Holroyd & Coles, 2002; Niv, 2009). Indeed, there is ample evidence for the implementation of Actor-Critic RL in the brain. In 1995, Barto hypothesized that the implementation of actor-critic architecture in the brain uses signals generated by dopamine neurons as both a prediction error as well and reinforcement. (Barto, 1995). Concurrently, Houk, Adams and Barto detailed the neural solution to the temporal credit assignment problem (James C. Houk, Adams, & Barto, 1995). Specifically, they postulated that dopaminergic (DA) and spiny neurons have reciprocal connections that allow the spiny neurons to be trained to recognize context that shortly lead to reinforcement, and subsequently the spiny neurons then control their own dopaminergic inputs. The DA signals are also sent to other spiny neurons such that the resulting system appears to implement an actor-critic architecture.

However, Joel, Niv, and Ruppin (2002) concluded that the models built on Barto's 1995 work were unlikely to work on a biological implementation level and suggested that the structure of an Actor-Critic system in the brain is likely much more complex than the anatomical model in these earlier works. More recently, Takahashi, Schoenbaum and Niv (2008) modelled the effects of cocaine sensitization on rat's brains using an Actor-Critic model. According to Takahashi et al.'s work, it appeared that, in rats, the basal ganglia (especially the striatum), limbic subcortical structures including the amygdala, the hippocampus, and prefrontal cortical areas are all involved in implementing an actor-critic model of RL, which is more extensive list of involved brain structures than previous models such Barto's or Houk et al.'s (Barto, 1995; James C. Houk et al., 1995).

1.4 Electroencephalograms, Event-Related Brain Potentials, and Learning

1.4.1 Electroencephalograms.

The human electroencephalogram (EEG) is composed of electrical signals generated by groups of neurons firing in synchrony, thereby creating voltage fluctuations that can be measured using electrodes either intracranially, or, more commonly, on the surface of the scalp (Luck, 2014). More specifically, the voltages observed in EEG result from the electrical difference between two ends of a neuron created when either an excitatory postsynaptic potential (EPSP) or an inhibitory postsynaptic potential (IPSP) is received by a neuron. EPSPs and IPSPs cause the ends of the neuron to form an electrical dipole, where one end of the neuron has a negative charge and the other end has a positive charge (Jackson & Bolger, 2014; Luck, 2014). The voltage fluctuations observed in EEG are thus the result of changes in these dipoles synchronously across many neurons.

Scalp-based EEG devices have been in use in humans for nearly one hundred years and are useful tools when non-invasive measurements of millisecond-level time resolution of observed signals is required (Berger, 1929; Luck, 2014). Distinctive waveform patterns have been observed in response to various stimuli in humans and animals; time-windowed epochs of evoked waveforms are called event-related brain potentials, or ERPs (Jackson & Bolger, 2014; Luck, 2014). A non-exhaustive list of ERPs observed and quantified to date include the N1, P1, N200, P200, ERN, reward positivity/FRN, and the P300 (Gehring, Liu, Orr, & Carp, 2012; Luck, 2014). ERPs are generally analyzed by averaging together many epochs of EEG data, each generated by the same stimuli, in order to reduce noise inherent in the signal (Luck, 2014). Noise sources include muscle movements, eye blinks and saccades, heartbeat, and external sources of electrical activity such as line noise and interference from other electromagnetic sources near the EEG recording equipment. For the EEG experiment detailed in this thesis, both the reward positivity and the P300 ERP components were analyzed and reported upon and will therefore be discussed in the following subsections.

1.4.2 Reward Positivity.

The reward positivity ERP component was identified during investigation into neural responses to errors. The first research on error-related components was published by two independent groups of researchers: Gehring, Coles, Meyer, and Donchin (1990); and Falkenstein, Hohnsbein, Hermann, and Blanke (1991). During speeded choice reaction time tasks, both teams observed an ERP component that appeared as an abrupt, negative deflection in the EEG waveform that was largest at the front and midline electrodes, and peaked approximately 100 ms after the onset of muscle movement (to be clear, the component both Gehring et al. and Falkenstein et al. observed was response-locked; i.e. it occurred after a

participant had responded in error, not after they received any feedback that their response was incorrect), and only occurred on error trials; thus they named it the Error-Related Negativity (ERN). Further investigation led Gehring and others to surmise that the ERN was a manifestation of activity in some system that was involved in monitoring accuracy of responses, as the negative deflection was not seen on correct trials (Gehring, Goss, Coles, Meyer, & Donchin, 1993). Gehring et al. also stated that the amplitude of the ERN was greater in cases where accuracy had been emphasized over speed in instructions to participants. Interestingly, the limb or response method used in committing the error does not seem to affect the ERN (Holroyd, Dien, & Coles, 1998). These early studies chiefly used experimental paradigms where the cognitive task was straightforward, and errors were due to performance mistakes rather than an incomplete understanding of the task; in combination with the timing of the component (onset approximately 50 ms post-response and peaking about 100 ms post-response), the ERN appeared to be the result of an internal error process evaluating the efference copy of motor commands as opposed to external feedback (Dehaene, Posner, & Tucker, 1994; Holroyd & Coles, 2002).

In their 1997 paper, Miltner, Braun and Coles (1997) reported on a time-estimation experiment where participants received correct/incorrect feedback on their actions in one of three sensory modalities: visual, auditory, or somatosensory. The experiment required participants to press a button one second after a cue. Participants then received feedback as to whether they had estimated the length of time correctly or incorrectly. Miltner et al. kept the correct/incorrect feedback at 50/50 by continually varying the time window where a response by the participant would receive correct feedback; each time the participant input a response within the time window, the window was narrowed by 10 ms, and each time they responded outside of the window, the window was increased by 10 ms. In this way, the number of correct and incorrect

trials were balanced. Similar to the aforementioned studies, Miltner et al. found a negative-going peak in the ERP on incorrect trials. However, the peak was maximal approximately 250 – 350 ms after feedback—not response—was given. Additionally, while this negativity had a similar scalp distribution as the ERN (fronto-central), source localization pointed to a different point of origin within the brain, unless the signal was generated by a combination of the ERN and the P300. Thus, Miltner and colleagues concluded that it was a manifestation of the ERN, but with increased latency. Further, they theorized that the ERN was evidence of a generic error detection system in the brain. Then, in 2002, Gehring and Willoughby designed a gambling task where participants could lose even having made the correct choice; losses elicited a negative deflection approximately 250 ms post-feedback, regardless of whether the feedback indicated a correct or incorrect choice, which implied that the observed negative deflection post-feedback was not related to performance errors (Gehring & Willoughby, 2002). Gehring and Willoughby termed this component the medial-frontal negativity, or MFN. Further studies continued to show that the MFN and the feedback-locked negative waveform that Miltner and colleagues' 1997 paper had reported shared several characteristics, including timing, direction of deflection in the waveform, and scalp topography (Hajcak, Moser, Holroyd, & Simons, 2007; Yeung, Holroyd, & Cohen, 2005). Notably, in 2002, Holroyd and Coles published a paper that sought to provide a unified theory of the ERN and error detection in the brain (Holroyd & Coles, 2002). Because of the evidence that both the ERN as well as the feedback-locked ERN are generated by the dopaminergic system, Holroyd and Coles put forth a model in their paper that both components are the result of the same error detection system, and that this system recognizes both internal errors via evaluation of an efference copy of motor commands as well as externals errors by

evaluating feedback from the agent's environment, e.g. correct or incorrect feedback in response to an action (Holroyd & Coles, 2002).

More recent papers have reframed the feedback-locked ERN as the *reward positivity*: a more positive deflection on correct trials rather than a more negative one on incorrect trials, that still functions as an error prediction signal (Holroyd, Pakzad-Vaezi, & Krigolson, 2008; Proudfit, 2015). Holroyd, Pakzad-Vaezi, and Krigolson postulated that the "negativity" previously reported is actually driven by the N200 ERP component, which has a similar latency and topographic presentation to the ERN (Coles & Rugg, 1995; Donchin, Ritter, & McCallum, 1978; Holroyd, 2004; Holroyd et al., 2008; Proudfit, 2015). Because ERP components are not representative of individual and separable neural processes, each one is the summation of the signals of multiple, overlapping processes, and so it is possible that what appears to be a negativity in response to incorrect feedback is simply the N200, which has been shown to respond to task-relevant stimuli in general (Kappenman & Luck, 2012). In their 2008 paper, Holroyd et al. reported on the results of an experiment that attempted to dissociate the neural "base" response to task-relevant stimuli from negative and from positive feedback. Holroyd et al. used a difference wave approach to show that what had been reported as a negative deflection on error trials (the fERN) was, in fact, a positive deflection on correct trials: that is, the difference between the waveforms generated by each type of feedback indicates that neural activity changes in response to correct feedback. Further work dissociating the effects of correct, incorrect, and neutral feedback provided more evidence that the error prediction signal was the positive deflection seen in cases of correct feedback; for example, in Foti and Hajcak's 2009 paper, they used Principle Component Analysis (PCA) to analyze ERPs generated by gains and losses in a gambling task (Foti & Hajcak, 2009). PCA is a factor analysis approach that can used to separate

out the underlying components that compose a spatiotemporal signal such as ERPs (Dien, 2010; Donchin & Heffley, 1978). Foti and Hajcak found that the difference in the ERPs for gains versus losses was explained by a PCA factor which was fronto-centrally situated, positive, and peaked approximately 300 ms after feedback that indicated a gain—and was reduced following feedback that indicated a loss (Foti & Hajcak, 2009). Foti, Hajcak, Weinberg, and Dien then performed a follow-up study that further teased apart the question of whether this component was a positivity added on correct trials, or a negativity on incorrect trials (Foti, Weinberg, Dien, & Hajcak, 2011). Foti et al. replicated the results of the Foti and Hajcak 2009 paper, finding that wins elicited a more positive PCA component in the time range of the feedback-locked ERN, thus providing more evidence that the feedback-locked ERN is a positive modulation on correct feedback rather than a negative modulation on incorrect feedback.

An important note is that multiple names have been given to this ERP component over the years: besides medial-frontal negativity/MFN, it has been called the feedback-related negativity (FRN), the feedback negativity (FN), the feedback correct-related positivity (fCRP), the feedback-error related negativity (fERN), and the reward positivity (Bress, Meyer, & Proudfit, 2015; Gehring & Willoughby, 2002; Holroyd & Coles, 2002; Holroyd et al., 2008; Miltner et al., 1997). To avoid confusion, for the remainder of this document, I will use the term *reward positivity* when referring to a positive-going deflection that occurs 250 – 350 ms post-stimulus on correct feedback and has a frontal-central maximum at the scalp (Holroyd et al., 2008; Proudfit, 2015).

The amplitude of the reward positivity has been linked to both the unexpectedness of the stimulus, as well as whether the task has been learned. In Holroyd, Krigolson, and Lee (2011), they replicated an experiment originally done by Potts et al, where participants played a passive

"gambling-like" task (Potts, Martin, Burton, & Montague, 2006). On each trial, participants were presented with an initial stimulus (S1), which was either a lemon or a gold bar (with 50% probability of each); S1 was followed by and predicted with 80% accuracy a second stimulus (S2), which was again either a lemon or a gold bar. When S2 was a lemon, participants won no money, however, when S2 was a gold bar, participants won $0.10 CAN. Participants did not respond other than pressing the space bar between blocks. Unlike in Potts et al., where only the ERPs following S1 were analyzed, Holroyd et al. examined the ERPs following both S1 and S2, looking at four conditions: S1 predicted a reward and S2 delivered one; S1 predicted no reward and S2 did not deliver one; S1 predicted a reward and S2 did not deliver one; and S1 predicted no reward and S2 delivered one. These four conditions can be categorized into consistent, or Expected, as in the first two; and inconsistent, or Unexpected, as in the latter two. Holroyd et al. found that the amplitude of the difference wave in the time period of the reward positivity for the Unexpected conditions was larger than for the Expected conditions, which is what RL theory posits.

Additionally, RL theory states that the prediction error will decrease in amplitude as a task is learned, and the value of the action more closely reflects the value of the reward. In their 2014 paper, Krigolson, Hassall, and Handy linked the amplitude of the reward positivity to whether the task is new or learned (Krigolson, Hassall, & Handy, 2014). Krigolson et al. had participants play a simple gambling game: on the first trial of each block, participants were presented with two uniquely-coloured squares, one on each side of a fixation cross positioned in the centre of the screen; participants then selected one or the other and received feedback in the form of the numbers "0", "1", or "2", indicating that they had won zero, one, or two cents (CDN), respectively. Note that if the coloured square chosen was the "no-reward" choice, they

always received zero, but if the square was the “reward” square, they could receive either one or two cents, with a 50% probability of each. On the second trial of each block, participants were presented with two uniquely coloured squares that were not the same colours as the squares from the first trial. Again, participants selected one of the squares, and received feedback indicating the reward or lack thereof. Importantly, the results of both gambles were deterministic, so that participants could use the feedback from each gamble to learn the correct colour square to receive a reward for each one. Following the first two gambles, participants were then shown the pair of coloured squares from either the first gamble or the second. If participants had learned from the initial trial where that gamble was presented, then they knew which was the correct (rewarding) choice. To complete a block of trials, participants were then presented with three more gambles chosen randomly from either the first or second pair of coloured squares. From this, Krigolson et al. hypothesized that the amplitude of the reward positivity time-locked to the feedback would initially be high and then decrease as the participant learned the correct choice, and that an ERP component similar in time and topology to the reward positivity but time-locked to the choice presentation would appear as the correct choice was learned within each block. Upon analysis, Krigolson et al. found a difference in the reward positivity amplitude between the three feedback conditions (No Reward, Reward 1, and Reward 2) on the first trial of each block. They also found that, as predicted by reinforcement learning theory, there was a decrease in amplitude of the reward positivity ERP component with each trial: Trial 1 resulted in the largest amplitude, and Trial 3 resulted in the smallest amplitude. However, Trial 2 amplitude did not significantly differ from Trial 3. Simultaneously, an ERP component presenting like the reward positivity increased in amplitude from Trial 1 to Trial 2 and 3 (the difference between the latter two did not reach statistical significance). Thus, Krigolson et al. demonstrated that the reward

positivity presents post-feedback in a new, unlearned task, but once the task is learned the amplitude of the reward positivity at feedback decreases, while the amplitude of an identical ERP component time-locked to the choice cue that predicts a reward increases. Their result both aligns with RL theory predictions as well as with the physiological evidence from neuroimaging studies where phasic dopamine signals have been recorded during learning and performing tasks (Matsumoto, Matsumoto, Abe, & Tanaka, 2007; Schultz, Dayan, & Montague, 1997; Schultz et al., 1995). All this evidence further points to the reward positivity as the resulting neurophysiological signal generated by a reward prediction error system in the brain.

Although the work discussed in the rest of this thesis is largely theoretical in nature, it is worth noting the anatomical areas of the brain involved in error and reward prediction signals. In terms of the anatomical originators of the ERN and feedback-locked ERN, Schultz and others' work in the 1990s on the properties of midbrain dopamine neurons was crucial as it described the behaviour of these neurons in behaving primates (Romo & Schultz, 1990; Schultz, 1998; Schultz et al., 1995). The signals produced by these neurons were consistent with a prediction error as hypothesized by RL theory. Montague, Dayan, and Sejnowski then advanced a theory that these phasic dopamine signals, produced in the ventral tegmental area (VTA), function as a temporal-difference error signal that is sent to other brain regions (Montague, Dayan, & Sejnowski, 1996). Montague et al.'s model of error prediction in the brain was again consistent with physiological data recorded from primates. Further, Montague and colleagues speculated that the substantia nigra, nucleus accumbens, and possibly the amygdala might be involved in storing the weight changes that result from these TD error signals. In a 1997 paper by Schultz, Dayan, and Montague, the authors propose that the phasic dopamine signal from the VTA is consistent with a scalar reward prediction error signal (Schultz et al., 1997). This conclusion resulted from

Schultz et al.'s work modeling results from learning experiments in rats. The researchers further postulated that the dopamine signal from the VTA and substantia nigra is received by the striatum, where it would have an effect on behavioural choices by modulating competition among excitatory cortical inputs. The anatomical origin of the feedback-locked ERN/reward positivity have since been investigated using human functional magnetic resonance imaging (fMRI). Studies have shown a correlation between a Blood Oxygen Level Dependent (BOLD) contrast signal in the ventral striatum and a prediction error signal (Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008). Additionally, several papers have reported on source localization for each component, and both ERN and feedback negativity/MFN appeared to be localized to the medial dorsal anterior cingulate cortex (Holroyd & Coles, 2002; Holroyd, Yeung, Coles, & Cohen, 2005). Further research has continued to provide evidence that these midbrain dopamine signals constitute a prediction error signal that is sent to diverse areas of the brain, including the frontal cortex, thalamus, striatum, globus pallidus, and the amygdala (Doya, 2008; Schultz, 2015).

