Reward Processing when Evaluating Goals:

Insight into Hierarchical Reinforcement Learning

Chad C. Williams

University of Victoria

Honours Thesis

April 2016

Supervisors:

Dr. Clay Holroyd

Dr. Olave Krigolson

Abstract

Reinforcement learning can be very complex. Although early research defined it fairly simply, with the advancement of technology, the conceptual understandings and measures of this construct has evolved. Even so, there is still much that is not understood about the reward processing mechanisms that underlie reinforcement learning, and there are still many questions that need examination. One such question is what information is considered feedback when learning. Specifically, here we sought to investigate whether the actions of others would be perceived as feedback and recruit reward processing mechanisms. Another unanswered question pertains to whether the mechanisms of reinforcement learning would be affected by goal-proximity. As reinforcement learning theory would predict an augmentative function of reward processes when approaching a goal, we also sought to examine this claim. We tested these questions by having participants play a series of Tic Tac Toe games against a computer. We manipulated valence by having the computer play good and bad moves and we manipulated goal-proximity by having the participant begin within an on-going game in which it was the computers first, second, or third move. Our results indicated that there was no reward positivity in response to the actions of the computer (good and bad moves). The P300 was present in response to the actions of the computer, but no effect of valence was found. Consequently, goal-proximity had no effect on either of the components. Interestingly, we found a reward positivity and an effect of valence on the P300 to the outcomes of the game (win and loss). This is in support of hierarchical reinforcement learning that extends traditional reinforcement learning to higher-level evaluations of a series of behaviours that lead to a goal.

Reward Processing when Evaluating Goals:
Insight into Hierarchical Reinforcement Learning

The Conceptual Understanding of Reinforcement Learning

The conceptual findings of reinforcement learning by pioneer research still plays a key role in its definition today. Although these first findings are adequate in understanding the general concept of reinforcement learning, electroencephalography (EEG) and intra-cranial measures led to the discovery of the processes and biological mechanisms that guide this type of learning. Accordingly, the evolution of the conceptual understanding of reinforcement learning was a series of modifications from the original definition rather than complete overhauls of the concept.

B.F. Skinner (1958) described reinforcement learning to be a type of operant conditioning that changed a behaviour's likelihood of occurrence due to its association with an event or stimulus. Furthermore, he explained a reinforcement to be any event or stimulus that gives a sense of pleasure. The reinforcement becomes associated with the behaviour and thus an organism learns that their behaviour may lead to a pleasurable outcome (Skinner, 1958). As humans instinctually approach pleasure, these reinforcement-behaviour associations motivate their behaviour (Thorndike, 1935). Although this is true and important for future research, it does not discern the underlying processes that are involved in this type of learning.

With the emergence EEGs, a different approach to measuring reinforcement learning evolved. This technique allowed researchers to associate reinforcement learning to specific neural activity with a high temporal resolution (Luck, 2014). As this field of research progressed, it became apparent that at least two separate processes were involved in reinforcement learning: early and late error processing (Luu, Tucket, & Stripling, 2007; Luu, Shane, Pratt, & Tucket,

2009). Whereas early error processing would discern whether an unexpected error occurred (Holroyd and Coles, 2002), late error processing would then consciously update internal representations of behaviours and outcomes (Luu et al., 2007; Luu et al., 2009). This research was certainly dissecting the processes underlying reinforcement learning; however, a biological understanding of what drove them was presented by animal research.

In 1997, research by Schultz, Dayan, and Montague demonstrated a finding that re-conceptualized the way reinforcement learning was perceived. Using intracellular techniques to measure the neural firings of dopamine producing cells in monkeys, they discovered that dopamine firing increased in response to an unexpected reward and decreased to the lack of an expected reward. Dopamine is a neurotransmitter that is theorized to be involved with the pleasure system of the brain (Schultz et al., 1997); therefore, it is logically the driving force behind reinforcement learning. These findings led to research that described that the modulation of dopamine firing from baseline firing rates presented a detection where one's expectations did not match the actual outcome and indicated a need to update internal representations of events and outcomes (Holroyd and Coles, 2002). This was supported by Schultz et al. (1997) in that they found the increased dopamine firing to the reward decreased back towards baseline as the monkeys learned – and expected – the reinforcement.

With this and other emerging research, Holroyd and Coles (2002) conducted their own study comparing participants with a computational model of reinforcement learning. They described early error processing to be a process of prediction errors- when one's expectation did not match the actual outcome. Meanwhile, they defined late error processing to be the allocation of attentional resources to update internal representations of behaviour-outcome contingencies. As supported by Schultz et al. (1997), this error detection system was theorized to be driven by

the firing of dopamine which then recruited other brain mechanisms to process the discrepancy and update one's expectations. This again improved the conceptual understanding of reinforcement learning; however, one final insight was necessary to reach the modern definition. Although, at the time Holroyd and Coles (2002) described this process to be dependent on prediction errors elicited by error feedback, further investigations determined that this was in fact driven by reward feedback (Foti, Weinberg, Dien, & Hajcak, 2011; Holroyd, Pakzad-Vaezi, & Krigolson, 2008). Therefore, the processes involved in reinforcement learning were in fact early and late reward processing rather than error processing (Holroyd et al., 2008).

Due to these findings, and for the remainder of this article, reinforcement learning will be conceptually defined as the change in the likelihood of a behaviour due to reward prediction error signals which recruit attentional resources to update behaviour-outcome contingencies. Although somewhat complicated, this definition is just a modification of the one B.F. Skinner first proposed.

While some research examined the processes of reinforcement learning, other research explored whether it functioned similarly across contexts. As humans are social beings, the majority of scenarios in which they find themselves are likely to be within a social setting. One limitation of a lot of reinforcement learning research is that it overlooks the social context that is ingrained in the human species. Fortunately, there is a field of research that does explore reinforcement learning in a social context. This research has found considerable evidence that humans learn from each other in a similar way – and likely using the same processes – as they do in these isolated environments (Yu & Zhou, 2006). The literature on reinforcement learning in a social context too has diverged into subfields which include, but are not limited to, learning in cooperation and competition. This research involves observing the actions and outcomes of

another individual. This can be done in that the success of the partner has no benefit to the participant, has a benefit to the participant, or has an impedance to the participant. The latter two, of course, refer to research focusing on learning in cooperation and competition, respectively. In all three of these approaches, observing the actions and outcomes of another has shown to influence learning (Armantier, 2004).

When observing another, even without any consequence to oneself, learning still occurs (Fukushima & Hiraki 2009; Kang, Hirsh, & Chasteen, 2010). For instance, Kang et al. (2009) had participants complete a Stroop task. The task was to identify the colour of the presented word. The difficulty of this task varied as the presented word was either the same as the colour (congruent), different from the colour (incongruent), or a non-colour word (neutral). Participants must then inhibit the word itself and focus on the colour. In this study, participants were to complete the task themselves, watch a friend complete the task, and watch a stranger complete the task. Using neural measures to identify reward processing, they found that observing others still recruited reinforcement learning mechanisms. Furthermore, they found that early reward processing was more powerful when the participant was watching a friend rather than a stranger, implying that observational reinforcement learning mechanisms were dependent on self-other overlap (the similarity between the two). Other research, however, has failed to find this distinction between friend and stranger in early reward processing but found it in late reward processing (Leng & Zhou, 2010). Thus it is unclear how interpersonal relationship further affects reward processing.

Other research found similar observational reward processing in which the outcome of the partner worked in cooperation with the participant (Bellebaum & Colosio, 2014; Bellebaum, Kobza, Thiele, & Daum, 2010; Koban, Pourtois, Bediou, & Vuilleumier, 2012; Kobza, Thoma,

Daum, & Bellebaum, 2011; Yu & Zhou, 2006) or in competition with the participant (Itagaki & Katayama, 2008; Marco-Pallarés, Krämer, Strehl, Schröder, & Münte, 2010). A study observed early reinforcement learning mechanisms in a cooperative and competitive setting across two experiments (Itagaki & Katayama, 2008). In both of the experiments, the task was to select one of two cards. On each trial, it was either the choice of the participant or of the partner (which was a computer player). Once a card was selected, both of the cards changed colours to indicate which one was a win and which one was a loss to identify whether the selected card had been rewarding. In the cooperative experiment, the partner's winnings and losses contributed to the participant's total earnings in a parallel fashion. In the competitive experiment, the partner's winnings and losses contributed to the participant's total earnings in an inverse fashion. The results indicated that early reward processing mechanisms were influenced by whether the participant was actively playing or observing their partner. Specifically, both experiments found present, yet reduced, early reward processing when observing their partner in comparison to when participating actively. In the competitive setting, however, reward processing was reversed in that the loss of the partner was regarded as a rewarding event, and the win of the partner was regarded as a punishing event. This then indicated that reinforcement learning is sensitive to how the outcome of an event affects oneself and not to the objective valence of the outcome.

