Neural Correlates of Temporal Credit Assignment

Matthew M. Walsh (mmw187@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University, 342C Baker Hall
Pittsburgh, PA 15213

John R. Anderson (ja@cmu.edu)
Department of Psychology, Carnegie Mellon University, 345D Baker Hall
Pittsburgh, PA 15213

Abstract

When feedback follows a sequence of decisions, how do people assign credit to intermediate actions within the sequence? To explore this temporal credit assignment problem, we recorded event-related potentials (ERPs) as participants performed a sequential decision task. Our ERP analyses focused on feedback-related negativity (FRN), a component thought to reflect neural reward prediction error. The experiment showed that FRN followed negative feedback and negative intermediate states. This outcome suggests that participants evaluated intermediate states in terms of expected future reward, and that these evaluations guided acquisition of earlier actions within sequences. We compared these results to the predictions of three reinforcement learning models that address temporal credit assignment: Actor-critic, Q-Learning, and SARSA.

Keywords: Actor-critic; ERP; Q-Learning; SARSA; Temporal credit assignment; Temporal difference learning.

Introduction

To behave adaptively, humans and animals must learn to predict the outcomes of their actions. Reinforcement learning (RL) provides a mechanism for acquiring this knowledge through trial-and-error interactions with an environment (Sutton & Barto, 1998). According to many RL models, the difference between expected and actual outcomes, or “reward prediction error”, provides a learning signal. By revising estimates based on prediction error, humans and animals learn to anticipate outcomes, and consequently, to select actions that maximize reward and minimize punishment.

RL methods have influenced contemporary neuroscientific theories. For example, one popular RL method, temporal difference (TD) learning, has been used to characterize the phasic response of midbrain dopamine neurons to rewarding and punishing events (Schultz, Dayan, & Montague, 1997). Several studies have confirmed that the response of these neurons depends on reward magnitude and reward likelihood (Tobler, Fiorillo, & Schultz, 2005). Rather than responding directly to experienced outcomes, however, these neurons respond to the *difference* between expected and actual rewards. Thus, midbrain dopamine neurons convey information about TD prediction error.

Recent ERP research with humans has revealed a frontocentral negative component that appears 200-300 ms after the display of error feedback (Gehring & Willoughby, 2002; Miltner, Braun, & Coles, 1997). Three features of this feedback-related negativity (FRN) indicate that it too reflects neural reward prediction error. First, FRN is larger after unexpected than expected outcomes (Holroyd et al., 2009). Second, FRN correlates with behavioral adjustment (Cohen & Ranganath, 2007). Third, neuroimaging experiments, source localization studies, and single cell recordings suggest that FRN originates from the anterior cingulate cortex (ACC), a region implicated in goal-direct behavioral selection (Holroyd et al., 2009). These ideas have been synthesized in the reinforcement learning theory of the error-related negativity (RL-ERN), which proposes that midbrain dopamine neurons transmit a prediction error signal to the ACC, and that this signal strengthens or weakens the actions that precipitated outcomes (Holroyd & Coles, 2002).

Although the RL-ERN theory has stimulated a great deal of research (for review, see Nieuwenhuis et al., 2004), feedback immediately follows actions in most studies of FRN. Similarly, although RL methods have stimulated a great deal of psychological research (for review, see Fu & Anderson, 2006), most studies of RL in humans also involve relatively simple tasks. These scenarios contrast with complex control problems we face in daily life. One such problem is temporal credit assignment. When feedback follows a sequence of decisions, how should credit be assigned to intermediate actions within the sequence?

Here, we consider three TD learning methods that address the temporal credit assignment problem: Actor-critic, Q-Learning, and SARSA. These methods evaluate actions in terms of immediate and future reward. For example, an action may bring an individual into direct contact with reward. Alternatively, an action may bring an individual into a state associated with a high probability of future reward. How should future reward be calculated? In the actor-critic model, future reward is treated as the value of potential options, weighted according to the probability of selecting each (Sutton & Barto, 1998). In Q-Learning, future reward is treated as the value of the best potential option (Watkins & Dayan, 1992). Finally, in SARSA, future reward is treated as the value of the future option that is actually selected (Rummery & Niranjan, 1994).