1.4.3 P300.

The second ERP component that is relevant for this thesis is the P300, which has been linked to context switching, memory, and the use of attentional resources (Polich, 2007). The P300 was first identified in the 1960's by Sutton, Braren, Zubin, and John, who identified a positive deflection in the difference wave when comparing average evoked potentials in a cued multi-modal stimulus task where the initial stimulus varied in the certainty with which it predicted the second stimulus (S. Sutton, Braren, Zubin, & John, 1965). Sutton et al. found a positive deflection that peaked approximately 300 ms after the second stimulus, and was larger in amplitude for trials with the 'uncertain' first stimulus. Over the following years, multiple

research teams examined this positive waveform, labeled the P300 (Ritter & Vaughan, 1969; Ritter, Vaughan, & Costa, 1968; S. Sutton, Tueting, Zubin, & John, 1967). Notably, in 1975, Nancy Squires, Kenneth Squires, and Stephen Hillyard determined that what had been referred to as the P300 was in fact two separate components: the P3a and P3b (N. Squires, K. Squires, & Hillyard, 1975). Squires et al. disentangled the two and showed that the P3a was maximal over fronto-central regions with a peak latency of 220 to 280 ms, while the P3b was maximal over parieto-central regions and had a peak latency between 310 to 380 ms. For the remainder of this thesis, we will be looking only at the P3b, and thus I will refer to it simply as the "P300", as is common in ERP literature (Luck, 2014).

Surprisingly, for all the research done on the P300, there is not yet consensus on the underlying neural processes that cause it (Polich, 2007). However, in his 1981 paper, Donchin developed a preliminary model of what underlying processes the P300 may reflect, examining the effects of stimulus probability, task relevance, and attention (E. Donchin, 1981). Called the *context-updating theory*, Donchin stated that the P300 is generated when the brain revises or modifies its internal representation of the environment based on incoming stimuli (E. Donchin, 1981). The P300 is thought to reflect a "strategic" process that is involved in long-term planning and behavioural control, as well as probability mapping within the environment and updating biases (E. Donchin & Coles, 1988). Thus, the P300 occurs after initial sensory processing and may be caused by inhibitory neural activity in order to focus attentional resources on task-relevant stimuli, and subsequently to transfer those stimuli to working memory in service of context updating (E. Donchin & Coles, 1988; Polich, 2007).

The amplitude of the P300 is affected by several factors, including stimulus probabilities, the amount of attentional resources engaged by the task, and target-to-target interval length

(Polich, 2007). Both global as well as local stimulus probabilities have been shown to affect the amplitude of the P300; as the probability of the target stimulus decreases, the amplitude of the P300 increases (Duncan-Johnson & Donchin, 1977, 1982; Johnson Jr. & Donchin, 1982; Squires, Wickens, Squires, & Donchin, 1976). Multiple studies have also shown that the P300 amplitude depends on the cognitive load of the subject, which can be manipulated by setting both a primary task in which the cognitive difficulty is increased and decreased over time, and a secondary task involving discriminating between frequent and infrequent stimuli such as an oddball paradigm (e.g. Isreal, Chesney, Wickens, & Donchin, 1980; Kramer, Wickens, & Donchin, 1985; Wickens, Kramer, Vanasse, & Donchin, 1983). Thus, as the cognitive load of the primary task is increased, the amplitude of the P300 decreases, and conversely, when the cognitive load of the primary task is decreased, the amplitude of the P300 increases. Finally, the target-to-target interval length affects P300 amplitude, with a longer interval between presentations of the stimulus correlated with a higher amplitude (Polich, 1990, 2007).

The latency of the P300 component is also affected by several factors, including stimulus evaluation timing, task processing demands, and individual cognitive capabilities (Polich, 2007). Stimulus evaluation timing is the amount of time required to detect, evaluate, and categorize a stimulus; as the P300 occurs after those processes have completed, stimuli that are harder to perceive or are more complex will increase the latency of the P300 (Polich, 2007). Task processing demands also can increase latency of the P300 when the response requires effort (Polich, 2007). Finally, the latency of the P300 is highly dependent on individual cognitive capabilities, with better cognitive performance correlated with shorter latency (Polich, 2007). Genetics, age, and brain health all have been shown to impact P300 latency (Polich, 2007).

1.5 Analysis of EEG Using Machine Learning Methods

In the following sections, I provide an overview of prior research using machine learning (ML) analysis methods on EEG data Additionally, as applying machine learning to EEG presents a number of challenges, I will elaborate on those issues, provide a few potential solutions that have evidence to date, and explain the algorithms that were used in this current work.

Machine learning analysis of EEG data has been investigated as an alternative to event-related potential analysis for decades; Donchin, for instance, published "Discriminant analysis in average evoked response studies: the study of single-trial data" in 1969, which detailed a single-trial analysis of EEG data from an evoked response experiment. In this paper, Donchin sought to show that differences between trials of EEG data was a meaningful endeavor in determining neural function, and presented a method using Step Wise Discriminant Analysis (SWDA) to calculate the Euclidean distance between samples to categorize those samples into one of four groups. Subsequently, Horst and Donchin used SWDA on a different EEG dataset; this dataset was generated by presenting a checkerboard pattern to either the upper or the lower half of participants' visual field (Horst & Donchin, 1980). The researchers divided the recorded data into training and test sets, using the former to train discriminant functions, and the latter to test the performance when the functions were fed the heretofore unseen testing data. By examining various channels, Horst and Donchin were able to get mean accuracy of up to 87.8% within subjects, and 78.1% mean accuracy between subjects.

More recent research on analysis of EEG data using various machine learning techniques has continued to show the validity of this type of approach. Ford, White, Lim, and Pfefferbaum investigated the P300 in people diagnosed with schizophrenia; there was strong evidence that schizophrenics had smaller amplitude P300s in averaged ERP studies, and Ford et al. wanted to

see whether those results were caused by variability in the amplitude of the P300 between trials, consistently small amplitude of the P300 across all trials, or latency variation of the P300 (Ford, White, Lim, & Pfefferbaum, 1994). By using single-trial analysis of EEG data from a two-tone auditory oddball experiment, they found that all three cases were true. Another example is Stahl, Pickles, Elsabbagh, and Johnson's work on identifying infants at risk of autism (Stahl, Pickles, Elsabbagh, Johnson, & The BASIS Team, 2012). Stahl et al. reanalyzed an EEG dataset collected from infants using two different machine learning algorithms, regularized discriminant function analyses and support vector machines; their work was motivated by the difficulties with collecting a suitably large number of trials required for averaged ERP analysis, with the goal of validating EEG analysis using machine learning as an alternative. Stahl et al. achieved above-chance classification accuracies on the dataset with 6-fold cross-validation, thereby showing that these methods are useful in differentiating at-risk groups where collecting large numbers of trials is prohibitive.

From these examples, it has become clear over the past couple decades that machine learning can be applied to EEG data and be used to solve outstanding questions with a high degree of success. Thus, for this present experiment, I chose to apply machine learning techniques to analyze the EEG data generated by the first experiment discussed in this thesis to determine whether I could classify segments of EEG by the states that generated those segments. The following section provides background and summary of the specific machine learning methods that were employed in this present work.

1.6 Support Vector Machines and EEG Data

Machine learning (ML) algorithms can be split into three groups: unsupervised, supervised, and reinforcement learning. As noted in section 1.2, reinforcement learning is neither a supervised nor an unsupervised method: learning occurs via trial-and-error experimentation by the system while maximizing a reward function (R. S. Sutton & Barto, 1998). Unsupervised learning methods generally look for previously unrecognized patterns in the data; principle component analysis and clustering algorithms are two examples of unsupervised learning (Hinton, Sejnowski, & Poggio, 1999). Supervised learning systems, on the other hand, rely on expert instruction, usually in the form of a training set composed of labelled examples (Kotsiantis, Zaharakis, & Pintelas, 2007). For classification algorithms, these labels indicate the category to which the example data belongs. Linear discriminant analysis, support vector machines, artificial neural networks, and kernel estimation are all examples of supervised classification methods. Many machine learning techniques have been used with success for performing analysis of EEG data, especially within the brain-computer interface field (e.g. Bashashati, Fatourechi, Ward, & Birch, 2007; Kumar, Dewal, & Anand, 2014; Lotte, Congedo, Lécuyer, Lamarche, & Arnaldi, 2007; Nicolaou & Georgiou, 2012). Consistently, Support Vector

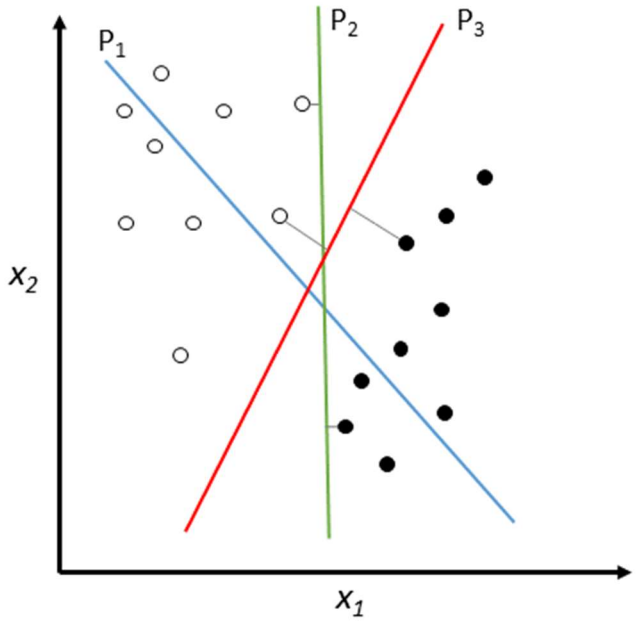

Figure 2. Support Vector Machine Separating Hyperplanes. P_1 does not separate groups; P_2 does but with a very small margin; P_3 separates the groups with the maximum margin.

Machines (SVMs) have been among the best-performing ML algorithms for classifying EEG data, taking into account accuracy, robustness, and many freely available implementations (Lotte et al., 2007; Quitadamo et al., 2017). The robustness of SVMs means they are able to handle several characteristics of EEG data that make some other ML methods inadvisable: the non-stationarity of signal, low signal-to-noise ratio, and high dimensionality (Blankertz, Lemm, Treder, Haufe, & Müller, 2011; Lotte et al., 2007).

The support vector machine technique evolved out of earlier work done by Fisher's work on linear discriminant analysis for pattern recognition and was first published in its current form by Russian researchers Vapnik and Chervonenkis in 1974, although the fundamental algorithm used in SVMs—the Generalized Portrait algorithm—was published in 1963 (Fisher, 1936; Vapnik & Chervonenkis, 1974; Vapnik & Lerner, 1963). The Generalized Portrait algorithm was later extended to a nonlinear form and incorporated into the SVM method. Currently, there are many variants of the SVM method: least-squares, fuzzy, multiple-kernel, spatially-weighted, and more (Quitadamo et al., 2017). However, the essential idea behind SVMs is to define a hyperplane dividing a set of data points into two groups, such that we maximize the margin between the hyperplane and data points on each side closest to the hyperplane (Alpaydin, 2014). The closest data points to the hyperplane are called support vectors, hence the name of the method. An example in two dimensions with three potential separating hyperplanes is depicted in Figure 2. By using support vectors, much of the data can be ignored, thus reducing computational effort.

SVMs have been used extensively to classify EEG data. Much of the work to date has focused on detecting neurological conditions such as epilepsy, Parkinson's, or other neurological disorders (e.g. Kumar et al., 2014; Lotte et al., 2007; Nicolaou & Georgiou, 2012; Pestian et al.,

2016; Tawfik, Youssef, & Kholief, 2016; Tibdewal & Tale, 2017; Yuvaraj, Rajendra Acharya, & Hagiwara, 2018). SVMs have also seen substantial use within the brain-computer interface (BCI) research community (Lotte et al., 2007; Nicolaou & Georgiou, 2012; Panda et al., 2010; Pestian et al., 2016). However, less research has been done on examining epoched neural responses in neurologically intact humans and other animals, although a recent study in rats by Hyman, Holroyd, and Seamans used SVMs to classify data recorded from the anterior cingulate cortex (ACC) during a two-choice probabilistic gambling task analogous to one used in humans (Hyman, Holroyd, & Seamans, 2017).

Advantages to using SVMs for EEG as opposed to other classification methods are: one, they can handle high-dimensional inputs, including the case where the dimensionality of the data exceeds the number of samples; two, they maintain reasonable accuracy in the face of noisy inputs; and three, SVMs always find the global solution to a problem and not local minima (Burges, 1998; Vapnik, 1998). In terms of this thesis, the most relevant disadvantages to SVM are that it can be difficult to determine the best kernel and optimal parameters, and that the compute time can be lengthy for large, high-dimensional datasets (Suykens, Van Gestel, De Brabanter, De Moor, & Vandewalle, 2003). The former can be mitigated, however, by using a grid search algorithm to automatically assess a variety of parameter values for highest accuracy in a methodical manner (Quitadamo et al., 2017). Compute time can be limited by reducing the dimensionality of the data or the number of samples.

1.7 Summary

The preceding sections of this chapter have laid out the background and previous work that I build on in the experiments detailed in the following two chapters. Sections 1.2, 1.3, and 1.4 detailed the development of reinforcement learning theory, the evidence to date of an

implementation of reinforcement learning in the brain, and finally discussed several computational models of reinforcement learning that have been utilized to model electroencephalogram results. In section 1.5, the history of machine learning analysis of electroencephalogram data was summarized, and several motivations for this type of analysis were put forward. Finally, in section 1.6, the background for the method of ML analysis performed as part of the work detailed in this thesis was explained, as well as the rationale for its use.

The primary research question addressed by this thesis is to ascertain whether or not the responses of a neural reinforcement learning system mirror theoretical predictions. Specifically, while it is fairly well established the neural responses evoked by feedback appear to reflect RL prediction errors (Holroyd & Coles, 2002; Holroyd et al., 2005; Krigolson et al., 2014; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003), it is considerably less clear whether the theoretical values associated with mid-value states are indeed conveyed within the human EEG spectra and whether entering each of those states can also evoke RL prediction errors. Is there a detectable neural signal when an action results in a shift to an environmental state that is closer to a goal state but is still an intermediate step on the way to achieving that goal? For instance, if your goal is to get home while driving a car from work, in theory arriving at the intersection closest to your house has more value that the first intersection after you leave the office. And, imagine you found that you could not take the most valuable (shortest) route home—in principle this would elicit a prediction error as you would be transitioning from a higher value state to a lower value state. Likewise, can we see a change in the ERP components when an action results in a shift the other way, to a state that is further from the goal state? These are the types of questions this research aims to answer.

Based on my primary research question, herein I present the results of one EEG experiment and a subsequent machine learning analysis of the EEG data. In the EEG experiment detailed in Chapter 2, I predicted that there would be a difference in the reward positivity between 'Expected' and 'Unexpected' events, and that Good-Unexpected events would cause a larger reward positivity amplitude than Good-Expected events. Additionally, I predicted that there would be a difference in the P300 when expectancy is violated: unexpected events should produce larger P300 components. Furthermore, I hypothesized that these differences in the reward positivity will align with prior results and RL theory (Krigolson et al., 2014). For the machine learning experiment, detailed in Chapter 3, I predicted that I would be able to successfully classify epoched, feedback-locked EEG data according to which of the seven states the participant was currently in (e.g. State 1 versus State 2 versus State 3, etc.), thus showing that it is possible to differentiate between intermediate states in a multi-step task. I also predicted that I would be able to classify, with above-chance accuracy, epoched, feedback-locked EEG data from each of the pairs of conditions examined and analyzed with ERP methods in Chapter 2, such as "Good-Expected" versus "Good-Unexpected", and "Bad-Expected" versus "Bad-Unexpected", among others.

Chapter 2: Experiment One – Reward Positivity and P300 ERP Components in a Learned Non-Deterministic Task

2.1 Introduction

It is well established in prior literature that outcomes that are more unexpectedly rewarding elicit a larger neural response than those than are expected (Bress et al., 2015; Gehring & Willoughby, 2002; Holroyd & Coles, 2002; Holroyd et al., 2008; Miltner et al., 1997; Proudfit, 2015). Early research using event-related brain potentials (ERPs) such as Miltner, Braun and Coles (1997) framed the difference as a negative deflection in response to feedback that indicated an error; this same negative deflection was seen in Gehring and Willoughby's 2002 study where participants could "lose" in a gambling game, even though they had chosen the correct response (i.e. the one that was statistically more likely to be a win). Gehring and Willoughby's results showed that the ERP component in question was not due to performance errors but instead due to the external feedback received as part of the task. More recently, it has been recognized that this ERP component is a positive deflection on correct—or rewarding—trials, rather than a negative deflection in response to unfavourable feedback (Holroyd et al., 2008; Proudfit, 2015). Many studies have now confirmed that the reward positivity appears in response to rewarding feedback in unlearned tasks in many animals, including humans, and is absent in the case of negative feedback (e.g. Bayer & Glimcher, 2005; Holroyd & Coles, 2002; Holroyd, Nieuwenhuis, Yeung, & Cohen, 2003; Holroyd et al., 2008; Krigolson, Pierce, Holroyd, & Tanaka, 2009).

It has also been shown that as learning progresses, the amplitude of the reward positivity in response to positive feedback decreases (Krigolson et al., 2009; 2014). Once a task is learned, and the subject is familiar with the reward schedules for each cue, the reward positivity is

elicited by cues that predict the reward, rather than by the reward itself (e.g. Krigolson et al., 2014; Romo & Schultz, 1990; Schultz, Apicella, & Ljungberg, 1993; Schultz & Romo, 1990). Reinforcement learning theory posits that as values propagate back to cues or decision points that predict future rewards, prediction errors occur when these predictors are presented, something again seen in human neural data (Holroyd et al., 2011; Krigolson et al., 2014; R. S. Sutton & Barto, 1998). In a learned task, when the outcome of a choice is non-deterministic, the feedback-locked reward positivity will still appear in cases of an "unexpected" outcome; i.e. if there are two outcomes of a choice indicated by a specific cue, the reward positivity amplitude will be greater for the less frequent reward outcome (Holroyd & Krigolson, 2007; Sambrook & Goslin, 2015). Additionally, the reward positivity has been shown to be larger when feedback is more positive than expected in a learned task, such as when an erroneous response results in positive feedback (Holroyd & Krigolson, 2007b; Holroyd et al., 2011). These appearances of the reward positivity have been interpreted to indicate the updating of action selection values as per reinforcement learning theory (Holroyd et al., 2008; Proudfit, 2015).