There is less research on late reward processing in observational settings. Similar to early reward processing, late reward processing has been shown to be reduced when observing another's feedback than when receiving feedback of one's own actions (Bellebaum & Colosio, 2014; Leng & Zhou, 2010). Furthermore, Leng and Zhou (2010) found that late reward processing was larger for gain (reward) trials than loss trials in both active and observational conditions. This is in congruence with studies that have found an effect of valence in late reward

processing (Wu & Zhou, 2009, Cano, Class, & Polic, 2009); however, other research has not supported this claim (Sato et al., 2005; Yeung & Sanfey, 2004). Thus even though there is evidence that late reward processing is reduced in an observational setting in correspondence to an active setting, there is no clear consensus as to whether it is affected by valence.

The wealth of research on reinforcement learning, while using a variety of measures, has helped us understand a lot about reward processing. There is still, however, much to learn. One similarity across all of the research thus far defined is that the feedback was delivered explicitly. For example, Schultz et al. (1997) delivered juice in response to a sensory cue, Holroyd and Coles (2002) presented an image of various vegetables to indicate whether each response was correct or incorrect, and Miltner et al. (1997) presented auditory, visual, or somatosensory feedback to the performance of their participants. Although these approaches have given a lot of insight as to how reinforcement learning functions, they are not always realistic. Often in real world settings, our behaviours are not paired with explicit feedback. For example, when conversing with another, there is no juice, symbol, or sound that emerges to indicate whether the individual is enjoying the conversation. Instead, we search for ambiguous cues to use as feedback. If the other is smiling or providing all of their attention, this could be an indication that they enjoy the conversation. Thus a limitation to much of the research on reinforcement learning is that it does not require the processing of ambiguous feedback. In the present research, we sought to examine reward processing when feedback was interpreted by the actions of an opponent.

Another interest of reinforcement learning would be to examine whether reward processing was affected by goal-proximity. Although reinforcement learning theory would theorize prediction errors to increase linearly to goal-proximity (Sutton & Barto, 1998), evidence

supporting this claim is only recently arising and is thus far restricted to animal research (Enomoto et al., 2011; Yamada et al., 2013). Alternatively, related yet different research with humans has demonstrated that reward prediction error signals propagate backwards from the reward stimulus to the cues that can predict the reward (Niv, Duff, & Dayan, 2005; Krigolson, Hassall, & Handy, 2014; Schultz et al., 1997). If there were a series of cues that led to the rewarding state, the prediction errors would propagate backwards one cue at a time, beginning with the cue that immediately preceded the rewarding stimuli. This means that early in learning one would expect to see larger prediction errors to cues that were closer to the rewarding goal than cues that were farther from the goal. In a completely learnable task, this would quickly propagate backwards and would result in prediction errors exclusively to the initial cue (Krigolson et al., 2014). However, if the cues were not completely reliable predictors of reward (if they do not always lead to a reward), then the value may not propagate back to the initial cue but would instead scale to reward-proximity as a function of the cues reliability as a predictor of that reward (Enomoto et al., 2011; Yamada et al., 2013). In this situation, it would be expected to find larger prediction errors to cues that immediately precede the reward (goal) in comparison to cues that were further from the reward. Here, we sought to explore this by analyzing neural activity when participants were processing actions at different proximities to the goal.

In summary, reinforcement learning is a mechanism that affects the likelihood of behaviours in response to rewarding events or stimuli. Processes involved in this complex learning mechanism include the computation of prediction errors and the allocation of resources to update ones behaviour. Reinforcement learning operates across contexts including when a participant is directly involved in a task, when a participant is observing a task, and when a participant is in a cooperative or competitive setting. Research has indicated that reward

processing is most sensitive when a person is directly involved in the task, thus one learns to a stronger degree when they are performing than when they are observing. The presence of reward processing in social contexts (observational, cooperative, and competitive); however, does indicate that humans do, in fact, learn from each other using reinforcement learning. This is important for many practical reasons, however, this research still relies on explicit feedback. The environment is often ambiguous when delivering feedback and so it is also important to explore whether humans learn when interpreting feedback from the actions of others. Furthermore, the back-propagation of prediction errors is important in understanding how learning occurs in more complex tasks. If a task were completely learned, one would expect to find prediction errors to an initial cue that is a reliable predictor of achieving a reward, but if the cues were less reliable, one would expect prediction errors to scale to goal-proximity.

Measures of Reinforcement Learning

The conceptual definition of reinforcement learning necessarily evolved with the progressing measurements used to test the construct. Measurements of reinforcement learning have been extensive and include behavioural measures such as acquisition trials, extinction trials, and accuracy rates, indirect neural measures such as EEGs, and direct neural measures such as intracranial recordings.

Early studies of reinforcement learning measured whether learning occurred by analyzing the number of behaviours that were performed. Acquisition trials were measured by tallying the number of times a reinforced behaviour occurred within a period of time (Skinner, 1958). The assumption was that the more often the behaviour occurred, the stronger the learned association between the behaviour and the reinforcement was (Skinner, 1958). Similarly, extinction trials were measured as the number of times a learned behaviour occurred once it was no longer being

reinforced (Skinner, 1956). Stronger behaviour-reinforcement associations would have more extinction trials than weaker associations. This was because the lasting pleasurable effects of the reinforcement had been attributed to the behaviour and thus the behaviour had become a reinforcement itself (Skinner, 1956). Without the presence of the reinforcement, however, a new association began between the behaviour and a non-rewarding outcome and this would, over trials, overcome the behaviour-reward associations.

As behavioural studies progressed, so did the measures used to analyze reinforcement learning. Moving away from counting behaviours, other studies looked at reinforcement measures in a different light. Researchers would now use accuracy as a measure of learning in response to reinforcements in a decision task (e.g., Krigolson, Pierce, Tanaka, & Holroyd, 2009). Generally, in this paradigm participants would make a choice about several stimuli in which correct choices were rewarded (Krigolson et al., 2009). With repeated trials, participants would learn the correct responses to the presented stimuli. Accuracy was then determined by how often participants chose the correct responses. For example, Krigolson et al., 2009 had their participants classify two-dimensional polygons (blobs) into two families. Participants were given no training and were to report whether the presented blob corresponded to the presented family. Theoretically, at first participants would guess at random, but with feedback they would be able to learn which set of blobs belonged to which family. This learning would be reflected in their accuracy rates. Interestingly, this was not the case for all participants. The researchers identified that, in fact, some of the individuals were unable to learn the task and their accuracy rates remained near chance. The other individuals were, as reinforcement learning theory would predict, able to learn the task and completed with high accuracy rates. Identifying this difference in individuals, Krigolson et al. (2009) classified participants into ‘high learner’ and ‘low learner’

groups to conduct their analysis. Thus, the measure of accuracy was essential in this research to identify whether the individuals were able to learn the task from feedback.

Although these behavioural measures allowed researchers to discern whether learning occurred, the use of EEGs became popular as a tool to analyze the processes that underlie reinforcement learning. These processes were described above as early and late reward processing and are measured in EEG studies as the reward positivity and the P300, respectively.

Prior to current investigations of early reward processing, the reward positivity was named the feedback error-related negativity (Holroyd, 2008; Proudfit, 2015). To reduce confusion, it will hereby be referred to as the reward positivity. The reward positivity is an event-related brain potential (ERP) component at the frontal-central area of the scalp in response to correct and incorrect feedback stimuli (Miltner, Braun, & Coles, 1997). The amplitude of the reward positivity is quantified as the difference between the most negative deflections in correct and incorrect trials within a 250 ms to 350 ms interval of following feedback stimuli onset (Holroyd & Coles, 2002; Holroyd & Krigolson, 2007; Krigolson et al., 2014; Schultz et al., 1997). Holroyd and Coles (2002) describe this component to reflect the computation of prediction errors – where one compares their expectations to the actual outcome. It is important to note that Schultz et al. (1997) were also able to measure the same process in a different way- with the use of intra-cranial measures. Importantly, this research does indicate that the signals being measured by the reward positivity are elicited by dopamine firing. This approach, however, is not very useful in human research which cannot easily use methods that are so invasive.