In the current experiment, we recorded ERPs as participants performed a sequential decision task. The initial decision in each sequence brought participants to an intermediate state associated with a high or a low probability of receiving positive feedback, and the final decision was followed by positive or negative feedback. Based on the idea that FRN reflects neural prediction error

(Holroyd & Coles, 2002), we tested two main hypotheses. First, FRN should be greater for unexpected than for expected outcomes. This follows from the fact that RL models anticipate probable outcomes. Consequently, model prediction error is greater for unexpected than for expected outcomes. Second, if credit assignment occurs "on the fly", as predicted by the TD model, negative feedback *and* negative intermediate states will evoke FRN. Alternate methods exist for performing temporal credit assignment (e.g. model-based RL, eligibility traces). If credit is only assigned at the end of the decision episode, as predicted by these alternate models, only negative feedback will evoke FRN. In addition to testing these two hypotheses, we compared predictions of three TD models, actor-critic, Q-Learning, and SARSA, to the behavioral and neural results of the experiment.

Experiment

Task

A pair of letters appeared at the start of each trial. A cue appeared after participants selected a letter. A second pair of letters followed the cue. Feedback appeared after participants selected a second letter. Participants completed 2 experiment blocks of 400 trials. 13 graduate and undergraduate students participated in the experiment.

Within each block, one pair of letters appeared at the start of all trials (Figure 1). When participants chose the correct letter in the first pair ("J" in this example), a positive and a negative cue appeared equally often. When they chose the incorrect letter in the first pair ("R"), a negative cue always appeared. A second pair of letters followed the cue. The correct letter in the second pair depended on the cue identity. The correct letter for the positive cue ("V" in this example) was rewarded with 80% probability, and the correct letter for the negative cue ("T") was rewarded with 20% probability. Incorrect letters were never rewarded. Consequently, optimal selections yielded positive feedback for 80% of trials involving the positive cue (0.8 Cue) and for 20% of trials involving the negative cue (0.2 Cue). The symbols "#" and "*" denoted positive and negative feedback.

Recording

The EEG was recorded from 32 Ag–AgCl sintered electrodes (10–20 system), and recordings were algebraically re-referenced offline to the average of the right and left mastoids. The vertical EOG was recorded as the potential between electrodes placed above and below the left eye, and the horizontal EOG was recorded as the potential between electrodes placed at the external canthi. The EEG and EOG signals were amplified by a Neuroscan bioamplification system with a bandpass of 0.1-70 Hz and digitized at 250 Hz. Eye blinks were corrected using ICA. 800 ms epochs were extracted from the continuous recording and these epochs were baseline corrected relative to the 200 ms prestimulus interval.

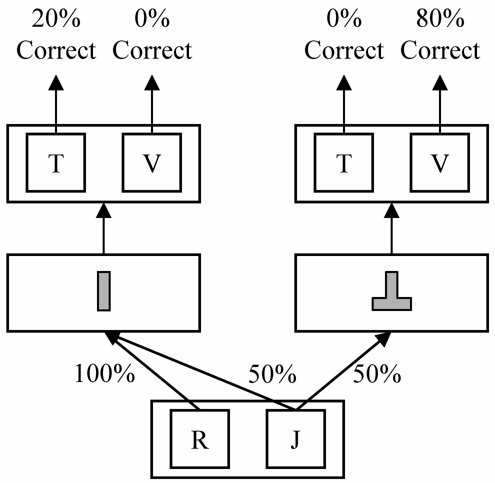

Figure 1. Experiment states, transition probabilities, and outcome likelihoods.

Feedback-locked ERPs were analyzed for trials where participants selected the correct letter for the cue, and FRN was calculated as the difference between ERP waveforms after losses and wins. FRN amplitude is often confounded by changes in P300 amplitude, a component that is also sensitive to event likelihoods. Consequently, we compared losses and wins that were equally likely by creating an "expected outcome" difference wave (0.2 Cue losses – 0.8 Cue wins), and an "unexpected outcome" difference wave (0.8 Cue losses – 0.2 Cue wins). FRN was measured as mean voltage of the difference waves from 200-300 ms after feedback onset, relative to the 200 ms prestimulus baseline. Cue-locked ERPs were analyzed for trials where participants selected the correct starting letter (after which the probability of receiving the 0.2 or the 0.8 Cue was equal). Cue FRN was measured as mean voltage of the cue difference wave (0.2 Cue – 0.8 Cue) from 200-300 ms after cue onset.