However, what happens when more than one sequential choice must be made—and made correctly—to achieve a reward? Using the example of a game of Tic-Tac-Toe, it is clear that this task involves multiple choices in sequence, one per turn. The reward is winning the game—it seems obvious that one's last turn where the winning mark is placed on the game board would be highly rewarding and thus elicit a reward positivity: goal achieved! But what about the preceding turn, when the correct choice is made? Does that decision cue (to place a mark in the correct square, or to place a mark in the incorrect square) at that point attain value? Is there a reward positivity in response to seeing that game state? RL theory states that the reward prediction error would appear at time of the earliest choice presentation that indicates a reward is available (R. S.

Sutton & Barto, 1998). Thus, we would expect to see the reward positivity at the first choice cue—in this case, the earliest game state—that indicates that one will win the game.

A simpler example, and one that is pertinent to the experiment described in this chapter, is a one-dimensional random walk. A random walk is a process that traverses a path through a

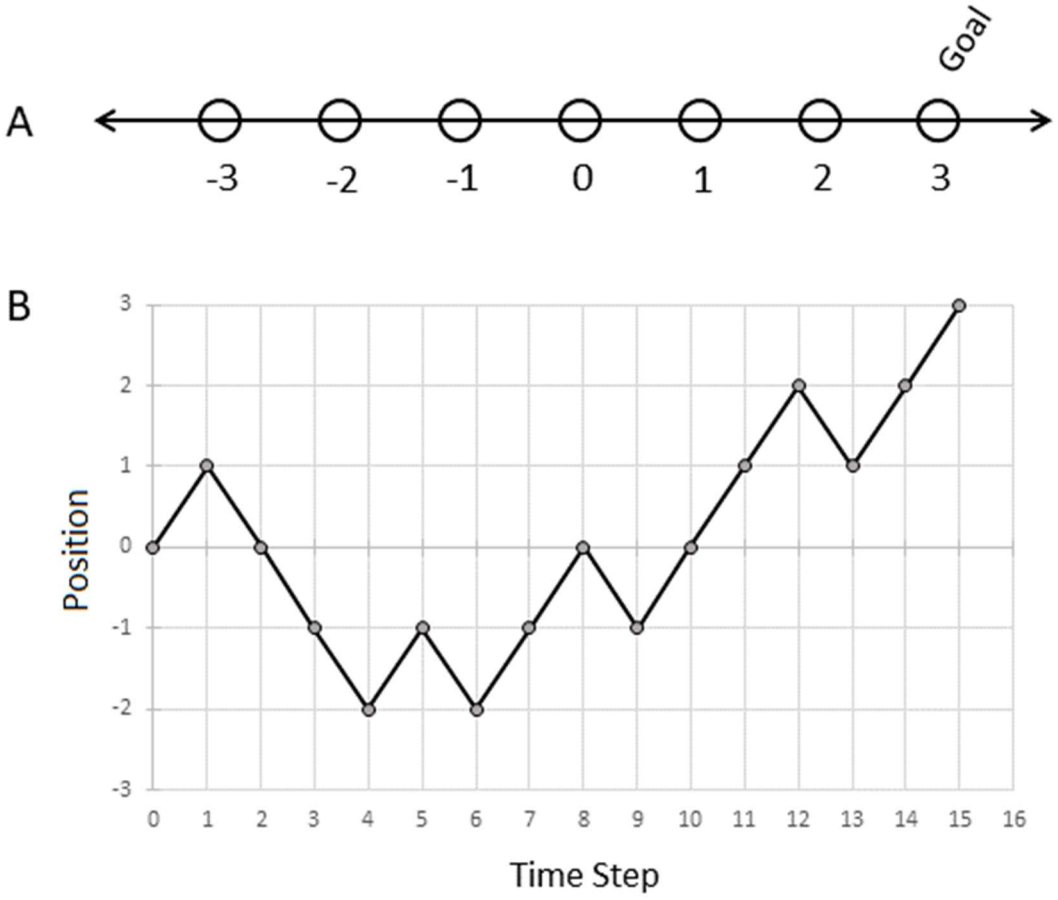

Figure 3. Illustration of a Random Walk. A: a one-dimensional mathematical space consisting of the integers from -3 to 3 inclusive. B: An example random walk through the above space.

mathematical space (Florescu, 2014). For this example, I will use the integer number line from -3 to +3 inclusive (see Figure 3A). As the name implies the process that generates this path is random, or stochastic, moving from one point in the space to an adjoining point randomly with equal probability. An important point about random walks is that eventually, every point within

the space will be visited; thus, if +3 is designated the goal state for the process, it is guaranteed to visit that point (Florescu, 2014). In Figure 3B, a sample random walk within this one-dimensional space is shown, with the process visiting a series of points and ending at the goal state. In this example, if the process traversing the depicted path is using reinforcement learning to learn the shortest path from the start state to the goal, the state value of position 2 would be updated after traversal to a more positive state value, indicating that it indicates that a reward is imminent as long as the correct subsequent choices are taken (R. S. Sutton & Barto, 1998). As the process performs successive iterations of random walks on this mathematical space, the value associated with position 1 will eventually be updated to a more positive state value as well (see Figure 4), and so forth. While this is an extremely simplistic case compared to the commuting-home example discussed earlier, it serves to clarify the questions I am attempting to answer in this thesis. Namely, I want to examine what happens in the human brain when transitioning between states that are nearer to or further away from a goal.

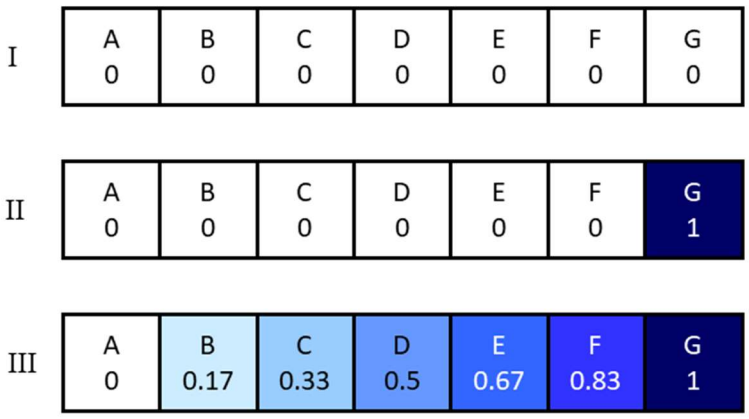

Figure 4. Illustration of State Value Acquisition During a Learning Task. I: State values are all zero at the beginning of learning. II: The goal state, *G*, is assigned the value of 1. III: After learning is complete, state values show a linear trend from 0 at the loss state, *A*, to 1 at the goal state, *G*.

In the experiment described in the remainder of this chapter (Experiment One), my goal was to see if reinforcement learning theory accurately predicted the behaviour of the reward positivity ERP component during a task that required multiple, sequential choices in

order to achieve a goal. I sought to extend existing research (Krigolson et al., 2014) into the reward positivity by looking at a learned multi-step task modeled after the one-dimensional random walk discussed above, instead of a two-armed bandit task as in Krigolson et al.'s study. I hypothesized that there would be a difference in the reward positivity amplitude depending on valence and expectancy of events and cues, in line with RL theory and prior literature on human learning and decision-making. Additionally, I predicted that the amplitude of the P300 ERP component would be larger for unexpected events than for expected events, as has been shown in prior literature.

2.2 Methods

2.2.1 Participants.

Thirty-three participants were recruited from the undergraduate population at the University of Victoria through the Department of Psychology's online system (SONA) for participation in experiments and via word-of-mouth. Undergraduate students recruited through SONA were compensated for their time with course credit; non-student participants received no compensation. Written, informed consent for each participant was obtained prior to beginning the experiment. This study was approved by the University of Victoria's Human Research Ethics Board (protocol number 16-428) and was conducted in accordance with the 1964 Declaration of Helsinki and all subsequent amendments.

Of the thirty-three datasets resulting from testing, one was removed from analysis due to a disqualifying neurological condition (epilepsy) divulged after testing had begun. Three were removed due to performance below minimum acceptable levels of accuracy. One was removed due to abnormal presentation of the reward positivity ERP component. This left twenty-eight

participants' data for inclusion in analysis (range = 18 to 62 years, M_{age} = 23.4 years, SD = 8.78[2], 20 females/8 males).

2.2.2 Apparatus and Procedure.

Upon arrival at the lab, participants were greeted by the researchers and shown to a testing room. Written informed consent was obtained before further action, and researchers were available to address any questions the participants had about the consent form and procedure. Participants were seated comfortably in a sound-dampened room in front of a standard computer equipped with a 22" widescreen LCD monitor, speakers, a mouse, and a keyboard. Distance between participant and monitor was approximately 100 cm.

Participants played a computerized game called Wizards' Duel. During Wizards' Duel, participants first learned how to traverse a linear random-walk-style task with seven discrete states including a loss state (1), a reward state (7), and five intermediate states. In the second phase of the task, after learning was complete, they played against a computer opponent that worked to make the participant lose. The aim of Phase One was for the participant to learn the correct sequence of actions to achieve their goal. The participant used the 'c' and 'm' keys on the keyboard to navigate through the series of states, each represented by a coloured fruit icon. Figure 5 shows an example of the game play in Phase 1; the participant starts in the state denoted by the apple icon, and must figure out that the 'c' key moves them to the state denoted by the lemon, and subsequently the 'm' key moves them to the state denoted by the plum, which for this example is the 'win' or goal state. In each of the intermediate states, participants had to learn

[2] One participant was 62 years old; the remainder were between 18 and 35.

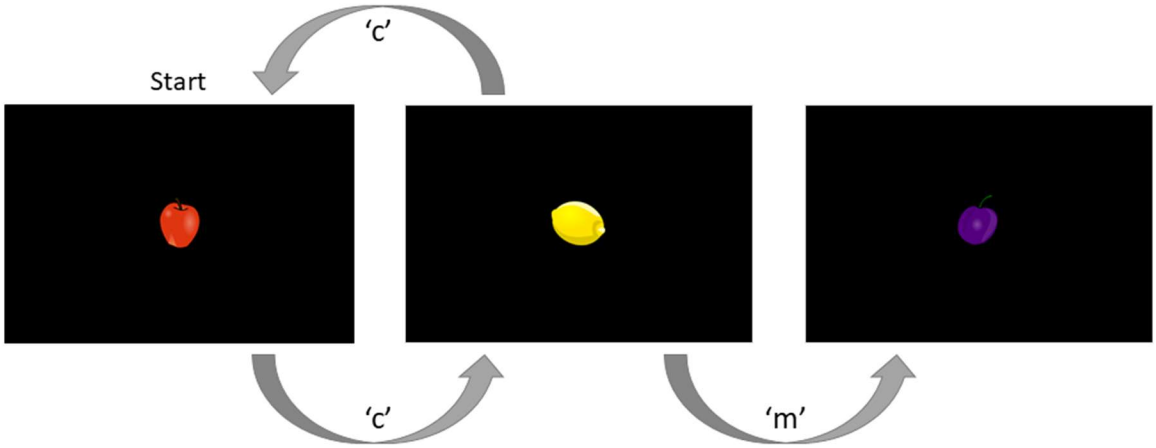

Figure 5. Wizards' Duel Phase One Example: goal state is 'plum', starting state is 'apple'. Participants must learn which key ('c' or 'm') will get them from apple to plum. In this case, 'c' moves from apple to lemon, and 'm' moves them from lemon to orange.

which key press moved them to a more desirable state, i.e. a state that was closer to their goal state. In other words, for each of the seven states there was a correct response and an incorrect response. There was no time limit on how long the participant could take before responding in this phase, nor was there any delay between stimulus presentations. Additionally, in this phase, choices were deterministic; that is, the chosen response always moved the participant in the intended direction—correct responses moved them one state closer to their goal and incorrect responses moved them one state further away from their goal. Phase One began with several screens of instructions for the participant to read at their own speed, and as part of the instructions, the participant was informed of the fruit icon that represented the goal state (a "win") as well as the fruit icon that represented the state to avoid (a "loss"). The instructions were as follows in Table 1.

Table 1

Wizards' Duel Phase 1 Instructions to Participant

Page	Text of Instructions
1	Welcome to magic school! Today you will learn to cast two simple transmogrify spells. Cast CONVERT by pressing the 'c' key. Cast MODIFY by pressing the 'm' key. (Press any key to proceed)
2	For safety reasons, you will be practicing on fruit. All fruit can be transformed into other fruit using the correct spell. Unfortunately, you do not yet know (and must learn) the effect of each spell.
3	During the following practice trials, your goal is to turn the starting fruit (which may be different each round) into <goal_fruit_name> while avoiding <avoid_fruit_name>. Note that you may not be able to turn the starting fruit into <goal_fruit_name> directly - you may have to pass through other fruit to get there.
4	Every time you see a fruit, cast either CONVERT ('c') or MODIFY ('m'). The fruit will then be transformed. Once you see <goal_fruit_name> you have succeeded. You will then hear an ascending tone, and the round will be over. If you see <avoid_fruit_name> you were unsuccessful. You will then hear a descending tone, and the round will be over.
5	Occasionally you will receive a grade that reflects the proportion of times, since you were last graded, that you picked the correct spell (spells that moved the fruit closer to <goal_fruit_name>).
6	Ready to begin spellcasting practice?

Upon reaching the goal state, an ascending pair of audio tones (a 500 Hz tone for 250 ms, followed by a 600 Hz tone for 250 ms) played and the round ended. If the participant ended up in the "avoid"/"loss" state, a descending pair of audio tones (a 300 Hz tone for 250 ms, followed by a 200 Hz tone for 250 ms) played, and the round ended. Every ten rounds, the participant was given feedback in the form of a percentage score for how many correct moves they made during that block of trials. The order of the fruit and the key presses to move between states were randomized across participants, but were consistent within each participant for both Phase One

and Phase Two. When participants consistently attained a 95% accuracy rate or better, and the capping process was complete, they moved into Phase Two of the experiment.

In Phase Two, participants 'dueled' against a computer opponent to reach the win state, while the computer attempted to reach the participant's loss state. As in Phase One, Phase Two started with several screens of written instructions as shown below in Table 2.

Table 2

Wizards' Duel Phase 2 Instructions to Participant

Page	**Text of Instructions**
1	You have now passed your transmogrification class and may join the dueling club. Congratulations! You will be competing in several wizards' duels. As in class, your goal is to turn the starting fruit into <goal_fruit_name>.
2	You will be playing against a computer-controlled opponent named Oloster Milner (also known as Oloster the Unflinching). Oloster's goal is to turn the starting fruit into <avoid_fruit_name>.
3	Dueling has several important differences from classroom spellcasting: • the identity of the current spellcaster will be indicated by a coloured border (you are <colour_one> and Oloster is <colour_two>) • before you may cast your spell, you must wait for a cue (a BEEP) • finally, spells cast while dueling may not always work (they may backfire and have the opposite effect)
4	Try to keep your gaze on the centre of the display at all times. Wait for the BEEP before responding. Try to win as many duels as possible.

The game was turn-based, and outcome feedback (win/loss) was indicated by the same audible tones as in phase one. However, there were two crucial differences between phase one and phase two: first, the participant was competing against an opponent, and second, both the participant's and the opponent's responses had a percentage chance to backfire, i.e. have the

opposite effect. The chance of the participants' responses backfiring was set at 20%, while the computer opponent's chance of backfiring was randomized between 20% and 40% based on a calculation at the beginning of each block in Phase Two for each participant. The variation of the computer's backfiring chance was done to make the participant slightly more likely to win than lose, and therefore more likely to remain engaged with the task; the percentage chance of backfiring was recorded in the behavioural file for each participant.