Late reward processing is commonly measured as the P300- an ERP component with a positive peak 200 ms to 600 ms following feedback that is maximal at parietal-central areas of

the scalp (Luu et al., 2007; Luu et al., 2009; Sato et al., 2005; Toyomaki & Murohashi, 2005; Wu & Zhou, 2009; Yeung & Sanfey, 2004). This component is measured as the maximal peak, and is compared between peak amplitudes. Although important to reinforcement learning, less is known about this components involvement as it is also seen in studies that do not include learning (e.g., Polich & Kok, 1995). Holroyd and Coles (2002) describe this component to reflect a recruitment of attentional resources to update behaviour-outcome contingencies. This explanation makes sense as many other non-reinforcement learning processes would require the recruitment of additional attentional resources and so this component would then be seen in those studies as well.

The reward positivity has been demonstrated in many experiments that involve the direct actions of participants; however, it has also been found in social contexts. Specifically, it has been revealed when participants were observing the actions and outcomes of others. Yu and Zhou (2006) were the first to discover this component that has since been named the observational reward positivity. In their experiment, a participant either engaged in or observed another engage in a gambling task. The task was to select one of two cards that either had the number five or 25 on it. The number corresponded to the amount they were gambling. Once chosen, the card changed colour to indicate which card resulted in a win and which card resulted in a loss. They found that the reward positivity was reduced, yet still existed, in the observational condition in contrast to the active condition indicating that participants still computed prediction errors in response to feedback delivered to others. Similar results have been found for the P300 in that it was also reduced in observational settings rather than active settings (Bellebaum & Colosio, 2014; Leng & Zhou, 2010).

Other research has elaborated this experiment to have the partner's outcome affect the participant. This research indicated that there was no difference between reward positivity amplitudes when observing another with no influence to the participant and when observing another when their performance was attributed to the participant's earnings (Bellebaum & Colosio, 2014; Bellebaum, Kobza, Thiele, & Daum, 2010; Koban, Pourtois, Bediou, & Vuilleumier, 2012; Kobza, Thoma, Daum, & Bellebaum, 2011; Yu & Zhou, 2006). The effects of prediction error signals were reversed in a competitive setting. Observing the loss of a competitor elicited a rewarding event and observing the win of a competitor elicited a punishing event (Itagaki & Katayama, 2008; Marco-Pallarés, Krämer, Strehl, Schröder, & Münte, 2010). This indicated that reward processing is sensitive as to how outcome affects oneself, and not just salient to the objective valence of the outcome feedback. There is considerably less research on the P300 in cooperative and competitive settings. Furthermore, as there is controversy as to whether the P300 is sensitive to valence, it is difficult to asses whether there should be any differences in these types of social settings. It is then necessary for future research to explore the P300 in cooperative and competitive settings to better understand its sensitivity to social contexts.

Research with reinforcement learning often uses explicit stimuli to indicate whether the actions performed were correct or incorrect. Furthermore, past research has found that reward processing occurs in response to feedback across modalities (Miltner et al., 1997). Although this is true, a lot of the research focuses on visual stimuli. Stimuli that has been used varies from symbols, to pictures, to words. Regardless of what is used, however, they all indicate the same explicit information: whether the action performed was the correct or incorrect behaviour. This has then defined the reward positivity, in particular, to be sensitive to explicit feedback. It is

important, however, to determine if this component – and prediction errors – only arise in response to explicit feedback. Here we sought to determine whether reward processing was sensitive to the actions of another – a more ambiguous form of feedback.

Finally, as reinforcement learning theory would predict a scaling of prediction errors to the proximity of goals (Sutton & Barto, 1998), it is important to examine this concept. Although there is no human research as to whether this occurs while utilizing EEGs, animal research has supported this claim (Enomoto et al., 2011; Yamada et al., 2013). For example, Enomoto et al. (2011) examined the firing of dopamine neurons in a multistep choice task. Monkeys learned to conduct a series of actions to receive multiple rewards. They were to depress a rewarding target button among two non-rewarding buttons. They then were able to receive additional rewards by pressing the button once or twice more, depending on the condition. Each trial was separated into two stages: exploration and exploitation. Whereas exploration referred to when the monkeys were searching for the target, exploitation was when the monkeys continued to select the target for additional rewards. The firing of dopamine cells scaled to the actions in the exploration phase in that the firing rate increased with each additional action. They concluded that this was because the monkeys learned that as they progressed through the buttons, the likelihood of them receiving a reward increased.

Alternatively, a related field of research with humans has investigated the mechanism of back propagation by examining the amplitude of the reward positivity. As learning occurs, the reward positivity would decrease for the reward state and increase for the cue state (Krigolson et al., 2014). In a deterministic task, Krigolson et al. (2014) presented two coloured blocks three times to their participants who were to choose between them. Across the three trials, one of the colours would always lead to a reward while the other would never lead to a reward. The

participant could immediately learn in this task because if they chose the rewarding block, they would then learn to select it again, but if they chose the non-rewarding block, they would learn to select the other. They found that the reward positivity to feedback stimuli was only present on the first trial, but in proceeding trials it was absent. Furthermore, they found a positive deflection in trials two and three to the onset of the two blocks (which is similar to a cue). They then concluded that the prediction error, as measured by the reward positivity, for the reward in the first trial propagated backward to the onset of the blocks. Thus if the task was not deterministic, one may expect to find that the reward positivity scales to the proximity of the goal in that the cues closer to the goal would have a larger reward positivity than the cues further from the goal.

In summary, the measures used to examine reinforcement learning has progressed considerably. Initially, research focused on the number of behaviours produced when reinforcing or extinguishing the behaviour to indicate whether, and to what strength, learning occurred. As behavioural measures developed, gambling tasks were used in that learning was identified as the number of times the participant chose the target stimuli among distractors. Although these were helpful for identifying the presence of learning, EEG research began to dissect the processes underlying reinforcement learning. Two components have been discovered, the reward positivity and the P300, which are theorized to compute prediction errors and allocate attentional resources, respectively. These components have also been examined in social contexts in that they are present, but reduced, when observing the actions and outcomes of others in comparison to processing feedback of one's own actions. Whether these components are still recruited when participants are to interpret feedback using ambiguous information (the actions of another), however, is still to be determined. Furthermore, it is still unclear as to whether these components scale to goal-proximity in humans.

Present Study

Although a lot is known about how reward processing occurs when being delivered feedback explicitly, or when being modelled by others, whether reward processing is involved when watching an opponents action is still poorly understood. Furthermore, whether reward processing is affected by goal-proximity is still in need of examination. In the present study, we sought to examine neural activity in response to the actions of one's opponent by utilizing EEGs. Participants were to play a computerized game of Tic Tac Toe against a computer which played good or bad moves. The actions of the computer which were good served as negative feedback and the actions that were bad served as positive feedback. This is because the computer playing a good move would disadvantage the participant while a bad move would benefit the participant. Furthermore, goal-proximity was manipulated by having participants begin at different stages of on-going games. Specifically, participants began each game when it was the computers first, second, or third move. Reward processing neural activity was measured as the peak amplitudes of the reward positivity and the P300. We hypothesize that we would see the reward positivity reflected by the actions of the computer, similar to when feedback is directly delivered or modeled. We also hypothesized to see a P300 in response to computers actions, however, due to the discrepancy of past research, we could not make a claim as to whether it would be affected by valence. Furthermore, in congruence with reinforcement learning theory, we sought to examine whether the reward positivity and P300 amplitudes would increase linearly to the proximity of the goal in that the third move condition would be the largest and the first move condition would be the smallest.

Methods

Participants

20 undergraduate students (15 female, mean age: 20) were recruited from the University of Victoria's Psychology Research Participation pool to participate in the experiment. One participant was removed due to issues with data collection, and another participant was removed as their component amplitudes were more than two standard deviations from the mean. This resulted in an analyses of 18 participants (14 female, mean age: 20). All participants had normal or corrected-to-normal vision, no known neurological impairments, and volunteered for extra course credit in a psychology course. All participants provided informed consent approved by the Human Research Ethics Board at the University of Victoria, and the study followed ethical standards as prescribed in the 1964 Declaration of Helsinki.

Apparatus and Procedure

Participants were seated in a sound dampened room in front of a 19" LCD computer monitor and used a standard USB mouse to play a series of Tic Tac Toe games against a computer (written in MATLAB [Version 8.6, Mathworks, Natick, U.S.A.] using the Psychophysics Toolbox extension (Brainard, 1997)).