Models

Actor-critic (Sutton & Barto, 1998)

The actor-critic (AC) model computed a state-action value function, $Q(s,a)$, and a state value function, $V(s)$. The state-action value function, which corresponded to the actor, enabled action selection. The state-value function, which corresponded to the critic, enabled evaluation of action consequences. Actions affected the transition from state s_t to s_{t+1}, and actions affected the presentation of reward, r_{t+1}. Following the selection of an action, a_t, the critic issued an evaluation in the form of prediction error, δ,

$$\delta = \left[r_{t+1} + \gamma \bullet V\left(s_{t+1}\right)\right] - V\left(s_t\right). \qquad (1)$$

The AC model maximized the combined immediate, r_{t+1}, and future reward, $V(s_{t+1})$, and future reward was discounted by γ ($\gamma < 1.0$). The value of the previous state, $V(s_t)$, was updated according to

$$V(s_t) \leftarrow V(s_t) + \alpha \bullet \delta, \qquad (2)$$

where α controlled the learning rate ($0.0 < \alpha < 1.0$). The value of the previous state-action pair, $Q(s_t,a_t)$, was updated according to

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \bullet \delta\,. \qquad (3)$$

Q-Learning (Watkins & Dayan, 1992)

The AC and Q-Learning models differed in two ways. First, the Q-Learning model used an action-state value function, $Q(s,a)$, to select actions and to evaluate outcomes. Second, the Q-Learning model treated future reward as the value of the optimal selection policy in state $t + 1$,

$$\delta = \left[r_{t+1} + \gamma \bullet \max_a Q(s_{t+1}, a)\right] - Q(s_t, a_t)\,. \qquad (4)$$

As in the AC model, future reward was discounted by γ, and the state-action value function was updated according to Equation 3.

SARSA (Rummery & Niranjan, 1994)

Like the Q-Learning model, the SARSA model only required an action-state value function, $Q(s,a)$. Unlike the Q-Learning model, however, the SARSA model treated future reward as the value of the actual state-action pair selected in state $t + 1$,

$$\delta = \left[r_{t+1} + \gamma \bullet Q(s_{t+1}, a_{t+1})\right] - Q(s_t, a_t). \qquad (5)$$

As with the AC and Q-Learning models, future reward was discounted by γ, and the state-action value function was updated according to Equation 3.

To summarize, all models used δ to learn the values of the state-action pairs that comprised the experiment task (Figure 1), and all models sought to select actions that maximized immediate and future reward. Although the initial selection in each trial was not followed by immediate reward (i.e. $r_{t+1} = 0$), the initial selection was followed by future reward associated with a subsequent state (AC model), or a subsequent state-action pair (Q-Learning and SARSA models). As such, prediction error for the initial selection was calculated as the difference between discounted future reward and the value of the first state (AC model), or the value of the first state-action pair (Q-Learning and SARSA models). Prediction error for the final selection was calculated as the difference between immediate reward and the value of the second state (AC model), or the value of the second state-action pair (Q-Learning and SARSA models). Positive feedback had a value of 1.0 and negative feedback had a value of 0.0[1].

Model predictions were based on 500 simulations. All state and state-action pairs began with values of 0.5 and values were updated according to prediction error. In each trial, logistically distributed noise was added to state-action values, and the state-action pair with the greatest value was selected. Two model parameters, learning rate ($\alpha = .05$) and the temporal discounting factor ($\gamma = 0.8$), were fixed according to values reported in Fu & Anderson (2006). Interestingly, when α and λ were treated as free parameters, mean squared error (MSE) for each model was minimized at values of α and γ within ±0.02 of their fixed values. Selection noise (t, defined as the standard deviation of the logistically distributed noise added to state-action pairs) remained as a free parameter. We compared model selections to participant performance. Additionally, we computed the difference in δ for expected feedback (0.2 Cue losses – 0.8 Cue wins), unexpected feedback (0.8 Cue losses – 0.2 Cue wins), and cues (0.2 Cue – 0.8 Cue) to derive model FRN. We then fit model FRN to observed FRN using a slope term (m) and a zero intercept.

Results

Behavioral Results

Selection accuracy varied by choice, $F(2,24) = 10.33$, $p < .001$, and selection accuracy increased by block half, $F(1,12) = 102.54$, $p < .0001$ (Figure 2). Selection times for correct responses did not vary by choice, $F(2,24) = 2.47$, $p > .1$, or block half, $F(1,12) = 2.55$, $p > .1$.