Each trial started with the participant being placed in a randomized starting state with a bias toward states with the lowest prior visit count; this was done by calculating the per-state percentage of total state visits for each of the intermediate states, then generating a random number between 0 and 1, and selecting the first state where the random number is less than the percentage of visits to that state. The border around the grey box containing the fruit icon (see

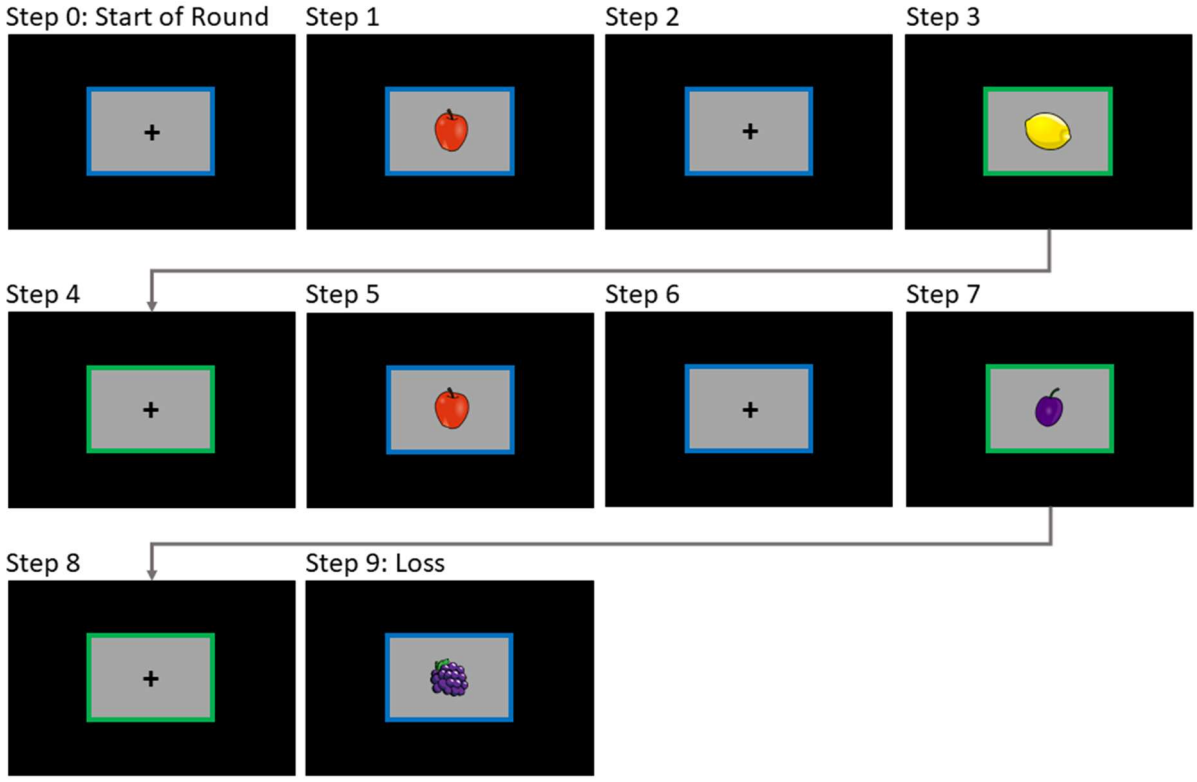

Figure 6. Wizards' Duel Phase Two Example.

Figure 6) indicated whose turn it was; the border colour was either blue or green, and the colours assigned to the participant and the computer opponent was randomized across participants. In the example shown in Figure 6, a blue border denotes that it is the participant's turn, while a green border denotes the computer's turn. At the start of the round, it is the participant's turn, and the state denoted by the apple icon (Step 1) is pseudo-randomly chosen for the starting state as described above. The participant responds with the correct choice, and moves to the state represented by the lemon (Step 3); now it is the computer's turn as indicated by the green border. The computer responds and the current state changes back to the apple. On the participant's turn (Step 5), they make the correct choice, but a backfire occurs and they are moved to the state represented by the plum (Step 7). The computer then takes its turn, and moves to the state denoted by the grapes (Step 9). Unfortunately for the participant, the grapes indicate that they have lost the round. As shown in Figure 6, a fixation cross was shown in between each presentation of the fruit icons for between 400 and 600 ms; this timing variability reduced the likelihood that the participant anticipated the onset of the stimulus. If it was the participant's turn, there was another 400 to 600 ms pause after stimulus presentation, and then a tone played (400 Hz for 150 ms) indicating that the participant could then respond. Participants had to input a valid response (either the 'c' or 'm' key as learned in phase one) within two seconds after the tone. Responses input after the two-second period were given the feedback "INVALID". Responses input before the tone were given the feedback "TOO FAST".

If the response results in a win for the participant (i.e. they have reached the goal state), a fixation cross was displayed for 400 to 600 ms, at which point an ascending pair of audio tones will be played (first tone 500 Hz for 250 ms, second tone 600 Hz for 250 ms). If the response resulted in a loss for the participant (i.e. they reached the avoid state), a fixation cross was

displayed for 400 to 600 ms, and then a descending pair of audio tones was played (first tone 300 Hz for 250 ms, second tone 200 Hz for 250 ms). Participants completed 10 rounds per block for as many blocks as possible within the time limit of the testing session. At the end of each block, as in Phase One, the participant was given feedback on their accuracy in the form of a percentage grade.

2.2.3 Data Acquisition.

In Phase One of the experiment, only behavioural data were collected. Reaction time in milliseconds, from the onset of the stimulus (the time at which a fruit icon was displayed on the monitor) to when the participant pushed a response key, was recorded by the MATLAB script. The script also recorded the participant's key choice and the correct/optimal response, which together indicated whether and how fast learning of the task occurred. This data was saved to a local text file formatted as tab-separated values.

In Phase Two of the experiment, both the above behavioural data as well as continuous EEG data were recorded. EEG data were recorded using Brain Vision Recorder software (Version 1.2, Brain Products GmbH, Munich, Germany) from 64 electrodes mounted in the standard 10-20 configuration in an ActiCAP (Brain Products GmbH, Munich, Germany), including two mastoid electrodes, at a sampling rate of 500 Hz. At time of data collection, electrodes were referenced to a common ground. Electrode impedances were kept below 20 kΩ on average. EEG data was amplified using a BrainProducts ActiCHamp 2 with an 8 kHz antialiasing low-pass filter. A DataPixx2 video unit was used to ensure accurate timing of markers in the EEG data stream and visual stimuli onset (“VPixx Technologies—DataPixx2 Display Driver,” 2017). The computerized task was coded in MATLAB and used the

Psychtoolbox library for timing of displayed stimuli (Brainard, 1997; Kleiner et al., 2007; Mathworks, 2017; Pelli, 1997).

2.2.4 Data Analysis.

Data processing post-recording was done as follows, using Brain Vision Analyzer software (Version 7.6, Brain Products GmbH, Munich, Germany): EEG data from each participant was visually inspected for overly noisy channels or flatline channels. Data was then downsampled to 250 Hz using spline interpolation, re-referenced to the averaged mastoid electrodes, and filtered using an infinite impulse response, dual-pass, zero phase shift Butterworth filter (bandpass 0.1 – 30 Hz, order 4, with a notch filter at 60 Hz). Subsequently, Independent Component Analysis (ICA) was run on whole data for each participant, and resulting components containing eye blinks were removed by visual inspection of the component waveforms, topographical head maps, and factor loadings of each component. Inverse ICA was then performed to recreate the original EEG waveforms minus the removed components. Following this, data was segmented into epochs of 1000 ms (-200 to 800 ms) around relevant markers and exported from Analyzer for further analysis in MATLAB. Within MATLAB, using custom code, an artifact rejection algorithm was done on each epoch, removing any trials that had voltage fluctuations of more than 100 uV.

Table 3

Labels for Categories of Events Used In ERP Analysis

Label	**Explanation**
Good	Any event in which the participant received feedback that they were closer to their goal state; i.e. the previous game state was State 3 and the new state is State 4.

Bad	Any event in which the participant received feedback that they were further away from their goal state (and therefore closer to the loss state) ; i.e. the previous game state was State 4 and the new state is State 3.
Expected	Any event where the participant received feedback that the action just taken by either the participant or the computer resulted in the most likely outcome.
Unexpected	Any event where the participant received feedback that the action just taken by either the participant or the computer resulted in the less likely outcome.

ERP waveforms were created by averaging the epoched EEG data for each condition (Good-Unexpected, Good-Expected, Bad-Unexpected, and Bad-Expected). Next, difference waveforms were made for each comparison (Good-Expected versus Good-Unexpected and Bad-Expected versus Bad-Unexpected) and scalp topographies were generated for conditions of interest (see Table 3 for the conditions and explanations). Visual inspection of the grand averaged conditional and difference waveforms indicated that the ERPs of interest (reward positivity and P300) had greater latency than usually seen in experiments; this was likely due to the complexity of the task and the feedback stimuli, which led to increased cognitive load (Krigolson, Hassall, Satel, & Klein, 2015; Krigolson, Heinekey, Kent, & Handy, 2012). Thus, based on prior findings in tasks with similar complexity, the maximum voltage on channel FCz within the window of 300 ms to 500 ms post-feedback stimulus was used to quantify the reward positivity, and the peak voltage on channel POz within the window of 300 ms to 600 ms post-feedback stimulus to quantify the P300. Effects of interest were examined using repeated-measures ANOVA and post-hoc paired samples *t* tests, with an alpha level of 0.05.

2.3 Results

2.3.1 Behavioural Results.

Participant behavioural data was analyzed to determine overall accuracy for both phases of the experiment. As can be seen in Figure 7A, percentage accuracy increased from around 50%

at the beginning of the learning phase to roughly 100%. Accuracy was largely maintained in Phase 2 (see Figure 7B). Average number of blocks to mastery (defined as receiving a score of 95% or better over a block of trials) was 5.6 (*SD* = 4.5); see Figure 8 for a per-participant breakdown of the number of blocks of training and testing, as well as the number of blocks to mastery.

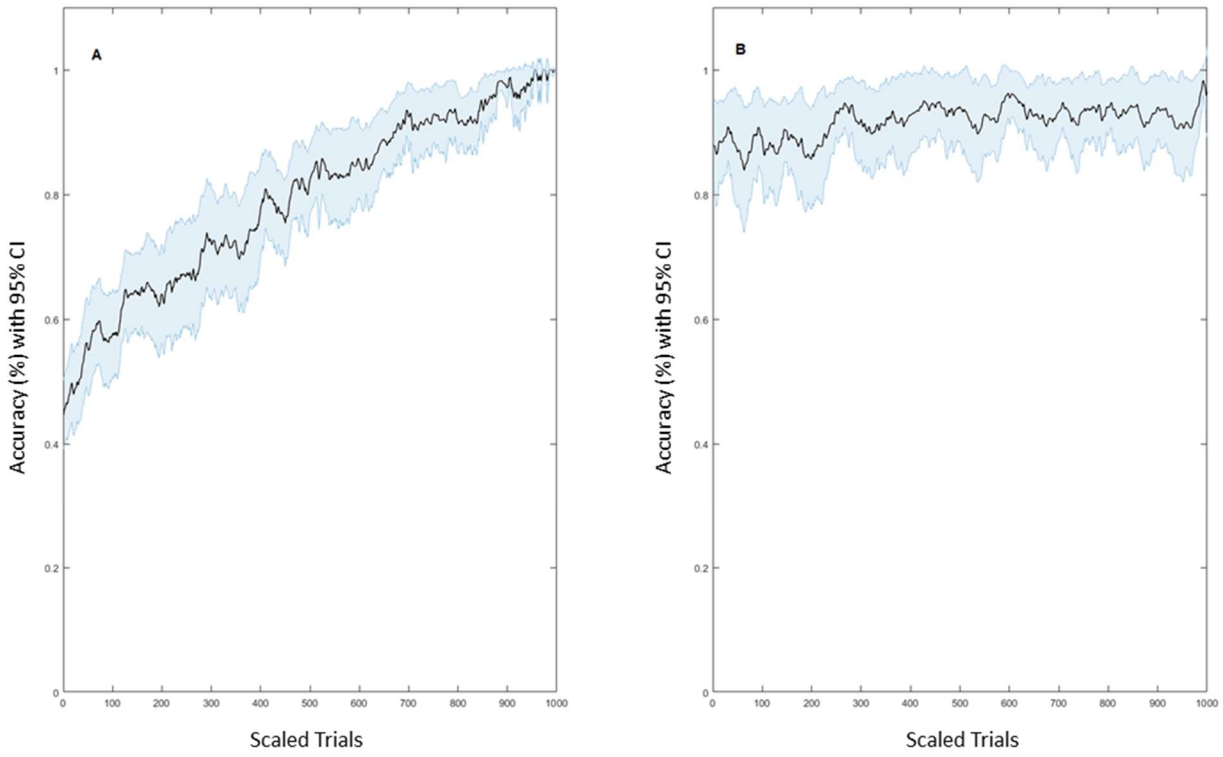

Figure 7. Response Accuracy in Phases One and Two. A: Phase 1 accuracy of responses until mastery achieved, scaled to 1000 trials. B: Phase 2 accuracy, scaled to 1000 trials. Shading indicates 95% CIs.

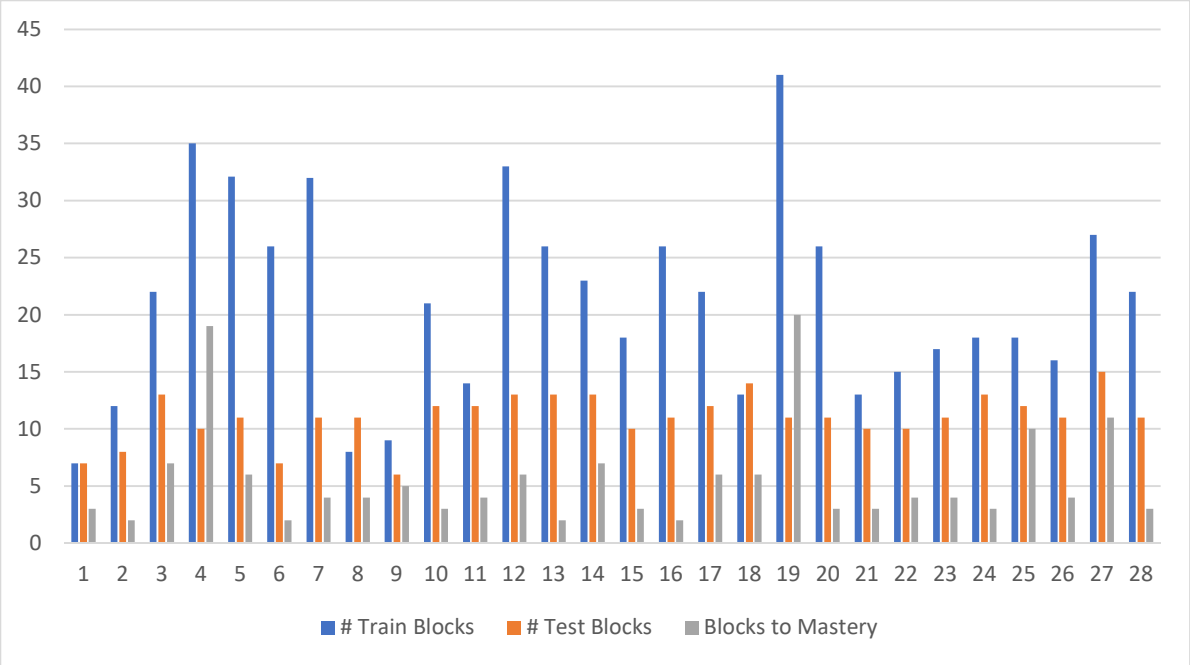

Figure 8. Number of Training Blocks, Test Blocks, and Blocks to Mastery by Participant. Note: participants generally completed more training blocks than required to achieve mastery, as they were instructed to continue until capping was complete.

2.3.2 Electroencephalogram Data.

Reward Positivity. Analysis of the ERP waveforms[3] for the reward positivity revealed an interaction between Valence and Expectancy, $F(1,27) = 6.681$, $p = 0.015$. Post-hoc decomposition of this interaction showed a difference between Good-Expected and Good-Unexpected events ($t = 6.29$, $p < 0.001$, Cohen's $d = 1.25$), as well as a difference between Bad-Expected and Bad-Unexpected events ($t = 3.19$, $p < 0.01$, Cohen's $d = 0.65$). Additionally, there was a difference between Expected and Unexpected events, $F(1,27) = 39.851$, $p < 0.001$. No difference was found between Good and Bad events, which is expected given that this data was recorded during a learned task, and as such, reflects the predictions of RL theory: in a learned

[3] Please note that all ERP results presented herein are plotted with positive-up, as recommended by Steven J. Luck (Luck, 2014).

task, the valence of the outcome does not affect the prediction error; only a violation of expectancy would do so (R. S. Sutton & Barto, 1998).

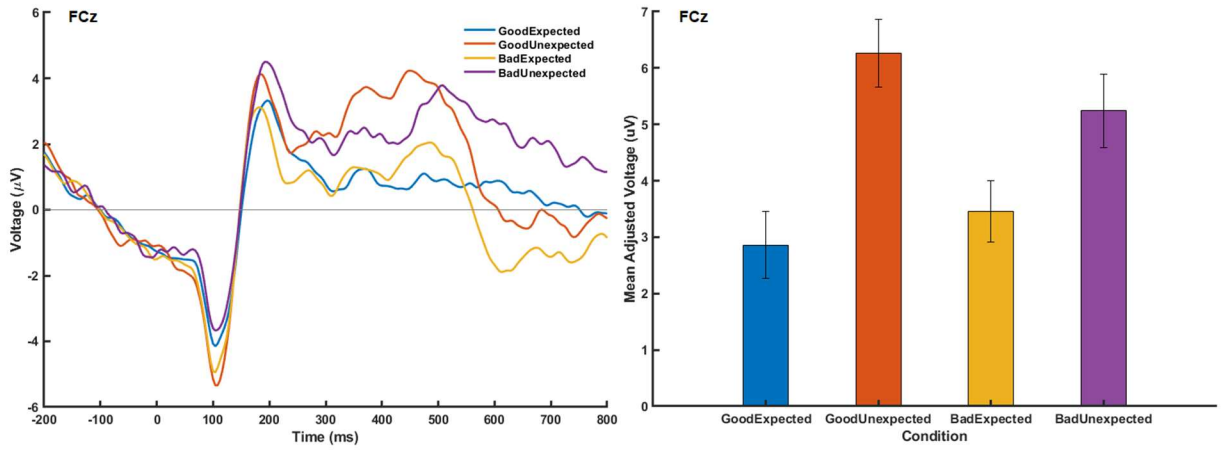

Figure 9. Left: ERPs at FCz for Good-Expected, Good-Unexpected, Bad-Expected, and Bad-Unexpected conditions. Right: peak mean adjusted voltages for each of those conditions with 95% CIs (between subjects). Means were taken over the window of ±20 ms centered on the peak voltage that occurred within the time range of 300 – 500 ms post-stimulus. Adjusted means were calculated to ensure individual variability in peak voltages did not skew the results.