Tic Tac Toe is a quick paced two player game. The game begins with an empty 3 x 3 grid. One player takes the role of 'X' and the other takes the role of 'O'. Each player takes turns (X starts) in which they select a square to occupy. The goal of the game is to occupy three squares in a row. These rows may be occupied vertically, horizontally, or diagonally. As soon as one of the players achieves this, the game is ended. If all nine squares are filled, and neither player has three in a row, the game is deemed a tie. In the present study, the participant played the role of X, yet always began within an ongoing game. Furthermore, the game would always begin on the computers turn in one of three conditions: where it was the computers first move, second move, or third move (see Figure 1). The conditions were presented randomly. The

computer would then play either a good or bad move (see Figure 2) and this event would be used to analyze reward processing of the participant. For the remainder of each game, the computer would play randomly unless an action would result in its win. In such a case, the computer would select this action. Each participant played 240 games which was evenly divided between the three conditions.

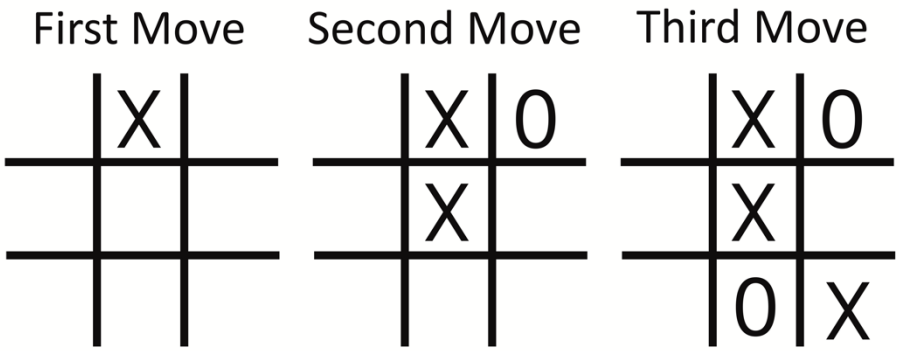

Figure 1. Manipulation of goal-proximity. Examples of the starting board state for the first, second, and third computer move conditions.

The game was played on a light grey background. The game board was 2.5 cm by 2.5 cm. The X's and O's were presented in a white 26 Monospace font. The game grid changed colour to indicate who's turn it was. Whereas white meant that it was the participants turn, black meant that it was the computers turn. Upon first presentation of the board state, the computer would respond between 1300-1500 ms to give the participant enough time to analyze the board so that they were able to discern good moves from bad moves. Once the game was won, lost, or tied the game board would disappear and a white 26 Monospace font fixation cross would appear in the middle of the screen for 800-1000 ms. Following fixation, the words 'WIN', 'LOSS', or 'TIE' appeared in the middle of the screen in the same size and font for 3000 ms. A new game would then load. Every 10 trials, there was a participant controlled rest break which delivered feedback on how many games they had thus far won.

To determine a good move from a bad move, we created a computational model in which

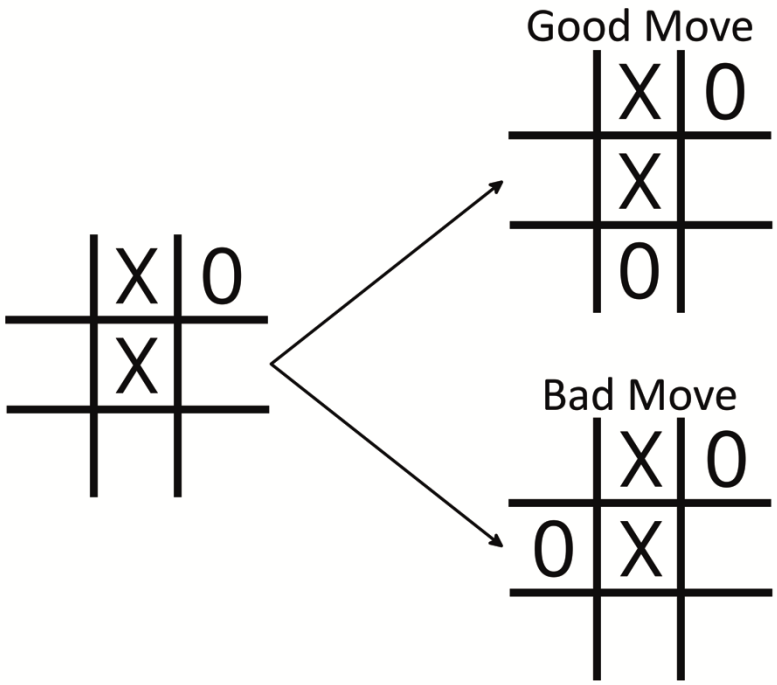

Figure 2. Manipulation of valence. Example of a good and bad move as played by the computer in the second move condition.

the computer would play itself and attach values to each game state by computing prediction errors (Schultz et al., 1997). If the computer had not encountered the game state it was currently in, it would add it to its repertoire with a value of 0. It updated the values of board states by computing the difference between the state of the current trial (e.g., 1 if it wins) and the previous value of the game state (e.g. 0 if it had been a new state), multiplying it by a learning rate coefficient (0.5), and adding it to the previous board state value. For example, if the value of the state was previously 0.5, and the computer won the game with that board state (which results in a current state value of 1), the difference would be 0.5 which would be multiplied by the learning rate and then added to the original value, resulting in a value of 0.75. Over time, this number would increase towards 1 as to indicate that it was a very good move. Thus low proportions (e.g., 0.002) were deemed bad moves and high proportions (e.g., 0.998) were deemed good moves. We removed any game states in which the computer could win immediately, and any game states in

which the participant could win in two ways. The former criterion was to ensure that the computer move was not confounded by the end of the game, and the latter criterion was to ensure that there was always a good and a bad move for the computer to play (if the participant could win in two ways, there was no good move). We then extracted the best and worst move of each game state using the max and min function, respectively. This resulted in every game state having a good and a bad move.

As there are only 9 possible game states in the first move condition, we computed the five best and five worst moves for each of the nine states and randomly chose four of the states to have five good and bad moves, and five of the states to have four good and bad moves (to total to 40 states, each with a good and bad move). We randomly chose 40 game states of the second and third move conditions with their respective good and bad moves. Thus, for the second and third move conditions, every starting state was novel. We randomly presented all game states twice. As the computers initial move was randomly chosen (50% chance of a good or bad move), this could result in seeing the same outcome for the same game state twice.

Data Acquisition

The electroencephalogram (EEG) data were recorded using 64 electrodes which were mounted in a fitted cap with a standard 10-20 layout (see www.neuroeconlab.com for the layout). Electrodes on the cap were initially referenced to a common ground. On average, electrode impedances were kept below 20 kΩ. The EEG data were sampled at 500 Hz, amplified (ActiCHamp, Revision 2, Brainproducts GmbH, Munich, Germany), and filtered through an antialiasing low-pass filter of 8 kHz.

Data Analysis

Offline EEG data analyses were conducted using Brain Vision Analyzer software (Version 7.6, Brainproducts, GmbH, Munich, Germany). First, excessively noisy or faulty electrodes were removed. Next, the sampling rate of the data were lowered to 250Hz. The EEG data were re-referenced to linked mastoids and filtered using a Butterworth filter with 0.1Hz to 30Hz and 60Hz notch criteria. The data were then segmented to a 3000 ms interval spanning 1000 ms prior to feedback stimulus to 2000 ms following feedback stimulus onset. Independent component analyses, as described by Luck (2014), were used to remove ocular artifacts. Channels that were previously removed were interpolated using spherical splines. The data were then re-segmented into an 800 ms epoch, 200 ms prior to feedback stimulus and 600 ms following feedback stimulus onset. The 200 ms prior to feedback stimuli was baseline corrected. Final segmentations with the same 800 ms interval as described above was conducted to separate the good moves from bad moves. Each of the conditional waveforms were put through artifact rejection with 10μV/ms gradient and 150μV absolute difference criteria. For each participant, ERP's were created by averaging the EEG data for each electrode and each conditional level (i.e., good moves and bad moves). Difference waveforms were created by subtracting good moves from bad moves. For each conditional and difference waveform, a grand average waveform was created by averaging corresponding waveforms across all participants. Although it was not initially the focus of the current study, the same processes were conducted for win and loss feedback. These difference waveforms were created by subtracting loss waveforms from win waveforms.