ERP Results

We first analyzed feedback-locked ERPs. Waveforms showed a pronounced negativity from 200-300 ms after loss feedback (Figure 3). This FRN (loss – win) appeared to be greater for unexpected than for expected outcomes. A 3 (site: Fz, Cz, Pz) by 2 (outcome likelihood: expected, unexpected) ANOVA on FRN amplitude revealed effects of site, $F(2,24) = 10.91$, $p = .005$, and outcome likelihood, $F(1,12) = 13.26$, $p = .003$. FRN was greater for unexpected than for expected outcomes at site Fz, $t(12) = 3.69$, $p = .003$. We also considered FRN over the first and the second halves of blocks (Figure 5). A 2 (outcome likelihood) by 2 (block half) ANOVA at Fz showed an effect of outcome likelihood, $F(1,12) = 11.68$, $p = .005$, but not block half, $F(1,12) = 0.10$, $p > .1$. Although the interaction was not significant, $F(1,12) = 3.13$, $p > .1$, experience caused FRN to increase for unexpected outcomes and to decrease for expected outcomes[2].

We then analyzed cue-locked ERPs. A 3 (site) by 2 (cue) ANOVA revealed a nonsignificant effect of site, $F(2,24) =$

[1] Because the model used a soft-max decision policy, choice proportions depended only on the absolute differences between Q-values. Consequently, changes to the noise parameter, t, can accommodate a wide range of positive and negative reward values.

[2] In a subsequent experiment with a larger sample size, this interaction reached significance.

0.24, $p > .1$, a marginal effect of cue, $F(1,12) = 3.05$, $p = .1$, and a nonsignificant interaction, $F(2,24) = 1.78$, $p > .1$. ERPs were relatively more negative for 0.2 than for 0.8 Cues at site Fz, but the effect failed to reach significance, $t(12) = 1.77$, $p = .1$. When we considered the first and the second halves of blocks separately, however, a different picture emerged (Figure 4). A 2 (cue) by 2 (block half) ANOVA at Fz revealed a significant interaction between cue and block half, $F(1,12) = 6.56$, $p = .025$. In the first half of blocks, ERPs did not vary by cue, $t(12) = .46$, $p > .1$, but in the second half of blocks, ERPs were relatively more negative for 0.2 than for 0.8 Cues, $t(12) = 2.76$, $p = .017$. The discovery of cue FRN indicates that participants evaluated intermediate outcomes in terms of future reward, as predicted by the temporal difference models.

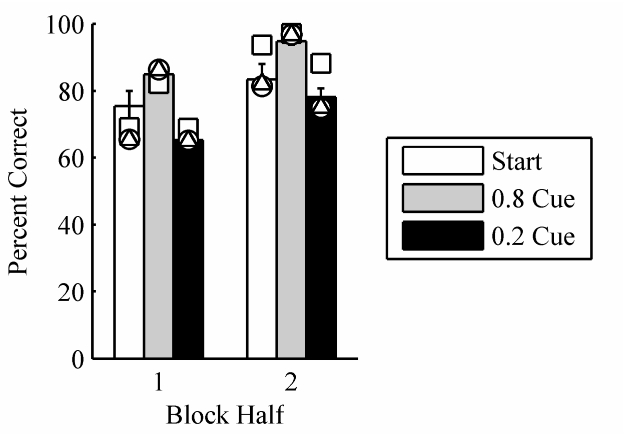

Figure 2. Selection accuracy for start pair, 0.8 Cues, and 0.2 Cues by block half and for participants (bars), AC (squares), Q-Learning (circles), and SARSA (triangles).

Model Performance

For each model, we estimated the value of noise, t, that best accounted for selection accuracy over the first and second halves of experiment blocks. For the Q-Learning and SARSA models, MSE was minimized at $t = 0.1$ (Q-Learning: MSE = 0.002, $r^2 = 0.90$; SARSA: MSE = 0.002, $r^2 = 0.90$). For the AC model, MSE was minimized at $t = 0.2$ (MSE = 0.004, $r^2 = 0.71$). As seen in Figure 2, all models displayed effects of choice and block half like those seen for participants. Additionally, the Q-Learning and SARSA models, which were structurally most similar, yielded nearly identical predictions to one another ($r^2 = 0.99$). Finally, the AC model outperformed participants and the other two models over the second half of blocks.