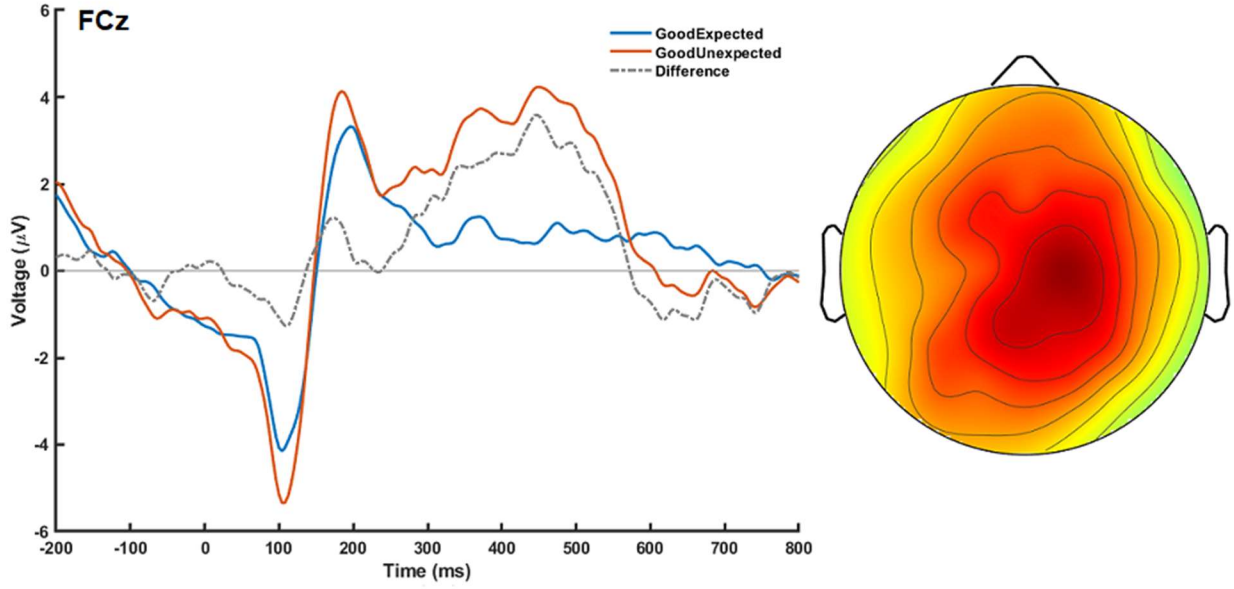

Figure 10. Left: ERPs and difference wave at FCz for Good-Expected versus Good-Unexpected conditions. Right: Scalp topography.

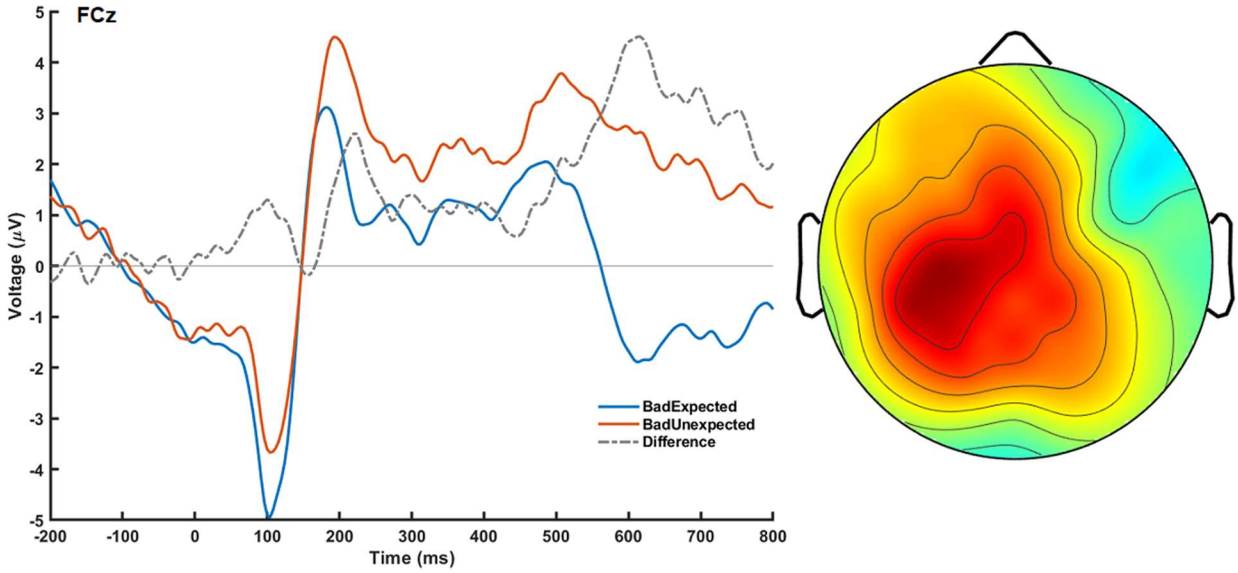

Figure 11. Left: ERPs and difference at FCz for Bad-Expected versus Bad-Unexpected conditions. Right: Scalp topography.

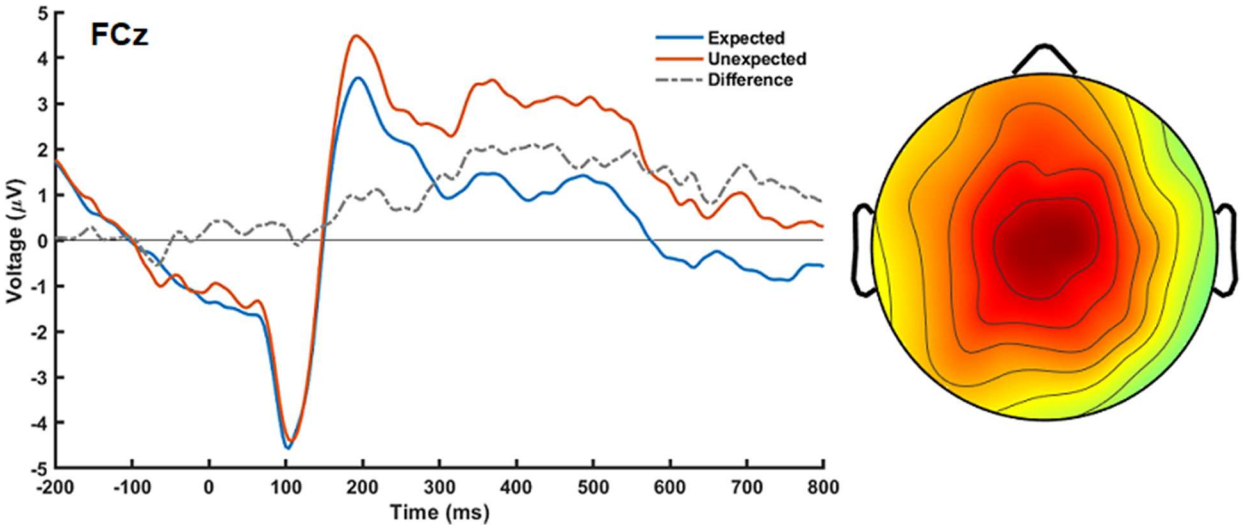

Figure 12. ERPs and difference wave at FCz for Expected versus Unexpected conditions. Right: Scalp topography.

P300. Analysis of the ERPs for the P300 revealed an interaction between valence and expectancy, $F(1,27) = 25.634$, $p < 0.001$. Post-hoc decomposition revealed a difference between Good-Expected and Good-Unexpected events ($t = 10.06$, $p < 0.001$, Cohen's $d = 1.90$), as well as

Good-Expected and Bad-Expected events ($t = 6.28$, $p < 0.001$, Cohen's $d = 0.35$). Additionally, there was a difference between Expected and Unexpected events, $F(1,27) = 42.116$, $p < 0.001$. As with the reward positivity, no difference was found between Good and Bad events.

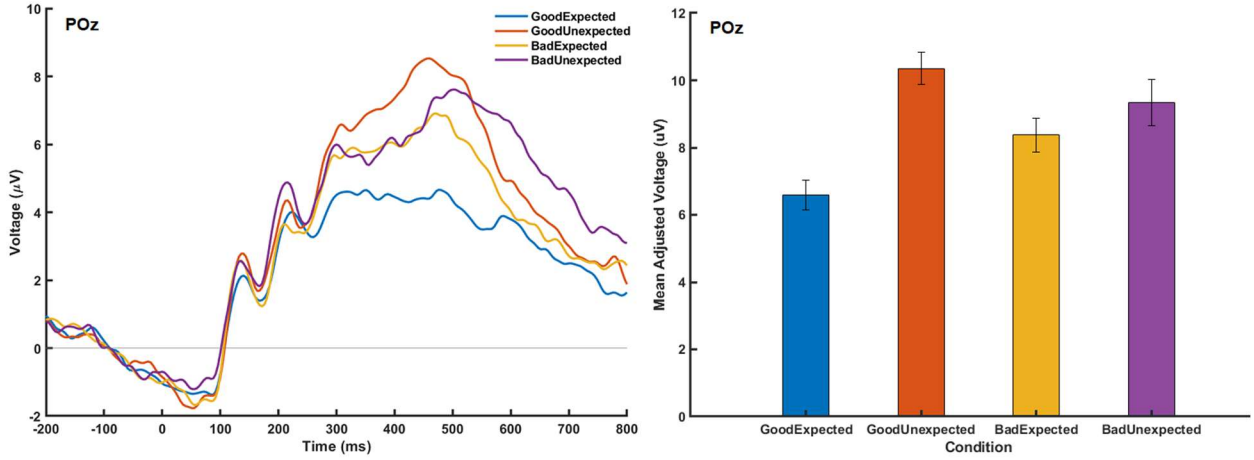

Figure 13. Left: ERPs at POz for Good-Expected, Good-Unexpected, Bad-Expected, and Bad-Unexpected conditions. Right: peak mean adjusted voltages for each of those conditions with 95% CIs (between subjects). Means were taken over the window of ±20 ms centered on the peak voltage that occurred within the time range of 300 – 600 ms post-stimulus. Adjusted means were calculated to ensure individual variability in peak voltages did not skew the results.

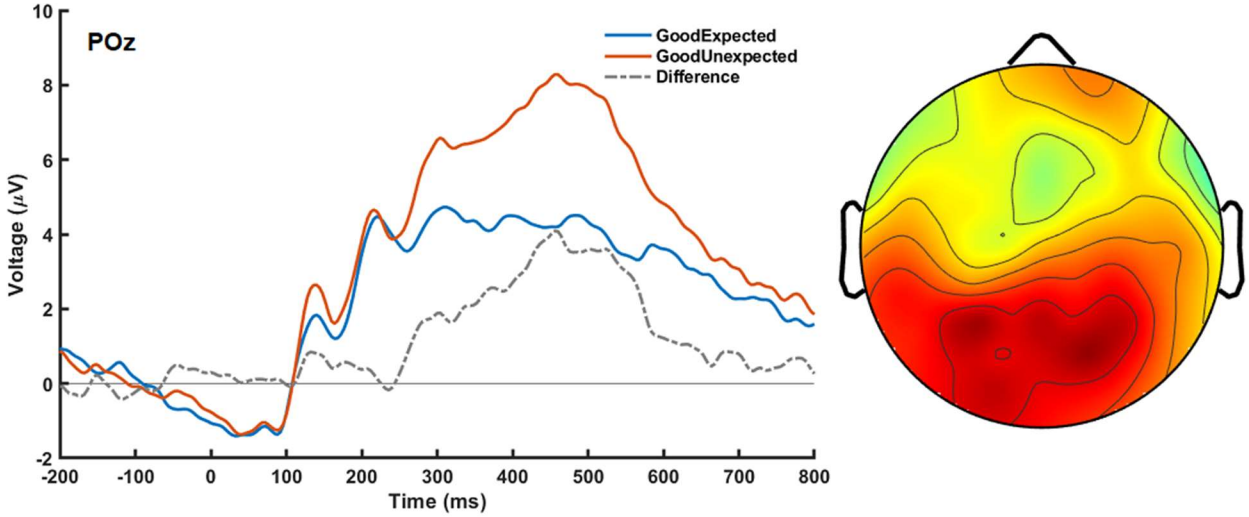

Figure 14. Left: ERPs and difference wave at POz for Good-Expected versus Good-Unexpected conditions. Right: Scalp topography.

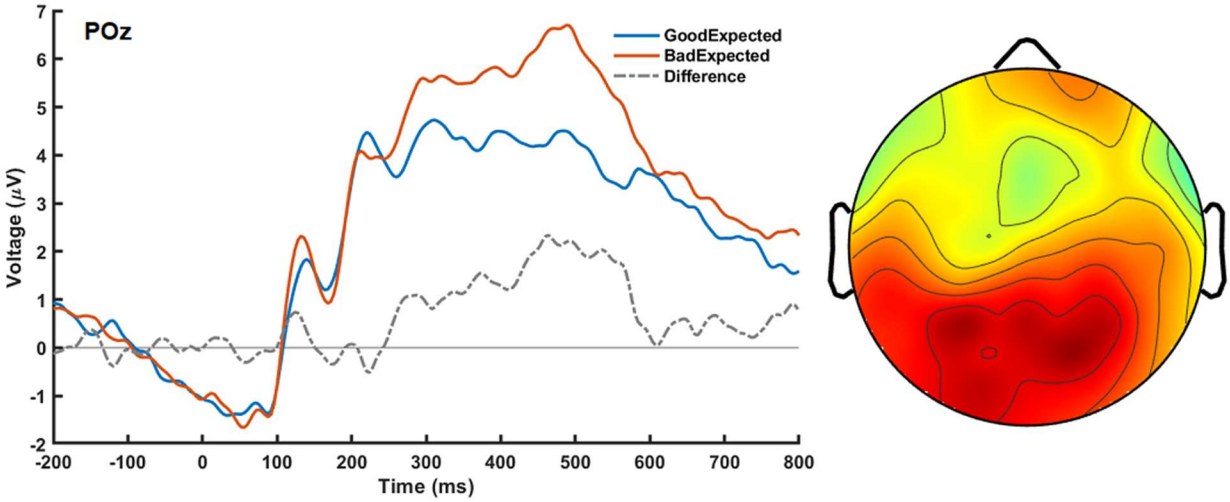

Figure 15. Left: ERPs and difference wave at POz for Good-Expected versus Bad-Expected. Right: Scalp topography.

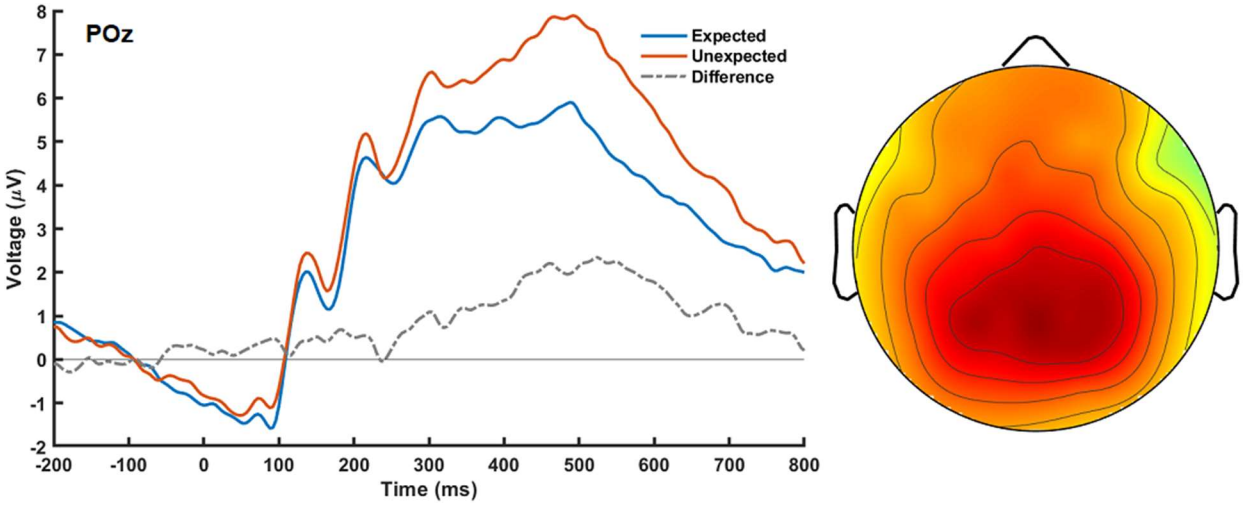

Figure 16. Left: ERPs and difference wave at POz for Expected versus Unexpected conditions. Right: Scalp topography.

2.4 Discussion and Summary

Results to date largely support the theory that reinforcement learning algorithms are implemented within the human brain to assist with learning and decision-making for goal-directed tasks (e.g. Holroyd & Coles, 2002; Holroyd et al., 2011, 2008; Proudfit, 2015). RL theory posits that, when the outcome of an action results in a better than expected reward, a reward prediction error is generated, which is used by the brain to update the value of the action, and increase the chance that that action will be taken in the future (R. S. Sutton & Barto, 1998). RL theory also predicts that, in a learned task, expected outcomes do not generate a reward prediction error, which has been borne out in many human neuroimaging experiments (e.g. Matsumoto et al., 2007; Schultz et al., 1997, 1995). In the experiment described in this chapter, I examined the effects of expected and unexpected events that could be either 'good' or 'bad' for the participants on the reward positivity and P300 components of ERPs.