The components of interest were the reward positivity and P300. The reward positivity is a component that deflects negatively 250-350 ms following feedback and is found on fronto-central areas of the scalp. The reward positivity amplitude for each condition (first, second, and

third move) and each participant was measured as the mean difference wave amplitude ± 25 ms centered on the average peak of the difference waveforms (264 ms) following the computers initial move at channel FCz. The time window was chosen as it was centered around the reward positivity based on visual inspection of the data and past literature (Holroyd & Coles, 2002; Holroyd & Krigolson, 2007; Krigolson et al., 2014; Schultz et al., 1997). The reward positivity to the win and loss feedback was quantified in a different manner due to a large variability in its peak timing (M = 200 ms, SD = 34 ms). Instead, the reward positivity was quantified as the most positive peak of the difference waveform (win - loss) within an interval of 150-250 ms. The P300 amplitude for each condition (first, second, and third move) and each participant was measured as the mean difference amplitude ± 50 ms centered on the average peak of the difference waveforms (384 ms) following the computers initial move at channel Pz. The time window was chosen as it was centered around the P300 peak based on visual inspection of the data and past literature (Luu et al., 2007; Luu et al., 2009; Sato et al., 2005; Toyomaki & Murohashi, 2005; Wu & Zhou, 2009; Yeung & Sanfey, 2004). The P300 to the win and loss feedback was also quantified in a different manner due to a large variability in its peak timing (M = 384 ms, SD = 60 ms). Instead, the P300 was quantified as the most positive peak of the difference waveform (win - loss) within an interval of 300-400 ms. Unexpectedly, there was a component within the time range of 80-180 ms following the computers initial action. This component was unusual for reinforcement learning tasks, thus we will here call it the 'early component'. The early component amplitude for each condition (first, second, and third move) and each participant was measured as the mean difference wave amplitude ± 25 ms centered on the average peak of the difference waveforms (130 ms) following the computers initial move at channel Cz, where the component was maximal in regards to the midline. The time window was

chosen as it was centered around the early component based on visual inspection of the data. The early component to the win and loss feedback was quantified in the same manner.

A single sample t-test was conducted on the difference waveforms (bad move – good move) against zero to determine whether there was an effect of valence on all components (reward positivity, P300, early component) in each condition of goal-proximity (first, second, and third move). For each component, a single sample t-test against zero was also conducted on the game outcome difference waveforms (win – loss). To assess whether there was an effect of goal-proximity on each component across conditions (first, second, third move), a One-Way repeated measures ANOVA was also conducted on the difference waveforms. An alpha level of 0.05 was assumed for all statistical tests. Error measures for descriptive statistics reflect 95% confidence intervals (Cummings, 2013).

Results

Our analyses of the grand average ERP waveforms revealed scalp topography and timing inconsistent with the reward positivity (see Figure 3) but consistent with the P300 (see Figure 4) in each goal-proximity condition. Furthermore, our analyses revealed an early component within the time range of 80-180 ms that was maximal at channel Cz in regards to midline electrodes in the goal-proximity conditions (see Figure 5). Interestingly, our analyses of the grand average ERP waveforms for the game outcome (win and loss) revealed scalp topography and timing consistent with both the reward positivity (see Figure 3) and the P300 (see Figure 4). Moreover, the scalp topography and timing of the early component was not observed in the game outcome analysis (see Figure 5).

We conducted a single sample t-test on the reward positivity for all conditions of goal-proximity (first, second, and third move) and game outcome feedback (win, loss) to identify

component existence (see Figure 6). We found that the reward positivity was not present when processing computer actions, but was present when processing game outcome feedback (see Table 1). Single sample t-tests were also conducted on the P300 difference waveforms for all

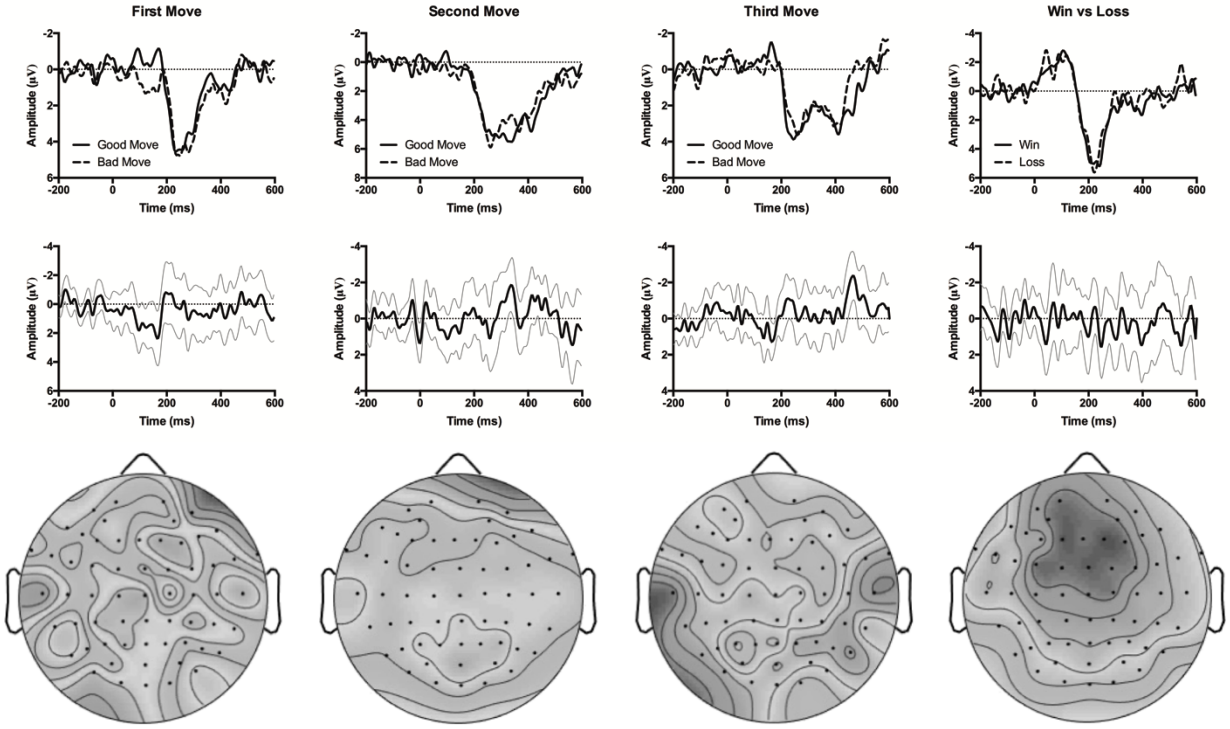

Figure 3. Conditional waveforms, difference waveforms, and topographic maps for the reward positivity across goal-proximity conditions (first, second, and third move) and for end game feedback (win, loss). The top ERP waveforms are conditional waveforms, and the bottom ERP waveforms are difference waveforms (bad moves – good moves; win – loss) with 95% confidence intervals at channel FCz. Scalp distributions for the reward positivity were created by averaging the difference waveform peak topographies of all participants.

conditions of goal-proximity (first, second, and third move) and game outcome feedback (win, loss) to identify effects of valence (see Figure 6). Similar to the reward positivity, we found that there was no difference between waveforms when processing computer actions, but there was a difference when processing game outcome feedback (see Table 1). Finally, single sample t-tests on the early component indicated that the good move waveform was significantly larger than the

bad move waveform in the first move condition, but not in any of the other conditions (second and third move) or to the end game feedback (see Table 1).

To test our hypothesis that goal-proximity (first, second, or third move) would effect observed components (reward positivity, P300, early component), a One-Way ANOVA was conducted for each component. For all components there was no main effect of goal-proximity ($p > .05$).