Next, we examined whether FRN related to model δ. To do so, we computed model FRN as the difference in δ for expected feedback, unexpected feedback, and cues. For each model, we estimated the value of the slope parameter, m, that best accounted for FRN over the first and second halves of experiment blocks. For the Q-Learning and SARSA models, MSE was minimized at $m = 2.6$ (Q-Learning: MSE = 0.295, $r^2 = 0.85$; SARSA: MSE = 0.294, $r^2 = 0.85$). For the AC model, MSE was also minimized at $m = 2.6$ (MSE = 0.262, $r^2 = 0.86$). As seen in Figure 5, all models predicted that cue FRN would increase with experience, and that FRN for unexpected outcomes would increase with experience while FRN for expected outcomes would decrease with experience. These trends were observed.

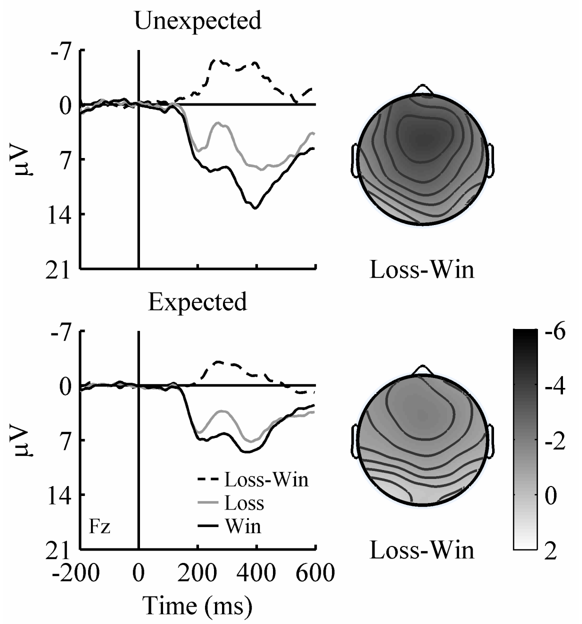

Figure 3. ERPs evoked by unexpected and expected losses and wins at site Fz (left panels). Scalp voltage topography for loss – win comparison from 200-300 ms (right panels).

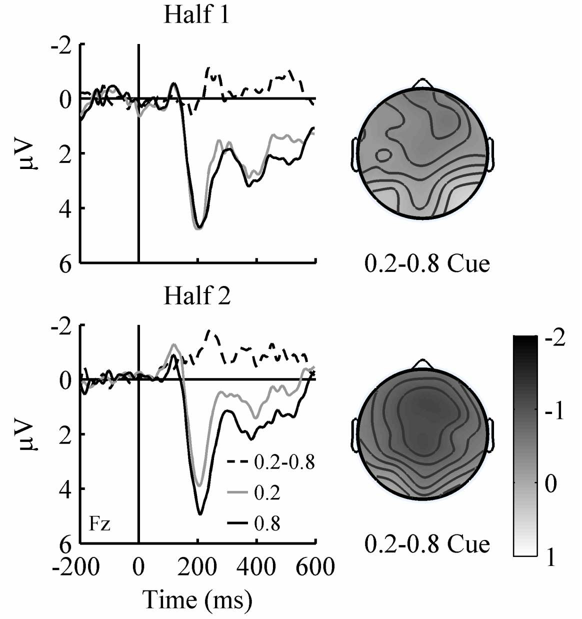

Figure 4. ERPs evoked by 0.2 and 0.8 Cues for the first and the second halves of blocks at site Fz (left panels). Scalp voltage topography for 0.2 Cue – 0.8 Cue comparison from 200-300 ms (right panels).

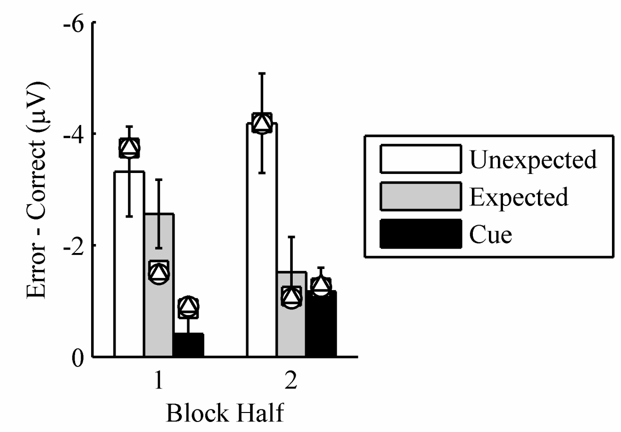

Figure 5. FRN for unexpected outcomes, expected outcomes, and cues by block half and for participants (bars), AC (squares), Q-Learning (circles), and SARSA (triangles).