For the reward positivity, I found an interaction between valence and expectancy, with post-hoc analysis finding significant differences between Good-Expected and Good-Unexpected events as well as between Bad-Expected and Bad-Unexpected events. I also found a difference between Expected and Unexpected events, regardless of valence. These results are in line with RL theory—in a learned task, unexpected outcomes will generate a larger reward prediction error than when the outcome is as expected (R. S. Sutton & Barto, 1998). Also in line with RL theory was the null finding for valence alone: in a learned task, there are no reward prediction errors generated for the expected outcome, whether that outcome is beneficial or harmful in pursuit of the goal (Krigolson et al., 2014; R. S. Sutton & Barto, 1998). However, the finding that Bad-Unexpected was greater than Bad-Expected as well as Good-Expected was not in agreement with

previous findings (Proudfit, 2015). One explanation for this could be the lower number of trials for the Bad-Unexpected condition; this discrepancy could be the results of sample sizes: the number of trials in the Bad-Expected case is much larger than for the Bad-Unexpected case, and thus any outliers will affect the grand average ERP less in the former condition than in the latter (Luck, 2014). It is also possible that this is a manifestation of the expectancy-deviation hypothesis, where an optimistic bias for one's actions being correct causes frequent prediction errors when negative feedback is received (Miller & Ross, 1975; Oliveira, McDonald, & Goodman, 2008).

An important note on these results is that the latency of the reward positivity in this experiment was much later than in simpler tasks such as two-armed bandits. The time window containing the positive deflection assumed to be the reward positivity in this current experiment appeared somewhat later (between 300 to 500 ms onset post-stimulus) than in much of the prior literature regarding this ERP component (Proudfit, 2015). I hypothesize that this is due to the complexity of the task and the feedback; participants must remember that a certain icon indicates a specific state—for instance, the apple icon represents State 4—and then determine if that is better or worse than the previous state; late presentation of the reward positivity due to increased cognitive load has been reported in prior work (e.g. Holroyd & Coles, 2002; Holroyd et al., 2005; Krigolson et al., 2015, 2012).

Prior research on the P300 has linked its amplitude to stimulus probabilities and attentional resources devoted to the task at hand, among other factors that are less relevant to the task performed in this experiment (Polich, 2007). Thus, I examined the effects of valence and expectancy upon the P300 ERP component in this experiment, finding an interaction between the two factors, with post-hoc analysis finding differences between each of Good-Expected versus

Good-Unexpected and Good-Expected versus Bad-Expected. As with the reward positivity, I also found a significant difference in the amplitude of the P300 between Expected and Unexpected events, as would be expected given the strong linkage in prior work between the P300 and stimulus probabilities (Polich, 2007). Additionally, the latency of the P300 observed in this experiment occurred later than is usually reported; prior studies have reported that the P300 latency reflects stimulus evaluation time (Duncan-Johnson & Donchin, 1982). It may also indicate additional processing steps necessitated by the complex task and stimuli used in this experiment, as suggested in Krigolson et al.'s 2012 paper. Latency of the P300 is affected by individual cognitive capabilities, genetics, and age (Polich, 2007). Thus, the results found in this analysis support and agree with previous research into and characterizations of the P300 ERP component.

In terms of the research question posed at the outset of this experiment—what do the neural responses to intermediate steps in a multi-step, goal directed task look like in comparison to single-step tasks, and do these responses line up with RL theory predictions—it appears that, at least in this experiment, both the reward positivity and the P300 components respond to intermediate steps of a task in much the same way as to single step tasks, and these results support the idea that the brain uses RL to learn and make decisions throughout multi-step activities. Unexpectedly, however, the difference between Bad-Unexpected and Bad-Expected events diverges from the findings of other studies, where Bad-Unexpected events did not cause an increase in the amplitude of the reward positivity.

Chapter 3: Experiment Two – EEG Data Classification Using Support Vector Machines

3.1 Introduction

While ERP analyses of EEG data excel at finding differences between the means of waveforms resulting from averaging many trials together, they cannot interrogate certain aspects of the data, such as trial-to-trial variability in amplitude or latency of signals that would be lost during averaging or where patterns present across multiple electrodes but do not show a significant difference on any one electrode (Pernet, Sajda, & Rousselet, 2011). Machine learning analysis methods, on the other hand, provide the solutions for analyzing variability, as well as several other problems: they can provide mapping between neural activity measured using one neuroimaging technique (e.g. EEG, MEG, or fMRI) and stimuli, behavioural data, or neural activity measured using another neuroimaging technique (deBettencourt, Goldman, Brown, & Sajda, 2011; Goldman et al., 2009; Ratcliff, Philiastides, & Sajda, 2009; Rousselet & Pernet, 2011; Schyns, 2010). In recent years, support vector machines (SVMs) have shown very good performance on a variety of classification problems, especially ones where there are many more features than data samples, which is common for EEG data (e.g. Bashashati et al., 2007; Blankertz, Curio, & Müller, 2002; Blankertz et al., 2011; Burges, 1998; Cortes & Vapnik, 1995; Kumar et al., 2014; Lotte et al., 2007; Nicolaou & Georgiou, 2012; Quitadamo et al., 2017; Tan, Steinbach, Karpatne, & Kumar, 2019).

In this experiment, I applied several variations on Support Vector Machine classifiers to the EEG data that resulted from Experiment One. The overarching goal was to determine if it was possible to automatically classify neural responses to intermediate states in a goal-directed activity as predicted by current theories regarding learning and decision-making in humans. Thus, the questions I set out to answer with the work described in the remainder of this chapter

are: first, can we differentiate between states in EEG data? That is, given an epoch of EEG data with the label withheld, can a classifier determine what state the participant just entered? Second, given an epoch of data that was not included in the training set, with the label withheld, can the SVM accurately predict which condition it belongs to (from Chapter 2; e.g. Good-Unexpected versus Good-Expected)? Finally, can we replicate the findings in Chapter 2 regarding differences between conditions with a machine learning analysis?

The first hypothesis I made for the work detailed in the following sections of this chapter was that, given epoched EEG data from each of the seven states, after training an SVM, it would be able to classify a heretofore unseen data sample with the label withheld according to what state the participant was in. For instance, given an epoch of EEG data time-locked to the instant a participant saw the cue for State 4, with the label withheld for testing, the SVM will be able to identify the sample as from State 4. Thus, the SVM would be able to determine which state the participants were in based solely on their neural activity, answering one of the main questions I set out in the beginning of this work. I also predicted that I could replicate the findings from the ERP analyses in Chapter 2, finding statistically significant differences in the epoched EEG data between the pairs of conditions analyzed earlier using standard ERP analysis techniques (e.g. Good-Unexpected and Good Expected, Good-Unexpected and Bad-Unexpected, etc.; see Section 2.2.4 for further details).

3.2 Methods

3.2.1 Preprocessing.

The raw, continuous EEG data was preprocessed before classification for efficiency of analysis. Data was first downsampled to 250 Hz (resulting in 200 data points for each of the 62 channels for each 800 ms-long epoch). Bad channels (designated as such by visual inspection of

the raw EEG data) were then removed and interpolated using spherical splines from surrounding channels. Channels marked as bad typically are either flat (i.e., no signal), contain large voltage spikes, an overabundance of electrical noise because of a faulty electrode or bad connection to the amplifier, or other artifacts that introduce an overwhelmingly large-amplitude signal compared to normal ranges for EEG data, and cause more than 10% of trials to be removed during artifact rejection. Next, I filtered the data using an Infinite Impulse Response (IIR) zero-phase-shift Butterworth filter, with a passband of 0.1 – 30 Hz (a standard range that preserves all the ERP components of interest) and a notch filter at 60 Hz to remove electrical line noise (Luck, 2014). All preprocessing up to this point was performed in BrainVision Analyzer. At this point, data was exported to comma-separated-values text files for further processing and analysis using Python 3 and the NumPy, Pandas, and Scikit-learn libraries (McKinney, 2010; Pedregosa et al., 2011; Walt, Colbert, & Varoquaux, 2011). Using these technologies, the continuous EEG data was segmented into 800 ms epochs around markers of interest, from -200 ms before the marker to +600 ms after it. Subsequently, for each EEG event marker (see Table *4* for a list of the markers in the EEG data), 5 trials across all participants were selected at a time without replacement and averaged together. When less than 5 trials remained for a given marker, those were averaged together. This was repeated for all markers listed in Table *4*. For a summary of preprocessing steps, see Table *5*.

Table 4

EEG Experiment Markers and Labels

Marker #	**Previous State**	**Current State**	**Who Just Moved**	**Comment**	**Good/Bad**	**Expected/Unexpected**
11	2	1	Opponent	Loss	Bad	Expected
12	2	1	Participant	Loss	Bad	Unexpected
20	N/A	2	N/A	Start	N/A	N/A

23	3	2	Opponent		Bad	Expected
24	3	2	Participant		Bad	Unexpected
30	N/A	3	N/A	Start	N/A	N/A
31	2	3	Opponent		Good	Unexpected
32	2	3	Participant		Good	Expected
33	4	3	Opponent		Bad	Expected
34	4	3	Participant		Bad	Unexpected
40	N/A	4	N/A	Start	N/A	N/A
41	3	4	Opponent		Good	Unexpected
42	3	4	Participant		Good	Expected
43	5	4	Opponent		Bad	Expected
44	5	4	Participant		Bad	Unexpected
50	N/A	5	N/A	Start	N/A	N/A
51	4	5	Opponent		Good	Unexpected
52	4	5	Participant		Good	Expected
53	6	5	Opponent		Bad	Expected
54	6	5	Participant		Bad	Unexpected
60	N/A	6	N/A	Start	N/A	N/A
61	5	6	Opponent		Good	Unexpected
62	5	6	Participant		Good	Expected
71	6	7	Opponent	Win	Good	Unexpected
72	6	7	Participant	Win	Good	Expected

Table 5

Summary of Preprocessing Steps for Creating Input Data to Classifiers

Step	**Operation**
1	Downsample raw, continuous EEG data from 500 Hz to 250 Hz
2	Remove bad channels (flat, excessive artifacts from electrical noise)
3	Interpolate removed channels using spherical splines
4	Filter EEG data using an Infinite Impulse Response (IIR) zero-phase-shift Butterworth filter, with a passband of 0.1 – 30 Hz and a notch filter at 60 Hz
5	Export continuous EEG from BrainVision Analyzer to comma-separated-value text files by marker (one file per marker, containing all trials across all participants)
6	Segment continuous EEG around each marker from -200 ms to + 600 ms
7	Average trials: for each marker, 5 trials across all participants were selected at a time without replacement and averaged together. When less than 5 trials remained for a given marker, those were averaged together.
8	Create datasets for each condition by combining the averaged trials for each marker contained in that condition (see Table *6*)

To create the datasets for each condition of interest for classification (see Table *6*), datasets for each of the EEG markers (as previously produced in the trial-averaging step) were combined into condition datasets; for example, to create the condition dataset "Good-Unexpected", the averaged datasets for markers 31, 41, 51, 61, and 71 were combined. The resulting condition datasets were then used as inputs to the classifier. Thus, the EEG markers (see Table *4* for a complete list) were collapsed in various ways for each of the classification tasks. The most important categories were:

- Events corresponding to state changes (states 1 through 7);
- "Good" events;
- "Bad" events;
- Expected events; and
- Unexpected events.

The valence ("good" versus "bad") are labelled from the participants' point of view; i.e. a good event is one that puts them closer to winning whereas a bad event is one that puts them closer to losing, regardless of whether it was the computer's action or their own action that caused the transition into that state.

Table 6

Groupings of Markers Used in Classification.

Label	**Included Markers**	**Number of Samples**
Good	31, 32, 41, 42, 51, 52, 61, 62, 71, 72	2646
Bad	11, 12, 23, 24, 33, 34, 43, 44, 53, 54	2426
Expected	11, 23, 32, 33, 42, 43, 52, 53, 62, 72	3569
Unexpected	12, 24, 31, 34, 41, 44, 51, 54, 61, 71	1503
Good Expected	32, 42, 52, 62, 72	1862
Good Unexpected	31, 41, 51, 61, 71	784
Bad Expected	11, 23, 33, 43, 53	1707

Bad Unexpected	12, 24, 34, 44, 54	719
State 1	11, 12	269
State 2	23, 24	594
State 3	31, 32, 33, 34	1028
State 4	41, 42, 43, 44	1161
State 5	51, 52, 53, 54	1067
State 6	61, 62	585
State 7	71, 72	368

Finally, the grouped sets were combined in the following manner to create the labelled datasets for classification:

- Good versus Bad
- Expected versus Unexpected
- Good Expected versus Good Unexpected
- Bad Expected versus Bad Unexpected
- Good Expected versus Bad Expected
- Good Unexpected versus Bad Unexpected
- State 1 versus State 2 versus State 3 versus State 4 versus State 5 versus State 6 versus State 7 (multiclass problem)

3.2.2 Classification.

Using the sklearn.svm.SVC class in the Python library Scikit-learn, models were built for each classification task. Two SVM kernels—one linear, one non-linear—were tested for each classification task, using a grid search to find the most accurate combination of kernel and regularization parameter for each dataset. Subsequently, 10-fold cross-validation was used to ensure reliability of the results. Using ten folds for cross-validation allows the use of the vast majority (90%) of overall data for training, and thus the expected performance of the resulting model should be closer to the performance of a model trained on 100% of the data than a model

trained with fewer folds (Tan et al., 2019). After running the classifier on these combinations, to test whether the results differed from random categorization, I ran permutation tests for each one. This involved randomizing the labels for all trials in a dataset and then running the grid search over the defined parameter space as in the classification step. Permutation testing provides a distribution of accuracies under the null hypothesis which can then be used to determine if we can reject that our measured accuracy was drawn from the null hypothesis distribution (Wasserman, 2013).

3.3 Results

As can be seen in Table *7*, several classifications yielded accuracy higher than chance; these included the Good versus Bad dataset with 61% accuracy, Good-Expected versus Bad-Expected with 59% accuracy, and Good-Unexpected versus Bad-Unexpected with 62% accuracy. Most importantly, the multiclass classification of individual states (states 1 through 7) yielded 37% accuracy. With all of these classifications, permutation testing gave a result equal to the sample split percentage for each category; note that the accuracy of permutation testing for the individual state classification was equal to the number of samples in the largest category at 23%. Thus, for these four tasks, the classifier was able to determine a separation between the groups. For all other tasks (Expected versus Unexpected, Good Expected versus Good Unexpected, and Bad Expected versus Bad Unexpected) the classifier was unable to achieve better than chance accuracy, with accuracy nearly equal to the number of samples in each category, in line with the permutation testing results.

Table 7

Accuracy of Classification for Each Comparison

Categories	**Kernel and Parameters**	**Accuracy**	**Sample Split Percentage**	**Permutation Test Results**
Good versus Bad	RBF (C = 1, gamma = auto)	0.61	52/48	*M* = 0.52, *SD* = 0.002
Expected versus Unexpected	RBF (C = 1, gamma = auto)	0.70	70/30	*M* = 0.70, *SD* = 0.00
Good Expected versus Good Unexpected	RBF (C = 1, gamma = auto)	0.70	70/30	*M* = 0.70, *SD* = 0.00
Bad Expected versus Bad Unexpected	RBF (C = 10, gamma = auto)	0.73	70/30	*M* = 0.70, *SD* = 0.00
Good Expected versus Bad Expected	Linear (C=0.1)	0.59	52/48	*M* = 0.52, *SD* = 0.0035
Good Unexpected versus Bad Unexpected	RBF (C = 10, gamma = auto)	0.62	52/48	*M* = 0.52, *SD* = 0.007
7-States (State 1 versus State 2 versus State 3 versus State 4 versus State 5 versus State 6 versus State 7)	RBF (C = 10, gamma = auto)	0.37	5/12/20/23 /21/12/07	*M* = 0.2303, *SD* = 0.0017

Note: M = Mean accuracy of model with permuted labels over 100 runs, SD = Standard Deviation.

3.4 Discussion and Summary

In summary, for the experiment described in this chapter, I created labelled datasets from the EEG data recorded in Experiment 1, using the event-locked markers to epoch the data and categorize it according to the type of event the data resulted from. I then used a machine learning classification algorithm to ascertain whether the epoched EEG could be categorized based on neural signals within the data. Four comparisons were classified with above chance accuracy, confirmed through permutation testing: Good versus Bad, Good-Expected versus Bad-Expected, and Good-Unexpected versus Bad-Unexpected, and States 1 through 7. The last is the most important, as it answers in part the question of whether we could determine which state a participant was in from their neural data only. This result means that it is indeed possible to deduce which state a participant was in using neural signals in the minimally preprocessed epoched EEG data. The three other comparisons that were statistically significant (Good versus Bad, Good-Expected versus Bad-Expected, and Good-Unexpected versus Bad-Unexpected) showed that the classifier was able to detect differences in the neural signals between these conditions without averaging across all trials as in the ERP analysis done in Chapter 2. However, the classification algorithm could not differentiate between two conditions in three other tasks: Expected versus Unexpected, Good Expected versus Good Unexpected, and Bad Expected versus Bad Unexpected.

For the latter tasks, there are several reasons why the algorithm was unable to find a solution, including inherent noise present in EEG, the high dimensionality of the data, or a comparatively small signal that was unique to each condition (Blankertz et al., 2011; Lotte et al., 2007). It is also possible that the averaging process (especially across subjects) may have

reduced or removed differences between each of the conditions, thus rendering it impossible for the SVM to find a solution (Luck, 2014; Vapnik, 1998). However, given that these three comparisons had the widest disparity in number of samples per category, and the fact that SVMs (along with many other ML methods) are susceptible to bias when given unbalanced datasets, it is possible that the unbalanced classes were the main cause of failure in these instances (e.g. Japkowicz & Stephen, 2002; Tan et al., 2019).