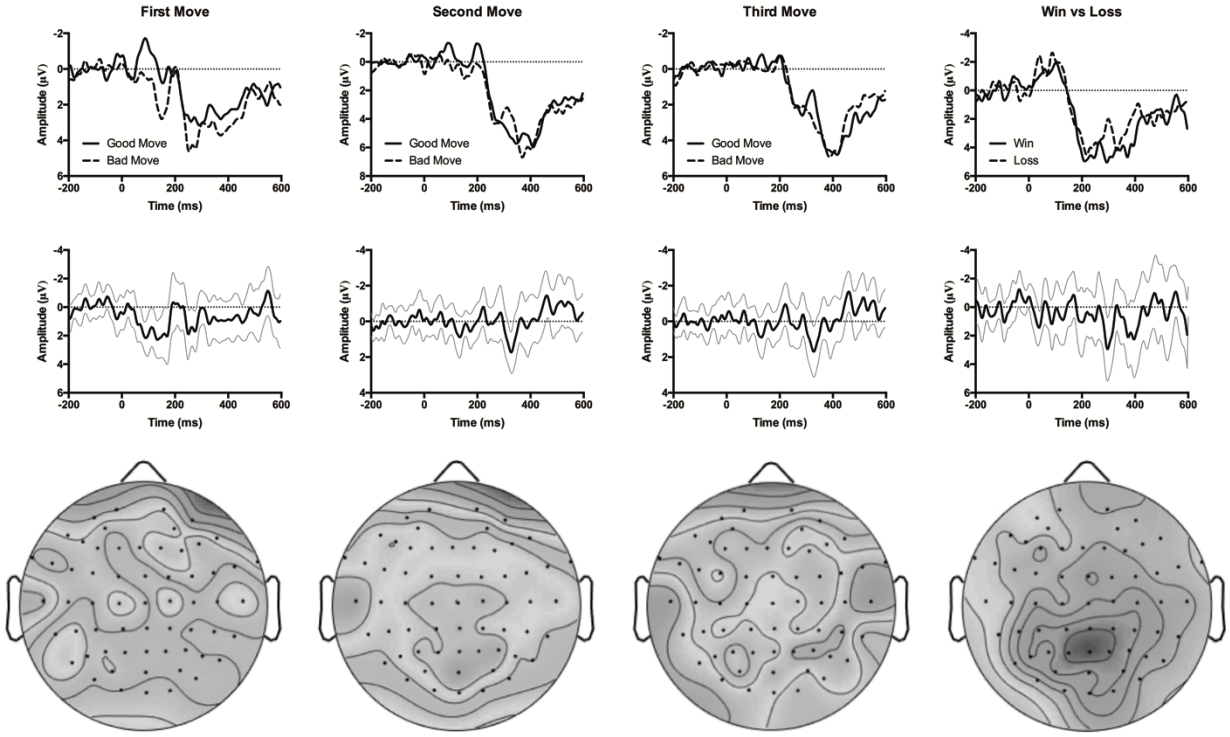

Figure 4. Conditional waveforms, difference waveforms, and topographic maps for the P300 across goal-proximity conditions (first, second, and third move) and for end game feedback (win, loss). The top ERP waveforms are conditional waveforms, and the bottom ERP waveforms are difference waveforms (bad moves – good moves; win – loss) with 95% confidence intervals at channel Pz. Scalp distributions for the P300 were created by averaging the difference waveform peak topographies of all participants.

Discussion

In the present study we have demonstrated that the actions of others did not recruit reward processing mechanisms (as measured by the reward positivity and the P300). Consequently, we were unable to analyze whether reward processing was affected by goal-

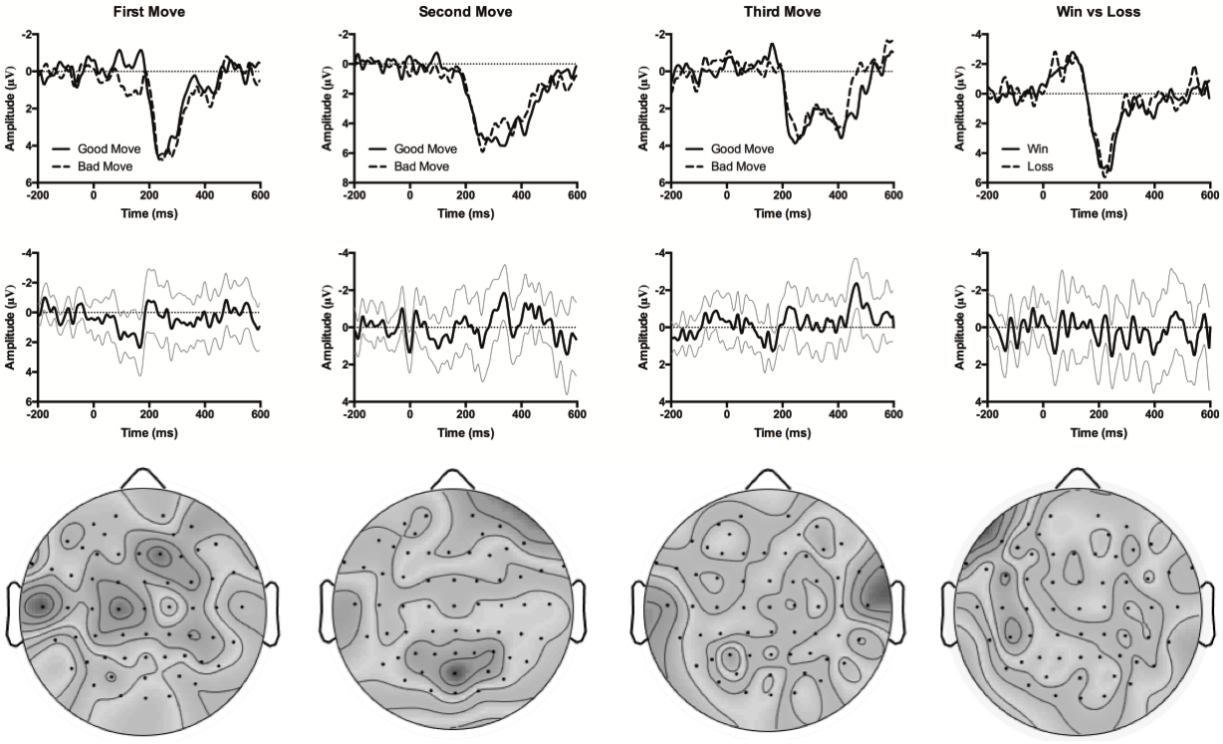

Figure 5. Conditional waveforms, difference waveforms, and topographic maps for the early component across goal-proximity conditions (first, second, and third move) and for end game feedback (win, loss). The top ERP waveforms are conditional waveforms, and the bottom ERP waveforms are difference waveforms (bad moves – good moves; win – loss) with 95% confidence intervals at channel Cz. Scalp distributions for the early component were created by averaging the difference waveform peak topographies of all participants.

proximity. Interestingly, the outcome of games (win, loss) did elicit a reward positivity and a difference in P300 amplitudes between valences. This then supports the hierarchical reinforcement learning theory (HRL) which extends traditional reinforcement learning to include overarching goals.

The current findings demonstrate a reward positivity and P300 difference in response to the goal of the game. There was not, however, reward processing when observing the actions of the computer. As it is still unclear whether the P300 would be affected by valence in the current task, it may not serve as strong evidence towards of HRL. Here, we did show that it was affected

by valence, however, and this finding can be used to help alleviate the controversy. Conversely, the presence of the reward positivity was encouraging support of HRL. The following conclusions, therefore, are in regards to the reward positivity.

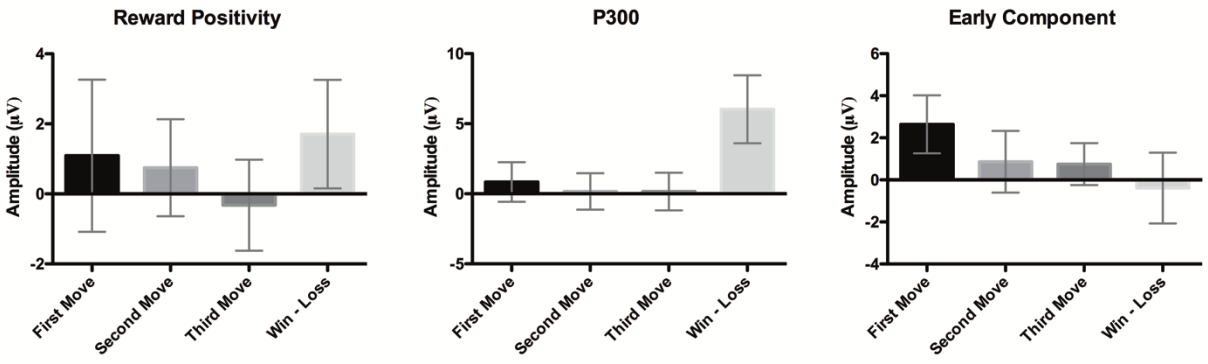

Figure 6. Peak difference waveform amplitudes with 95% confidence intervals for the reward positivity, the P300, and the early component across goal-proximity conditions (first, second, and third move) and for end game feedback (win, loss). The peak amplitudes were created by averaging the difference waveform (bad moves – good moves; win – loss) peak amplitudes of all participants.