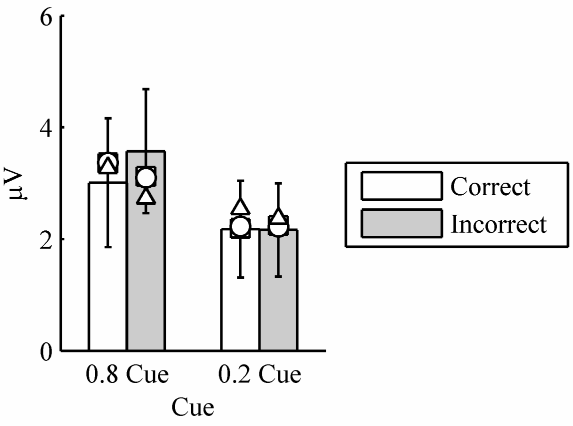

Figure 6. Cue-locked voltages preceding correct and incorrect responses by cue and for participants (bars), AC (squares), Q-Learning (circles), and SARSA (triangles).

The behavioral results favored the Q-Learning and SARSA models. The AC model outperformed participants and the other two models over the second half of blocks. Performance differences between models related to the nuanced meaning of state-action pairs, $Q(s,a)$, for each. In the Q-Learning and SARSA models, Q-values approximate values of state-action pairs. In the AC model, Q-values approximate selection preferences that maximize the state-value function, V(s). Because a deterministic selection policy maximized the state-value function, V(s), in our task, Q-values in the AC model became increasingly polarized until near-deterministic selections emerged. The same effect could be achieved in the Q-Learning and SARSA models by annealing the noise parameter.

To further distinguish between the Q-Learning and SARSA models, we re-analyzed cue-locked waveforms based on cue identity (0.2 Cue, 0.8 Cue) and the response that followed the cue. If prediction error depended on the value of future actions, as predicted by SARSA, we expected that cue-locked waveforms would be more negative before participants chose the incorrect response than before they chose the correct response. From 200-300 ms after cue presentation, average area under the 0.2 Cue waveform was less than area under the 0.8 Cue waveform at site Fz, $F(1,12) = 8.40$, $p = .013$ (Figure 6). Waveforms did not depend on the accuracy of the forthcoming response, however, $F(1,12) = 1.71$, $p > .1$.

We computed model δ for the same combination of factors[3]. Q-Learning and AC predictions were consistent with observations (Q-Learning: MSE = 0.237, $r^2 = .0.77$; AC: MSE = 0.225, $r^2 = .0.76$) in that they predicted an effect of cue but not response accuracy. In contrast, the SARSA model predicted a more negative signal before incorrect than correct responses (MSE = 0.419, $r^2 = .0.32$), owing to how the algorithm computed future reward (Eq. 5).

General Discussion

Although the RL-ERN theory has stimulated a great deal of research, feedback immediately follows actions in most studies of FRN. Similarly, although RL methods have stimulated a great deal of psychological research, most studies of RL in humans involve simple environments. In the current experiment, we examined learning in a more complex problem space. We asked how people assign credit to intermediate actions when making sequences of decisions.

The experiment yielded two clear results. First, FRN was greater for unexpected than for expected outcomes. Although some studies have reported a relationship between FRN and prediction error (Holroyd et al., 2009), others have not (Hajcak et al., 2005). This discrepancy has led to the proposal that FRN relates most strongly to prediction error when outcomes are contingent on behavior (Holroyd et al., 2009). In our experiments, feedback was contingent on behavior, and consistent with the proposal of Holroyd et al. (2009), we did observe a relationship between prediction error and FRN. Second, FRN also followed negative intermediate outcomes even though these outcomes did not directly signal reward. This result shows that people evaluated intermediate outcomes in terms of expected future reward. Although many theories propose that such evaluations underlie temporal credit assignment (Fu & Anderson, 2006; Holroyd & Coles, 2002; Schultz, Dayan, & Montague, 1997; Sutton & Barto, 1998), these results provide one of the clearest demonstrations of TD learning to our knowledge.