Chapter 4: Considerations and Discussion

4.1 Implications

The experiments described in this thesis aimed to answer two outstanding questions about the neural responses to intermediate states in a learned, multi-state, goal-directed task, specifically whether those responses are accurately predicted by reinforcement learning theory, and whether we can ascertain from the neural responses which specific state the participants were currently in. Experiment One investigated whether the reward positivity behaved as predicted by reinforcement learning theory during a multi-state task involving outcomes that varied in valence and expectancy. In RL theory, each intermediate state should acquire a value depending on how likely a reward is; thus, states closer to the win state should have a higher value than states further away from it. I wanted to see if those state values were encoded in the neural signals and therefore detectable at the scalp during the task. To that end, I performed an EEG experiment designed to elicit the reward positivity and P300 ERP components where participants performed a linear, random-walk-style task with seven discrete states, including both a defined win state and loss state. In Experiment Two, the goal was to determine if machine classification could be used to differentiate between state values conveyed in the EEG data, as well as to determine if a classifier could replicate the findings of the ERP analysis conducted in Experiment One. Thus, using the EEG data gathered in the first experiment, I trained Support Vector Machines to classify the epoched data into conditions and individual states.

4.1.1 Experiment 1.

In the first experiment, I performed an ERP analysis on epoched EEG data time-locked to state transitions, and examined differences in the reward positivity and the P300 between several combined conditions. For the reward positivity, I found a significant interaction between valence

and expectancy, with post-hoc statistical analysis showing significant differences in the amplitude of the reward positivity voltage when comparing Good-Expected versus Good-Unexpected conditions as well as Bad-Expected versus Bad-Unexpected conditions. I also found a difference in Expected versus Unexpected conditions. Reinforcement learning theory postulates that reward prediction errors will only be generated in a learned task (where the values for each state and action closely approximate the real values) when the outcome of an action is different than expected (R. S. Sutton & Barto, 1998). Therefore, my results for the reward positivity ERP analysis align with reinforcement learning theory: only outcomes that differ from expectations cause a reward prediction error (R. S. Sutton & Barto, 1998). However, the fact that Bad-Unexpected events elicited a larger amplitude reward positivity than either Bad-Expected or Good-Expected events is contrary to many prior findings in this area (e.g. Holroyd et al., 2008; Krigolson et al., 2014). Put another way, I found that unexpectedly positive outcomes had a higher amplitude reward positivity than expected outcomes, whether positive or negative, and unexpectedly negative outcomes generated a higher amplitude reward positivity than expected outcomes of either valence. While this does not align with much of the work done on RL in the brain, some researchers have observed a similar effect, where negative feedback elicits the reward positivity, and attributed it to an optimism bias where humans tend to evaluate their own performance overly optimistically, and thus generate a larger reward prediction error upon receiving the negative feedback (Miller & Ross, 1975; Oliveira et al., 2008).

In regards to the P300, my findings here were similar to the reward positivity: there was an interaction between valence and expectancy, and post-hoc investigation showed differences between Good-Expected versus Good-Unexpected and Good Expected versus Bad-Expected conditions. There was, again, a difference between the larger grouping of conditions into

Expected and Unexpected events. As the P300 amplitude is thought to be driven by processes responsible for stimulus probability and attentional resourcing, these results agree with prior work: the unexpected stimuli have a lower probability of appearing than the expected ones and thus should generate a larger P300 (Duncan-Johnson & Donchin, 1977; Johnson Jr. & Donchin, 1982; Polich, 2007; Squires et al., 1976). It is also possible, given previously published literature on the P300, that unexpected stimuli cause more attentional resources to be devoted to the task, possibly in support of context updating processes, thus eliciting an increase in the observed amplitude of the P300 (e.g. Isreal et al., 1980; Kramer et al., 1985; Wickens et al., 1983).

The results of this experiment support existing theories regarding the presentation of the reward positivity and the P300 in response to events with varying valence and expectancy combinations. Importantly, the results of the ERP reward positivity analysis suggest that reinforcement learning theory correctly predicts neural responses to intermediate events in a multi-step process. However, this analysis did not fully answer the research question I set out at the beginning of this work: can we differentiate between individual states based solely on neural responses as recorded in EEG? This is due to the limitations of ERPs: because many trials are averaged together, individual differences are lost, especially when averaging across participants, as latency and amplitude of components including the reward positivity (Bress et al., 2015; Colino et al., 2017; Proudfit, 2015) and the P300 are highly variable between persons (Polich, 2007). In this way, the creation of ERPs may erase neural signals present in the original data, especially if those signals are small or have variability in latency (Luck, 2014).

4.1.2 Experiment 2.

Given the above limitations of ERP analysis, I performed a method of machine classification analysis on the EEG data using a Support Vector Machine. Machine classification

analysis is able to answer questions about EEG data that ERP analyses cannot, such as how individual variability of latency or amplitude in neural signals is contributing to averaged results (Ford et al., 1994). For the work described in Chapter 3, I was interested in whether epoched EEG around the presentation of state cues could be accurately identified according to the specific state it was elicited by. Based on prior work (e.g. Bashashati et al., 2007; Blankertz et al., 2011; Lotte et al., 2007; Quitadamo et al., 2017), I chose to use Support Vector Machine algorithms due to their robustness, ability to handle high-dimensional data, and reasonable accuracy with low signal-to-noise-ratio data (Quitadamo et al., 2017; Vapnik, 1998). Epoched EEG data from Experiment 1 was split into training and test groups for each classification attempt, a grid search algorithm was employed to find the best parameters for each set of conditions, and 10-fold cross-validation was used to ensure reliability of the results. Finally, permutation testing was used to establish a base rate for classification with permuted labels such that we could determine if the null hypothesis could be rejected for the measured accuracy (Wasserman, 2013).

Using the same comparisons as in the ERP analysis, I found that an SVM could classify, with above-chance accuracy, Good versus Bad, Good-Expected versus Bad-Expected, and Good-Unexpected versus Bad-Unexpected. It is interesting to note that while the ERP analysis was unable to differentiate on valence alone, the classifier could (see Table *8*). This may indicate that valence is encoded in the EEG signal, but not within the reward positivity or the P300—or that the averaging of all trials during the process of creating ERPs removes the signal.

Most importantly, though, the classifier was able to detect differences between the seven states with above-chance accuracy. Thus, I conclude that individual states are detectable in neural signals recorded in EEG, and that there is evidence of the representation of the states in the neural signals. Of note, each analysis (reward positivity, P300, SVM) found differences between

some pairs of conditions but not others (see Table 8), but between the three analyses, each set of conditions was able to be differentiated at least once.

Table 8

Significant Findings By Method of Analysis: ERP analysis of reward positivity and P300, and SVM classifier.

Type of Analysis		**GE vs. GU**	**GE vs. BE**	**GU vs. BU**	**BE vs. BU**	**E vs. U**	**G vs. B**	**7 States**
ERP Analysis	RewP	X			X	X		
	P300	X	X			X		
Machine Learning	SVM		X	X			X	X

Note: GE = Good-Expected, GU = Good-Unexpected, BE = Bad-Expected, BU = Bad-Unexpected, G = Good, B = Bad, X = significant result.

4.2 Limitations and Future Directions

With regards to the EEG experiment described in Chapter 2, there are a few experimental/methodological points that bear consideration. First, the feedback stimuli (colored icons) were complex, and the seven-state concept (which was never explained to the participant, so they had to form this model on their own) increased the cognitive load required to understand and complete the task. The complexity added to the task by these two factors caused increased latency of the reward positivity, as discussed in section 2.3.2, and may have caused jitter or a "smearing" of the EEG signal across the time window (Luck, 2014).Thus, differences between the conditions may have disappeared when the ERPs were averaged together, and thus removed a detectable signal (Luck, 2014). Using simpler visual stimuli in any future attempt to reproduce these results is possibly warranted.

There are multiple avenues of investigation that could be explored using the results of the two experiments described in this thesis. It would have been instructive to record EEG during the

learning phase in order to model the neural responses as learning occurred. Modelling the neural signal during this type of multi-step task would add to existing research on prediction errors in the brain and may be highly instructive in determining how well RL models represent the neural activity of human brains while learning a complex task. An additional area for future work would be modifying the task so that state visits were more evenly distributed. In Experiment 1, due to the task design, participants spent much of their time in the middle three states. This resulted in many more trials from these states, compared to the outer states, especially the win/loss states. Changing the experimental design to prevent this type of "tug o' war" might provide positive results. One way this could be implemented is by having the unexpected events move the participant or computer opponent by either one or two states in a direction, thus moving towards the end states more quickly, and causing more visits to the outer states. The uneven distribution of visits across the seven states had implications for Experiment 2 as well; this is discussed in detail below.

For Experiment 2, several limitations must be acknowledged, starting with the decision to average a small number of trials together from each condition. While this was done for pragmatic reasons—compute time for SVMs is at least quadratic in relation to the number of samples—it does mean that smaller signals, or ones with jitter, may have been lost during the averaging process, much as in the ERP analysis (Tan et al., 2019). With additional computing resources, the analysis could be run again with no averaging, and may return differing results. Another limitation in this work was the unequal number of trials across conditions; for instance, the unexpected conditions had less than half the number of trials that the expected conditions did, and this was also true for the outer states in the 7-States classification: State 4 had 1161 samples, while State 7 had only 368. SVMs are sensitive to unbalanced classes, thus rerunning the

analysis using only a subset of the states such that there were more similar numbers of samples in each might avoid this problem (Tan et al., 2019).

However, with those limitations in mind, future work should prioritize investigating what the driving factors were for the 7-State classification. Examining the weights assigned to each feature and determining which were the highest is the obvious next step; this might allow dimensionality reduction and thus reduce training time, as well as boosting accuracy of classification. Alternate or additional preprocessing methods to increase signal-to-noise ratio may be worth studying; potential methods include principle component analysis (PCA), wavelet de-noising, or blind source separation techniques (Blankertz et al., 2011). Alternatively, switching to the time-frequency domain may prove fruitful (Bajaj & Pachori, 2013). Finally, other classification algorithms, such as a variant of Linear Discriminant Analysis (LDA) or even a combination of classification algorithms, have shown good results on EEG, and experimenting with one or more of those may provide higher accuracies than obtained in this work (Blankertz et al., 2011).

4.3 Conclusions

Detecting and differentiating neural activity between intermediate states in a task is an important step toward understanding decision making in humans. It is also necessary to determine whether—or how—the brain uses reinforcement learning to successfully interact with the environment. The work described in the preceding chapters examined these questions using two methods of analysis, and provided evidence that RL theory accurately predicts neural activity in a number of conditions. I showed that, when comparing expected and unexpected events, the reward prediction error signal—the reward positivity—is larger for unexpectedly events than for expected ones, and especially for unexpectedly positive events as compared to

negative events. Using a machine learning classifier, I also showed that it was possible to detect which state an epoched section of EEG data resulted from, which indicates that the state representation is encoded in neural activity, and can be detected from scalp-recorded EEG.

Given these results, two avenues of further investigation should be pursued: one, a follow-up EEG experiment to address the limitations of Wizards' Duel as described above; and two, continued analysis using machine learning techniques of the data gathered as part of this work. The logical next steps down this latter path should include balancing the classes as well as applying PCA in order to boost the signal-to-noise ratio in the data.

In summary, there is much we still don't know about how intermediate states are represented in the brain, but this work provides further evidence that reinforcement learning is used within the brain to learn optimal choices and make optimal decisions during complex tasks. It is exciting that it appears to be possible to differentiate the neural activity between states in these tasks, and this may allow us to further non-invasively probe decision-making processes in the brain, as well as have implications for brain-computer interfaces.

References

Alpaydin, E. (2014). *Introduction to machine learning*. MIT press.

Bajaj, V., & Pachori, R. B. (2013). Automatic classification of sleep stages based on the time-frequency image of EEG signals. *Computer Methods and Programs in Biomedicine*, *112*(3), 320–328. https://doi.org/10.1016/j.cmpb.2013.07.006

Barto, A. G. (1995). Adaptive Critics and the Basal Ganglia. In J. C. Houk, J. Davis, & D. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia* (1st ed., Vol. 7, pp. 215–232). https://doi.org/10.1.1.133.269

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-13*(5), 834–846. https://doi.org/10.1109/TSMC.1983.6313077

Bashashati, A., Fatourechi, M., Ward, R. K., & Birch, G. E. (2007). A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. *Journal of Neural Engineering*, *4*(2), R32–R57. https://doi.org/10.1088/1741-2560/4/2/R03

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

Berger, H. (1929). Uber das Elektrenkephalogramm des Menschen [On the use of the encephalogram in humans]. *Arch Psychiatr Nervenkr*, *87*, 527–570.

Blankertz, B., Curio, G., & Müller, K.-R. (2002). Classifying single trial EEG: Towards brain computer interfacing. *Advances in Neural Information Processing Systems*, *1*(c), 157–164. https://doi.org/10.1.1.19.8038

Blankertz, B., Lemm, S., Treder, M., Haufe, S., & Müller, K.-R. (2011). Single-trial analysis and classification of ERP components—A tutorial. *NeuroImage*, *56*(2), 814–825. https://doi.org/10.1016/j.neuroimage.2010.06.048

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Bress, J. N., Meyer, A., & Proudfit, G. H. (2015). The stability of the feedback negativity and its relationship with depression during childhood and adolescence. *Development and Psychopathology*, *27*(4), 1285–1294. https://doi.org/10.1017/S0954579414001400

Burges, C. J. C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167. https://doi.org/10.1023/A:1009715923555

Coles, M. G. H., & Rugg, M. D. (1995). Event–related brain potentials: An introduction. In M. Rugg & M. G. H. Coles (Eds.), *Electrophysiology of Mind* (pp. 1–26). Oxford: Oxford University Press.

Colino, F. L., Howse, H., Norton, A., Trska, R., Pluta, A., Luehr, S. J. C., … Krigolson, O. E. (2017). Older adults display diminished error processing and response in a continuous tracking task. *Psychophysiology*, (March), 1–8. https://doi.org/10.1111/psyp.12907

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The Good, The Bad and The Ugly. *Current Opinion in Neurobiology*, *18*(2), 185–196. https://doi.org/10.1016/j.conb.2008.08.003

deBettencourt, M., Goldman, R., Brown, T., & Sajda, P. (2011). Adaptive Thresholding for Improving Sensitivity in Single-Trial Simultaneous EEG/fMRI. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00091

Dehaene, S., Posner, M. I., & Tucker, D. M. (1994). Localization of a Neural System for Error Detection and Compensation. *Psychological Science*, *5*(5), 303–305.

Dien, J. (2010). Evaluating two-step PCA of ERP data with Geomin, Infomax, Oblimin, Promax, and Varimax rotations. *Psychophysiology*, *47*(1), 170–183. https://doi.org/10.1111/j.1469-8986.2009.00885.x

Donchin, E. (1981). *Surprise!...Surprise?* (Vol. 18). https://doi.org/10.1111/j.1469-8986.1981.tb01815.x

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, *11*(3), 357–374. https://doi.org/10.1017/S0140525X00058027

Donchin, E., & Heffley, E. F. (1978). Multivariate analysis of event-related potential data: A tutorial review. In D. Otto (Ed.), *Multidisciplinary perspectives in event-related brain potential research* (pp. 555–572). Washington, D.C.: U.S. Government Printing Office.

Donchin, E., Ritter, W., & McCallum, W. C. (1978). Cognitive psychophysiology: The endogenous components of the ERP. In E. Callaway, P. Tueting, & S. Koslow (Eds.), *Event-related brain potentials in man* (pp. 349–411). New York: Academic Press.

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

Duncan-Johnson, C. C., & Donchin, E. (1977). On quantifying surprise: The variation of event-related potentials with subjective probability. *Psychophysiology*, *14*(5), 456–467.

Duncan-Johnson, C. C., & Donchin, E. (1982). The P300 component of the event-related brain potential as an index of information processing. *Biological Psychology*, *14*(1–2), 1–52.

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of crossmodal divided attention on late ERP components. II. Error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*, *78*(6), 447–455. https://doi.org/10.1016/0013-4694(91)90062-9

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Florescu, I. (2014). *Probability and Stochastic Processes*. John Wiley & Sons.

Ford, J. M., White, P., Lim, K. O., & Pfefferbaum, A. (1994). Schizophrenics have fewer and smaller P300s: A single-trial analysis. *Biological Psychiatry*, *35*(2), 96–103. https://doi.org/10.1016/0006-3223(94)91198-3

Foti, D., & Hajcak, G. (2009). Depression and reduced sensitivity to non-rewards versus rewards: Evidence from event-related potentials. *Biological Psychology*, *81*(1), 1–8. https://doi.org/10.1016/j.biopsycho.2008.12.004

Foti, D., Weinberg, A., Dien, J., & Hajcak, G. (2011). Event-related potential activity in the basal ganglia differentiates rewards from nonrewards: Temporospatial principle component analysis and source localization of the feedback negativity. *Human Brain Mapping*, *32*(12), 2267–2269. https://doi.org/10.1002/hbm.21357

Gehring, W. J., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1990). The error-related negativity: An event-related brain potential accompanying errors. *Psychophysiology*, *27*(S34).