Although it was not originally the focus of the current study, our findings support HRL (Botvinick, 2012; Botvinick, Niv, & Barto, 2008; Diuk et al., 2013; Holroyd & McClure, 2015; Holroyd & Yeung, 2012; Ribas-Fernandes et al., 2011). HRL theories conjoin theories that reinforcement learning involves both trial-and-error processing and task motivation (Holroyd & Yeung, 2012). These concepts have been termed *primitive actions* and *options*, respectively. More specifically, learning of primitive actions includes examining the immediate outcome of a single specific action, but learning of options includes examining the outcome of a series of actions which together lead to a goal. Each concept requires standard computations of prediction errors using information about the outcomes of primitive actions or options. Thus there are multiple levels of feedback which can be used in learning. Importantly, Holroyd and Yeung (2012) propose that the completion of an option serves as a pseudo-reward which propagates

back to the series of primitive actions that led to the outcome. Therefore, primitive actions are then also affected by higher level reinforcements. This indicates an important characteristic of

Table 1. *Descriptive and inferential statistics of average peak differences, 95% confidence intervals, t-values, and p-values for the reward positivity, the P300, and the early component across goal-proximity conditions (first, second, and third move) and end game feedback (win, loss).*

Component	Condition	Peak Difference	95% CI Min	95% CI Max	T-Value	P-Value
Reward	First Move	1.09	-1.08	3.26	1.06	0.305
Positivity	Second Move	0.75	-0.64	2.13	1.14	0.270
	Third Move	-0.32	-1.62	0.98	0.52	0.609
	End Game Feedback	1.71	0.16	3.26	2.33	0.033
P300	First Move	0.84	-0.56	2.25	1.27	0.223
	Second Move	0.16	-1.14	1.47	0.27	0.792
	Third Move	0.16	-1.18	1.51	0.26	0.798
	End Game Feedback	6.03	3.60	8.46	5.24	<0.0001
Early	First Move	2.64	4.04	1.26	4.02	0.001
Component	Second Move	0.86	-0.60	1.24	1.24	0.231
	Third Move	0.75	-0.25	1.75	1.59	0.132
	End Game Feedback	0.39	-1.30	2.07	0.48	0.634

HRL which proposes a solution to the limitations of traditional reinforcement learning – that the overall goal of a task defines the adaptability of a series of actions. For example, if a game of chess was played exclusively using traditional reinforcement learning, losing a pawn would be a punishing outcome which would then be learned and avoided. Alternatively, with the goal of

defeating the king, losing a pawn may place a player in a better position to win the game and may be perceived as rewarding in the long run. Thus losing the pawn is punishing in the moment, but rewarding for the overall performance of the game.

In the current task, the specific actions performed and observed by the participant would elicit primitive action-outcome associations (Holroyd & Yeung, 2012). For example, if the computer was to block the participant from winning that may be perceived as punishing. With a game of Tic Tac Toe, however, there is an overarching goal beyond individual moves – that of getting three squares in a row to win the game. In correspondence with HRL, we would then expect to see reward processing when achieving or failing to achieve this goal. This prediction error would then propagate back to the primitive actions that led to the outcome, adjusting their corresponding values. Our results show evidence that a higher level reinforcement learning system exists for options. There is, however, an inconsistency with our findings and HRL. An assumption of HRL is that primitive action reward processing would be specific to the higher level goal. This then means, we would still expect to see reward processing to primitive actions that are congruent with the option. In our task, we would still expect to see prediction error signals to the actions of the computer that are relevant to the goal of winning. We found no reward processing to the actions of the computer, but there may have been a few limitations of the current study which explain why.

One limitation could be that participants may have simply not perceived good and bad moves as punishing and rewarding, respectively. This may be because the actions of the computer were not entirely determinant of whether one would complete their goal (win the game) and so the actions may have been too ambiguous to use in higher-level reward processing. Specifically, this may be the case because if a computer plays a good move, it does not

necessarily inhibit the participant from winning the game, it just prolongs it. Thus even though it is a punishing event in that it would cause the game to last longer, and decrease the likelihood of winning, it could still result in a win (which it often did). Furthermore, if the computer played a bad move, it may not directly lead to a win either. For example, in the one move condition, regardless of whether the computer plays a good or a bad move, it would not directly lead to a win in the participants next turn. This is also the case for many of the board states in the second and third move conditions. This may then indicate that actions were not reliable predictors of the outcome of the game, and so they may not have been processed in correspondence to the goal, as would be expected by HRL.

Although this is a likely explanation, there are alternatives that may also explain our findings. For instance, perhaps the actions of another was not adequate stimuli to represent feedback in reward processing. This would then indicate that ambiguous feedback is not appropriate for learning. Another possible explanation is that the feedback was not actually linked to the participants behaviour. Each game began within a board state in which it was the computers turn. We then analyzed how participants processed this initial move. As these turns either place the participant in a better or worse board state, we presumed that it would function similar to feedback. There was, however, a disconnection in that these initial moves were not actually linked to the actions of the participant but instead to the board state that the participant had been placed into. Thus it was not feedback of their actions, but of the environment. It is then possible that reinforcement learning requires explicit feedback, or a causal link between actions and feedback.

Another limitation could be that participants had no need to recruit reinforcement learning mechanisms. Tic Tac Toe is not a very complicated game, and there are not very many

defining moves that will determine the outcome. Furthermore, in North America, Tic Tac Toe is fairly popular and it is likely that many, if not all, of our participants have had a lot of previous experience with the game. If this was true, it is possible that they had already learned these defining moves and did not need to learn anything more to win the game. This may be indicated by the engrossed ratio of wins the participants procured in comparison to both losses and ties. If they truly needed to learn about the game, we would expect to see many more losses than what was observed in our sample. This seems unlikely, however, as there was reward processing to the evaluation of the end game outcome. Thus if the participants were not learning, we would also expect to see no reward processing to the win and loss feedback.

A final limitation could be that the computer moves were predictable. If the participants were able to predict the move of the computer, then we would also expect not to see a reward positivity. The reward positivity is theorized to reflect the computation of prediction errors – where one's expectations of an event did not match the actual outcome (Holroyd and Coles, 2002). As one learns, their expectations more closely reflect the outcome so that their prediction errors – and the reward positivity – diminishes. Thus, if an event is completely expected, there would be an absence of prediction errors and there would be no reward positivity. As the likelihood of the computer to play a good move or a bad move was equiprobable, this is not likely as they would not be able to predict the outcome. Past research has used paradigms in which reward and non-reward outcomes were random, and still found a reward positivity (e.g., Holroyd, Krigolson, Baker, Lee, & Gibson, 2009). Thus, a reward positivity should still be present as computer actions were impossible to expect.

The early component observed in the current study was unusual in respect to reinforcement learning tasks thus it is difficult to make any conclusions about why it was

affected by valence in the first move condition. To ensure it was not an eye movement, we conducted several processes and examined the scalp distribution of the component. Combined with independent component analysis and artifact rejection, we manually scanned each trial for each condition in search of eye movement artifacts. Furthermore, the topographic maps of this component were not consistent with eye movements. We then believe that it is unlikely that this component reflects eye movements.

The early component was in the time range of attentional components such as the N1. Although a lot of literature on the N1 component observes the negativity at parietal locations of the scalp, there is evidence for a frontal N1 (Vogel & Luck, 2000). The N1 has been attributed to motor preparation processes (Blakemore, Hyland, Hammond-Tooke, & Anson, 2013; Vogel & Luck, 2000) and cognitive effort (Enge, Fleischhauer, Lesch, & Strobel, 2011). In the present study, the effect of computer move valence on the component amplitude could possibly be explained by these theories. When the board state appears, perhaps participants select their next action before the computer performs its initial move. This selected action is likely to be the position that the computer would select when playing a good move. Thus if the computer does not select that square, there is little need for motor preparation and cognitive effort as the participant simply executes their planned action. On the other hand, if the computer does select the square, then there is a higher need for motor preparation and cognitive effort as the participant must assess the new board state and select an alternative action. This effect of valence was only observed in the first move condition, and this may be due to the large variety of possible actions the participant would have to assess in that condition compared to the other two conditions (second and third move). Although this may be an encouraging explanation, in the current study we are unable to confidently explain the early component. Future research,

however, could explore whether this component is related to motor preparation and cognitive effort by using a paradigm that enables the participant to pre-select an action, yet on occasion forces them to select a new action with varying degrees of possible options.