We also examined three TD methods: Actor-critic, Q-Learning, and SARSA. A recent neuroimaging study provided support for the AC model by showing that activity in the dorsal and ventral striatum of the basal ganglia corresponded to the behavior of the actor and the critic in the AC model (O'Doherty et al., 2004). Alternatively, single-cell recordings from midbrain dopamine neurons in monkeys have supported SARSA (Morris et al., 2006), and

[3] This analysis was based on the area under individual waveforms rather then FRN. Consequently, we computed new slope and intercept terms to compare model δ to observations.

recordings from dopamine neurons in rats have supported Q-Learning (Roesch, Calu, & Schoenbaum, 2007). An integrative account of these findings is hindered by the between species comparison. Consequently, it is unclear, as of yet, which form of TD control is most applicable to humans. The behavioral and neural results of the current experiment were consistent with Q-Learning. This considerations not withstanding, the current data do not definitively distinguish between TD variants. The more valuable contribution of this work is the demonstration that intermediate states inherit value, a feature central to each TD model. Future studies should aim to elucidate the precise TD algorithms that underlie neurological computations.

Our simulations demonstrated that the core Q-Learning model could account for the behavioral and neural data. Additionally, our computational instantiation clarified two nuanced features of the experiment results. First, FRN decreased for expected outcomes and increased for unexpected outcomes. Model FRN changed in the same manner. Because utility estimates began at 0.5, δ was initially -0.5 (0.0 – 0.5) for all losses, and δ was initially 0.5 (1.0 – 0.5) for all wins. As the model learned, the utility of the correct response for the 0.2 Cue approached 0.2 and the utility of the correct response for the 0.8 Cue approached 0.8. Consequently, δ magnitude decreased for expected wins and losses, and δ magnitude increased for unexpected wins and losses, giving rise to the observed changes in FRN.

Second, cue FRN increased with experience. The Q-Learning model (and in fact all TD models) also showed an experience-dependent increase in cue FRN. The models only distinguished between positive and negative cues after the values of the states and actions that followed those cues (e.g. future reward) became polarized. As this result demonstrates, the TD models learn the utility of actions that are near to rewards before learning the utility of actions that are far from rewards. Humans and animals also exhibit this learning gradient (Fu & Anderson, 2006).

Do the results of this experiment indicate that TD methods alone are sufficient for coping with temporal credit assignment? We think not. Although participants faced a discrete Markov decision process (MDP) in our experiment, people must sometimes identify current states *and* recall past transitions. Violations of the Markov property may be problematic for TD methods. Additionally, although TD learning reduces the delay between action selection and credit assignment, TD learning does not typically eliminate delays in continuous time domain tasks. An important question for future research is how people integrate TD learning with other RL methods, like eligibility traces and model-based RL, to behave proficiently in complex environments.

Acknowledgments

This work was supported by training grant T32GM081760 and NIH training grant MH 19983 to the first author and NIMH grant MH068243 to the second author.

References

Cohen, M.X., & Ranganath, C. (2007). Reinforcement learning signals predict future decisions. *The Journal of Neuroscience, 27,* 371-378.

Fu, W.T., & Anderson, J.R. (2006). From recurrent choice to skill learning: a reinforcement-learning model. *Journal of Experimental Psychology: General, 135,* 184-206.

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

Hajcak, G., Holroyd, C.B., Moser, J.S., & Simons, R.F. (2005). Brain potentials associated with expected and unexpected good and bad outcomes. *Psychophysiology, 42,* 161-170.

Holyroyd, 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,* 679-709.

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,* 59-70.

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,* 788-798.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., & Bergman, H. (2006). Midbrain dopamine neurons encode decisions for future action. *Nature Neuroscience, 9,* 1057-1063.

Nieuwenhuis, S., Holroyd, C.B., Mol, N., & Coles, M.G.H. (2004). Reinforcement-related brain potentials from medial frontal cortex: origins and functional significance. *Neuroscience and Biobehavioral Reviews, 28,* 441-448.

O’Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R.J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science, 304,* 452-454.

Roesch, M.R., Calu, D.J., & Schoenbaum, G. (2007). Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature Neuroscience, 10,* 1615-1624.

Rummery, G.A., & Niranjan, M. (1994). On-line Q-learning using connectionist systems. (Tech. Rep. CUED/F-INFENG/TR166). Cambridge University.

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

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

Tobler, P.N., Fiorillo, C.D., & Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science, 307,* 1642-1645.

Watkins, C.J., & Dayan, P. (1992). Q-learning. *Machine Learning, 8,* 279-292.