Gehring, W. J., Goss, B., Coles, M. G. H., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*(6), 385–390. https://doi.org/10.1111/j.1467-9280.1993.tb00586.x

Gehring, W. J., Liu, Y., Orr, J. M., & Carp, J. (2012). The Error-Related Negativity (ERN/Ne). In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford Handbook of Event-Related Potential Components* (pp. 231–291). https://doi.org/10.1093/oxfordhb/9780195374148.013.0120

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science*, *295*(5563), 2279–2282.

Goldman, R. I., Wei, C.-Y., Philiastides, M. G., Gerson, A. D., Friedman, D., Brown, T. R., & Sajda, P. (2009). Single-trial discrimination for integrating simultaneous EEG and fMRI: Identifying cortical areas contributing to trial-to-trial variability in the auditory oddball task. *NeuroImage*, *47*(1), 136–147. https://doi.org/10.1016/j.neuroimage.2009.03.062

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2007). It's worse than you thought: The feedback negativity and violations of reward prediction in gambling tasks. *Psychophysiology*, *44*(6), 905–912. https://doi.org/10.1111/j.1469-8986.2007.00567.x

Hare, T. A., O'Doherty, J. P., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the Role of the Orbitofrontal Cortex and the Striatum in the Computation of Goal Values and Prediction Errors. *Journal of Neuroscience*, *28*(22), 5623–5630. https://doi.org/10.1523/JNEUROSCI.1309-08.2008

Hinton, G. E., Sejnowski, T. J., & Poggio, T. A. (Eds.). (1999). *Unsupervised learning: Foundations of neural computation*. MIT Press.

Holroyd, C. B. (2004). A Note on the Oddball N200 and the Feedback ERN. *Errors, Conflicts, and the Brain: Current Opinions on Performance Monitoring*.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679–709. https://doi.org/10.1037//0033-295X.109.4.679

Holroyd, C. B., Dien, J., & Coles, M. G. H. (1998). Error-related scalp potentials elicited by hand and foot movements: Evidence for an output-independent error-processing system in humans. *Neuroscience Letters*, *242*(2), 65–68. https://doi.org/10.1016/S0304-3940(98)00035-4

Holroyd, C. B., & Krigolson, O. E. (2007). Reward prediction error signals associated with a modified time estimation task. *Psychophysiology*, *44*(6), 913–917. https://doi.org/10.1111/j.1469-8986.2007.00561.x

Holroyd, C. B., Krigolson, O. E., & Lee, S. (2011). Reward positivity elicited by predictive cues. *Neuroreport*, *22*(5), 249–252. https://doi.org/10.1097/WNR.0b013e328345441d

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2003). Errors in reward prediction are reflected in the event-related brain potential. *Cognitive Neuroscience and Neuropsychology*, *14*(18), 2481–2484. https://doi.org/10.1097/01.wnr.0000099601.41403.a5

Holroyd, C. B., Pakzad-Vaezi, K. L., & Krigolson, O. E. (2008). The feedback correct-related positivity: Sensitivity of the event-related brain potential to unexpected positive feedback. *Psychophysiology*, *45*(5), 688–697. https://doi.org/10.1111/j.1469-8986.2008.00668.x

Holroyd, C. B., Yeung, N., Coles, M. G. H., & Cohen, J. D. (2005). A Mechanism for Error Detection in Speeded Response Time Tasks. *Journal of Experimental Psychology: General*, *134*(2), 163–191. https://doi.org/10.1037/0096-3445.134.2.163

Horst, R. L., & Donchin, E. (1980). Beyond averaging. II. Single-trial classification of exogenous event-related potentials using stepwise discriminant analysis. *Electroencephalography and Clinical Neurophysiology*, *48*(2), 113–126. https://doi.org/10.1016/0013-4694(80)90298-9

Houk, James C., Adams, J. L., & Barto, A. G. (1995). A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement. In James C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of Information Processing in the Basal Ganglia* (pp. 249–270). https://doi.org/10.7551/mitpress/4708.003.0020

Hyman, J. M., Holroyd, C. B., & Seamans, J. K. (2017). A Novel Neural Prediction Error Found in Anterior Cingulate Cortex Ensembles. *Neuron*, *95*(2), 447-456.e3. https://doi.org/10.1016/j.neuron.2017.06.021

Isreal, J. B., Chesney, G. L., Wickens, C., & Donchin, E. (1980). P300 and Tracking Difficulty: Evidence For Multiple Resources in Dual-Task Performance. *Psychophysiology*, *17*(3), 259–273. https://doi.org/10.1111/j.1469-8986.1980.tb00146.x

Jackson, A. F., & Bolger, D. J. (2014). The neurophysiological bases of EEG and EEG measurement: A review for the rest of us. *Psychophysiology*, *51*(11), 1061–1071. https://doi.org/10.1111/psyp.12283

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*(5), 429–449. https://doi.org/10.3233/IDA-2002-6504

Joel, D., Niv, Y., & Ruppin, E. (2002). Actor–critic models of the basal ganglia: New anatomical and computational perspectives. *Neural Networks*, *15*(4–6), 535–547. https://doi.org/10.1016/S0893-6080(02)00047-3

Johnson Jr., R., & Donchin, E. (1982). Sequential Expectancies and Decision Making in a Changing Environment: An Electrophysiological Approach. *Psychophysiology*, *19*(2), 183–200. https://doi.org/10.1111/j.1469-8986.1982.tb02545.x

Kappenman, E. S., & Luck, S. J. (2012). ERP components: The ups and downs of brainwave recording. In S. J. Luck & E. S. Kappenman (Eds.), *Oxford Handbook of ERP Components* (pp. 3–30). New York, NY: Oxford University Press.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, *36*(14), 1.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. , 160, 3-24. In *Emerging artificial intelligence applications in computer engineering* (pp. 3–24). Amsterdam, The Netherlands: IOS Press.

Kramer, A. F., Wickens, C., & Donchin, E. (1985). Processing of stimulus properties: Evidence for dual-task integrality. *Journal of Experimental Psychology: Human Perception and Performance*, *11*(4), 393–408. https://doi.org/10.1037/0096-1523.11.4.393

Krigolson, O. E., Hassall, C. D., & Handy, T. C. (2014). How We Learn to Make Decisions: Rapid Propagation of Reinforcement Learning Prediction Errors in Humans. *Journal of Cognitive Neuroscience*, *26*(3), 635–644. https://doi.org/10.1162/jocn_a_00509

Krigolson, O. E., Hassall, C. D., Satel, J., & Klein, R. M. (2015). The impact of cognitive load on reward evaluation. *Brain Research*, *1627*, 225–232. https://doi.org/10.1016/j.brainres.2015.09.028

Krigolson, O. E., Heinekey, H., Kent, C. M., & Handy, T. C. (2012). Cognitive load impacts error evaluation within medial-frontal cortex. *Brain Research*, *1430*, 62–67. https://doi.org/10.1016/j.brainres.2011.10.028

Krigolson, O. E., Pierce, L. J., Holroyd, C. B., & Tanaka, J. W. (2009). Learning to become an expert: Reinforcement learning and the acquisition of perceptual expertise. *Journal of Cognitive Neuroscience*, *21*(9), 1833–1840. https://doi.org/10.1162/jocn.2009.21128

Kumar, Y., Dewal, M. L., & Anand, R. S. (2014). Epileptic seizure detection using DWT based fuzzy approximate entropy and support vector machine. *Neurocomputing*, *133*, 271–279. https://doi.org/10.1016/j.neucom.2013.11.009

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, *4*(2), R1–R13. https://doi.org/10.1088/1741-2560/4/2/R01

Luck, S. J. (2014). *An introduction to the event-related potential technique*.

Mathworks. (2017). Statistics and Machine Learning Toolbox. Retrieved October 30, 2017, from https://www.mathworks.com/products/statistics.html

Matsumoto, M., Matsumoto, K., Abe, H., & Tanaka, K. (2007). Medial prefrontal cell activity signaling prediction errors of action values. *Nature Neuroscience*, *10*(5), 647–656. https://doi.org/10.1038/nn1890

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 51–56.

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, *82*(2), 213–225. https://doi.org/10.1037/h0076486

Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-Related Brain Potentials Following Incorrect Feedback in a Time-Estimation Task: Evidence for a "Generic" Neural System for Error Detection. *Journal of Cognitive Neuroscience*, *9*(6), 788–798. https://doi.org/10.1162/jocn.1997.9.6.788

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

Nicolaou, N., & Georgiou, J. (2012). Detection of epileptic electroencephalogram based on Permutation Entropy and Support Vector Machines. *Expert Systems with Applications*, *39*(1), 202–209. https://doi.org/10.1016/j.eswa.2011.07.008

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

O'Doherty, J. P., Dayan, P., Friston, K. J., Critchley, H. D., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *28*(2), 329–337. https://doi.org/10.1016/S0896-6273(03)00169-7

Oliveira, F. T. P., McDonald, J. J., & Goodman, D. (2008). Performance Monitoring in the Anterior Cingulate is Not All Error Related: Expectancy Deviation and the Representation of Action–Outcome Associations. *Journal of Cognitive Neuroscience*, *19*(12), 11.

Panda, R., Khobragade, P. S., Jambhule, P. D., Jengthe, S. N., Pal, P. R., & Gandhi, T. K. (2010). Classification of EEG signal using wavelet transform and support vector machine for epileptic seizure diction. *2010 International Conference on Systems in Medicine and Biology*, 405–408. https://doi.org/10.1109/ICSMB.2010.5735413

Pavlov, I. P. (1897). The work of the digestive glands (translated by W. H. Thompson, 1902). In *Classics of Medicine Library* (Reprinted). Birmingham.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Pernet, C. R., Sajda, P., & Rousselet, G. A. (2011). Single-trial analyses: Why bother? *Frontiers in Perception Science*, *2*(November), 1–2. https://doi.org/10.3389/fpsyg.2011.00322

Pestian, J. P., Sorter, M., Connolly, B., Bretonnel Cohen, K., McCullumsmith, C., Gee, J. T., … Rohlfs, L. (2016). A Machine Learning Approach to Identifying the Thought Markers of Suicidal Subjects: A Prospective Multicenter Trial. *Suicide and Life-Threatening Behavior*, 1–10. https://doi.org/10.1111/sltb.12312

Polich, J. (1990). P300, probability, and interstimulus interval. *Psychophysiology*, *27*(4), 396–403. https://doi.org/10.1111/j.1469-8986.1990.tb02333.x

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology*, *118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Potts, G. F., Martin, L. E., Burton, P., & Montague, P. R. (2006). When things are better or worse than expected: The medial frontal cortex and the allocation of processing resources. *Journal of Cognitive Neuroscience*, *18*(7), 1112–1119.

Proudfit, G. H. (2015). The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology*, *52*(4), 449–459. https://doi.org/10.1111/psyp.12370

Quitadamo, L. R., Cavrini, F., Sbernini, L., Riillo, F., Bianchi, L., Seri, S., & Saggio, G. (2017). Support vector machines to detect physiological patterns for EEG and EMG-based human–computer interaction: A review. *Journal of Neural Engineering*, *14*(1). https://doi.org/10.1088/1741-2552/14/1/011001

Ratcliff, R., Philiastides, M. G., & Sajda, P. (2009). Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(16), 6539–6544. https://doi.org/10.1073/pnas.0812589106

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II Current Research and Theory*, *21*(6), 64–99. https://doi.org/10.1101/gr.110528.110

Ritter, W., & Vaughan, H. G. (1969). Averaged evoked responses in vigilance and discrimination: A reassessment. *Science*, *164*(3877), 326–328.

Ritter, W., Vaughan, H. G., & Costa, L. D. (1968). Orienting and Habituation to Auditory Stimuli: A study of short term changes in average evoked responses. *Electroencephalography and Clinical Neurophysiology*, *25*(6), 550–556.

Rousselet, G. A., & Pernet, C. R. (2011). Quantifying the Time Course of Visual Object Processing Using ERPs: It's Time to Up the Game. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00107

Rummery, G. A., & Niranjan, M. (1994). On-Line Q-Learning Using Connectionist Systems. *Cambridge University Engineering Department*, *37*(September), 1–20.

Sambrook, T. D., & Goslin, J. (2015). A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. *Psychological Bulletin*, *141*(1), 213–235. https://doi.org/10.1037/bul0000006

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

Schultz, W. (2015). Neuronal reward and decision signals: From theories to data. *Physiological Reviews*, *95*(3), 853–951. https://doi.org/10.1152/physrev.00023.2014

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *The Journal of Neuroscience*, *13*(3), 900–913. https://doi.org/10.1523/JNEUROSCI.13-03-00900.1993

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

Schultz, W., & Romo, R. (1990). Dopamine neurons of the monkey midbrain: Contingencies of responses to stimuli eliciting immediate behavioral reactions. *Journal of Neurophysiology*, *63*(3), 607–624. https://doi.org/10.1152/jn.1990.63.3.607

Schultz, W., Romo, R., Ljungberg, T., Mirenowicz, J., Hollerman, J. R., & Dickinson, A. (1995). Reward-related signals carried by dopamine neurons. In James C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (1st ed., pp. 233–248). https://doi.org/10.7551/mitpress/4708.003.0019

Schyns, P. G. (2010). Grand challenges in perception science: Modeling the future. *Frontiers in Psychology*, *1*. https://doi.org/10.3389/fpsyg.2010.00010

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. Appleton-Century.

Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.

Squires, K. C., Wickens, C., Squires, N. K., & Donchin, E. (1976). The effect of stimulus sequence on the waveform of the cortical event-related potential. *Science*, *193*(4258), 1142–1146. https://doi.org/10.1126/science.959831

Squires, Nancy, K., Squires, Kenneth, C., & Hillyard, Stephen, A. (1975). Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and Clinical Neurophysiology*, *38*(4), 387–401.

Stahl, D., Pickles, A., Elsabbagh, M., Johnson, M. H., & The BASIS Team. (2012). Novel Machine Learning Methods for ERP Analysis: A Validation From Research on Infants at Risk for Autism. *Developmental Neuropsychology*, *37*(3), 274–298. https://doi.org/10.1080/87565641.2011.650808

Sutton, R. S. (1988). Learning to Predict by the Method of Temporal Differences. *Machine Learning*, *3*(1), 9–44. https://doi.org/10.1023/A:1018056104778

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

Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Evoked-Potential Correlates of Stimulus Uncertainty. *Science*, *150*(3700), 1187–1188. https://doi.org/10.1126/science.150.3700.1187

Sutton, S., Tueting, P., Zubin, J., & John, E. R. (1967). Information Delivery and the Sensory Evoked Potential. *Science*, *155*(3768), 1436–1439.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2003). Support Vector Machines: Least Squares Approaches and Extensions. In J. A. K. Suykens, G. Horvath, S. Basu, C. Micchelli, & J. Vandewalle (Eds.), *Advances in*

Learning Theory: Methods, Models, and Applications (pp. 155–178). Amsterdam, The Netherlands: IOS Press (NATO-ASI Series in Computer and Systems Sciences).

Takahashi, Y., Schoenbaum, G., & Niv, Y. (2008). Silencing the critics: Understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an Actor/Critic model. *Frontiers in Neuroscience*, *2*(1), 86–89. https://doi.org/10.3389/neuro.01.014.2008

Tan, P.-N., Steinbach, M., Karpatne, A., & Kumar, V. (2019). *Introduction to Data Mining* (2nd ed.). New York, NY: Pearson Education.

Tawfik, N. S., Youssef, S. M., & Kholief, M. (2016). A hybrid automated detection of epileptic seizures in EEG records. *Computers and Electrical Engineering*, *53*, 177–190. https://doi.org/10.1016/j.compeleceng.2015.09.001

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4), i – 109. http://dx.doi.org/10.1037/h0092987

Tibdewal, M. N., & Tale, S. A. (2017). Multichannel detection of epilepsy using SVM classifier on EEG signal. *Proceedings - 2nd International Conference on Computing, Communication, Control and Automation, ICCUBEA 2016*, 1–6. https://doi.org/10.1109/ICCUBEA.2016.7860106

Vapnik, V., & Chervonenkis, A. Y. (1974). *Teoriya raspoznavaniya obrazov: Statisticheskie problemy obucheniya*. Moscow, Russia: Nauka.

Vapnik, V., & Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, *24*, 774–780.

Vapnik, V. (1998). *Statistical Learning Theory*. New York: Wiley-Interscience.

VPixx Technologies—DataPixx2 Display Driver. (2017). Retrieved October 18, 2017, from VPixx Technologies website: http://vpixx.com/products/tools-for-vision-sciences/display-drivers/datapixx2/

Walt, S. van der, Colbert, S. C., & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*, (13), 22–30. https://doi.org/DOI:10.1109/MCSE.2011.37

Wasserman, L. (2013). *All of statistics: A concise course in statistical inference*. Springer Science & Business Media.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards* (PhD Thesis). King's College, Cambridge.

Wickens, C., Kramer, A. F., Vanasse, L., & Donchin, E. (1983). Performance of concurrent tasks: A psychophysiological analysis of the reciprocity of information-processing resources. *Science*, *221*(4615), 1080–1082. https://doi.org/10.1126/science.6879207

Yeung, N., Holroyd, C. B., & Cohen, J. D. (2005). ERP Correlates of Feedback and Reward Processing in the Presence and Absence of Response Choice. *Cerebral Cortex*, *15*(5), 535–544. https://doi.org/10.1093/cercor/bhh153

Yuvaraj, R., Rajendra Acharya, U., & Hagiwara, Y. (2018). A novel Parkinson's Disease Diagnosis Index using higher-order spectra features in EEG signals. *Neural Computing and Applications*, *30*(4), 1225–1235. https://doi.org/10.1007/s00521-016-2756-z