Conclusion

Reinforcement learning is a complex mechanism. There has been a lot of success in defining this concept and developing measures that better capture it. None-the-less, there are still many unexplored factors that may influence reinforcement learning. We sought to determine whether reward processing would occur to ambiguous feedback and whether it would scale to goal-proximity. Our results indicated that there was no reward processing to the actions of another, and so we were unable to examine any effect of goal-proximity. We did find, however, that there was reward processing to the outcome of each game. This supports HRL in that reinforcement learning occurs to a series of behaviours that lead to a goal. Although HRL would still expect to see reward processing to individual actions that are related to the goal, we may not have found this because the computer's actions were not reliable predictors of game outcome.

References

Armantier, O. (2004). Does observation influence learning?. *Games and Economic Behavior*, *46*(2), 221-239.

Bellebaum, C., & Colosio, M. (2014). From feedback-to response-based performance monitoring in active and observational learning. *Journal of cognitive neuroscience*, *26*(9), 2111-2127.

Bellebaum, C., Kobza, S., Thiele, S., & Daum, I. (2010). It was not MY fault: event-related brain potentials in active and observational learning from feedback. *Cerebral Cortex*, *20*(12), 2874-2883.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Blakemore, R. L., Hyland, B. I., Hammond-Tooke, G. D., & Anson, J. G. (2013). Distinct modulation of event-related potentials during motor preparation in patients with motor conversion disorder. *PloS one, 8*(4), e62539.

Botvinick, M. M. (2012). Hierarchical reinforcement learning and decision making. *Current opinion in neurobiology, 22*(6), 956-962.

Botvinick, M. M., Niv, Y., & Barto, A. C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition, 113*(3), 262-280.

Cano, M. E., Class, Q. A., & Polich, J. (2009). Affective valence, stimulus attributes, and P300: Color vs. black/white and normal vs. scrambled images. *International Journal of Psychophysiology, 71*(1), 17-24.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Diuk, C., Schapiro, A., Córdova, N., Ribas-Fernandes, J., Niv, Y., & Botvinick, M. (2013). Divide and conquer: hierarchical reinforcement learning and task decomposition in humans. *In Computational and robotic models of the hierarchical organization of behavior* (pp. 271-291). Springer Berlin Heidelberg.

Enge S, Fleischhauer M, Lesch K, Strobel A (2011) On the role of serotonin and effort in voluntary attention: Evidence of genetic variation in N1 modulation. *Behavioural Brain Research 216*: 122–128.

Enomoto, K., Matsumoto, N., Nakai, S., Satoh, T., Sato, T. K., Ueda, Y., ... & Kimura, M. (2011). Dopamine neurons learn to encode the long-term value of multiple future rewards. *Proceedings of the National Academy of Sciences, 108*(37), 15462-15467.

Foti, D., Weinberg, A., Dien, J., & Hajcak, G. (2011). Event-related potential activity in the basal ganglia differentiates rewards from nonrewards: Temporospatial principal components analysis and source localization of the feedback negativity. *Human Brain Mapping, 32*(12), 2207-n/a. doi:10.1002/hbm.21182

Fukushima, H., & Hiraki, K. (2009). Whose loss is it? Human electrophysiological correlates of non-self reward processing. *Social neuroscience, 4*(3), 261-275.

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. http://doi.org/10.1037/0033-295X.109.4.679

Holroyd, C. B., & Krigolson, O. E. (2007). Reward prediction error signals associated with a modified time estimation task. *Psychophysiology, 44*(6), 913–917.

Holroyd, C. B., Krigolson, O. E., Baker, R., Lee, S., & Gibson, J. (2009). When is an error not a prediction error? An electrophysiological investigation. *Cognitive, Affective, & Behavioral Neuroscience, 9*(1), 59–70. http://doi.org/10.3758/CABN.9.1.59

Holroyd, C. B., & McClure, S. M. (2015). Hierarchical control over effortful behavior by rodent medial frontal cortex: A computational model. *Psychological review, 122*(1), 54.

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. http://doi.org/10.1111/j.1469-8986.2008.00668.x

Holroyd, C. B., & Yeung, N. (2012). Motivation of extended behaviors by anterior cingulate cortex. *Trends in cognitive sciences, 16*(2), 122-128.

Itagaki, S., & Katayama, J. I. (2008). Self-relevant criteria determine the evaluation of outcomes induced by others. *Neuroreport, 19*(3), 383-387.

Kang, S. K., Hirsh, J. B., & Chasteen, A. L. (2010). Your mistakes are mine: self-other overlap predicts neural response to observed errors. *Journal of Experimental Social Psychology, 46*(1), 229-232.

Koban, L., Pourtois, G., Bediou, B., & Vuilleumier, P. (2012). Effects of social context and predictive relevance on action outcome monitoring. Cognitive, Affective, & Behavioral *Neuroscience, 12*(3), 460-478.

Kobza, S., Thoma, P., Daum, I., & Bellebaum, C. (2011). The feedback-related negativity is modulated by feedback probability in observational learning. *Behavioural brain research, 225*(2), 396-404.

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. http://doi.org/10.1162/jocn_a_00509

Krigolson, O. E., Pierce, L., Tanaka, J., & Holroyd, C. B. (2009). Learning to become an expert: Reinforcement learning and the acquisition of perceptual expertise. Journal of Cognitive *Neuroscience, 21*(9), 1834-1841.

Leng, Y., & Zhou, X. (2010). Modulation of the brain activity in outcome evaluation by interpersonal relationship: an ERP study. *Neuropsychologia, 48*(2), 448-455.

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique 2nd ed*. Cambridge, MA: MIT Press.

Luu, P., Shane, M., Pratt, N. L., & Tucker, D. M. (2009). Corticolimbic mechanisms in the control of trial and error learning. *Brain Research, 1247*, 100–113. http://doi.org/10.1016/j.brainres.2008.09.084

Luu, P., Tucker, D. M., & Stripling, R. (2007). Neural mechanisms for learning actions in context. *Brain Research, 1179*, 89–105. http://doi.org/10.1016/j.brainres.2007.03.092

Marco-Pallarés, J., Krämer, U. M., Strehl, S., Schröder, A., & Münte, T. F. (2010). When decisions of others matter to me: an electrophysiological analysis. *BMC neuroscience, 11*(1), 86.

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. http://doi.org/10.1162/jocn.1997.9.6.788

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

Polich, J., & Kok, A. (1995). Cognitive and biological determinants of P300: an integrative review. *Biological psychology, 41*(2), 103-146.

Proudfit, G. H. (2015). The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology, 52*(4), 449-459.

Ribas-Fernandes, J. J., Solway, A., Diuk, C., McGuire, J. T., Barto, A. G., Niv, Y., & Botvinick, M. M. (2011). A neural signature of hierarchical reinforcement learning. *Neuron, 71*(2), 370-379.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science, 275*(5306), 1593–1599.

Sato, A., Yasuda, A., Ohira, H., Miyawaki, K., Nishikawa, M., Kumano, H., & Kuboki, T. (2005). Effects of value and reward magnitude on feedback negativity and P300. *Neuroreport, 16*(4), 407–411.sa

Skinner, B. F. (1956). A case history in scientific method. *American Psychologist, 11*(5), 221-233. doi:10.1037/h0047662

Skinner, B. F. (1958). Reinforcement today. *American Psychologist, 13*(3), 94-99. doi:10.1037/h0049039

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

Thorndike, E. L. (1935). *The psychology of wants, interests, and attitudes*. New York: Appleton-Century-Crofts.

Toyomaki, A., & Murohashi, H. (2005). The ERPs to feedback indicating monetary loss and gain on the game of modified "rock–paper–scissors." *International Congress Series, 1278*, 381–384. http://doi.org/10.1016/j.ics.2004.11.032

Vogel, E. K., & Luck, S. J. (2000). The visual N1 component as an index of a discrimination process. *Psychophysiology, 37*(02), 190-203.

Wu, Y., & Zhou, X. (2009). The P300 and reward valence, magnitude, and expectancy in outcome evaluation. *Brain Research, 1286*, 114–122. http://doi.org/10.1016/j.brainres.2009.06.032

Yamada, H., Inokawa, H., Matsumoto, N., Ueda, Y., Enomoto, K., & Kimura, M. (2013). Coding of the long-term value of multiple future rewards in the primate striatum. *Journal of neurophysiology, 109*(4), 1140-1151.

Yeung, N., & Sanfey, A. G. (2004). Independent Coding of Reward Magnitude and Valence in the Human Brain. *The Journal of Neuroscience, 24*(28), 6258–6264. http://doi.org/10.1523/JNEUROSCI.4537-03.2004

Yu, R., & Zhou, X. (2006). Brain responses to outcomes of one's own and other's performance in a gambling task. *Neuroreport, 17*(16), 1747-1751.