# How We Learn to Make Decisions: Rapid Propagation of Reinforcement Learning Prediction Errors in Humans

Olav E. Krigolson[1], Cameron D. Hassall[1], and Todd C. Handy[2]



## Abstract

■ Our ability to make decisions is predicated upon our knowledge of the outcomes of the actions available to us. Reinforcement learning theory posits that actions followed by a reward or punishment acquire value through the computation of prediction errors—discrepancies between the predicted and the actual reward. A multitude of neuroimaging studies have demonstrated that rewards and punishments evoke neural responses that appear to reflect reinforcement learning prediction errors [e.g., Krigolson, O. E., Pierce, L. J., Holroyd, C. B., & Tanaka, J. W. Learning to become an expert: Reinforcement learning and the acquisition of perceptual expertise. *Journal of Cognitive Neuroscience, 21,* 1833–1840, 2009; Bayer, H. M., & Glimcher, P. W. Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron, 47,* 129–141, 2005; O'Doherty, J. P. Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology, 14,* 769–776, 2004; Holroyd, C. B., & Coles, M. G. H. The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109,* 679–709, 2002]. Here, we used the brain ERP technique to demonstrate that not only do rewards elicit a neural response akin to a prediction error but also that this signal rapidly diminished and propagated to the time of choice presentation with learning. Specifically, in a simple, learnable gambling task, we show that novel rewards elicited a feedback error-related negativity that rapidly decreased in amplitude with learning. Furthermore, we demonstrate the existence of a reward positivity at choice presentation, a previously unreported ERP component that has a similar timing and topography as the feedback error-related negativity that increased in amplitude with learning. The pattern of results we observed mirrored the output of a computational model that we implemented to compute reward prediction errors and the changes in amplitude of these prediction errors at the time of choice presentation and reward delivery. Our results provide further support that the computations that underlie human learning and decision-making follow reinforcement learning principles. ■

## INTRODUCTION

Everyday we are faced with a myriad of decisions—simple decisions like what to eat or complex decisions like whether or not we should get married. But how do we actually decide what we want to do? Utilitarianism (Mill, 1879) suggests that we have an inherent desire to seek out and maximize rewarding outcomes, and thus, we tend to make decisions that maximize utility. For example, consider the choice of career. Do we choose a career doing something that we are passionate about—for instance, academia? Or, do we pursue a career such as medicine that would be more lucrative? Obviously, everyone has their own opinion about the relative value of these particular choices, but utilitarian theory suggests that each of us makes a choice that we believe maximizes utility. Indeed, our ability to decide what we want to do with our lives or to make any other decision for that matter is predicated upon our knowledge of the value of the actions available to us (Bernoulli, 1713).

So what does the value of an action reflect and how do we learn the value of an action? Reinforcement learning (RL) theory proposes that the value of an action is a prediction of the subsequent reward or punishment gained by selecting that action (Sutton & Barto, 1998; Rescorla & Wagner, 1972). Rescorla and Wagner also proposed that the value of an action is updated following a decision that leads to a reward or punishment. For example, imagine we are in a state Q with two choices—A and B. Initially, the values for these choices will be zero—we have no information about the outcome of choosing A or B, and thus the prediction of reward (or punishment) associated with each choice is zero. However, if we choose A and are then rewarded, we compute a prediction error—the discrepancy between the actual value of the reward and the predicted value of the reward. Importantly, the prediction error is then used to modify the value of choice A, such that over time, the value of choice A comes to accurately reflect the reward

[1]Dalhousie University, [2]University of British Columbia



*Journal of Cognitive Neuroscience 26:3, pp. 635–644*
*doi:10.1162/jocn_a_00509*

gained by making this choice. In other words, in the initial stages of learning one should observe prediction errors when rewards or punishments are encountered, as there will be discrepancies between the actual value of the reward and the value of the action (i.e., the prediction of reward: Rescorla and Wagner [1972] and Sutton and Barto [1998]). With learning, however, the magnitude of the prediction error computed at the time of reward delivery will diminish as the predicted value of reward (i.e., the value of the action) comes to approximate the actual reward value.

As the values of the actions available to us increase or decrease with learning, the value of the choice state where one can select potential actions also changes. Simply put, moving into a state that has actions with known values means that one is in a position to select an action that leads to a reward—or avoid an action that leads to a punishment. During learning, the value of the choice state is also changed by the reward prediction errors that are used to modify the values of actions. Thus, the value of the choice state also comes to reflect a prediction of the reward (or punishment) that can be gained by selecting actions within that state. Recall the previous example where we are in a state Q with two potential choices, A and B. If we choose A and are rewarded, we increase the value of the action, A, but we also increase the value of the choice state Q. Subsequently, when we encounter choice state Q after learning has occurred a prediction error is computed as we will have moved from a state with no value—the state before Q—into a state with value—state Q. Computational RL theories such as the method of temporal differences (Sutton & Barto, 1998) take this into account and posit that prediction errors are computed as the difference between the value and rewards of the current state and the value of the previous state. Consider another example—driving home from work. The actual reward is getting home, but arriving at an intersection close to home can be thought of as rewarding as it means we are in a state where we can select an action that will get us home. One can also summarize this as follows: Prediction errors occurring at the earliest indicator events are going to be better or worse than expected (i.e., anytime an agent moves into a state with greater or lesser value; Holroyd & Coles, 2002). Early in learning prediction errors occur at the time of reward delivery as the value of the choice state does not accurately reflect the value of the subsequent reward or punishment. However, after learning has occurred, prediction errors occur when we move into a choice state that has value from a prior state without value.

Studies in monkey measuring changes in the phasic firing rate of dopaminergic neurons in the substansia nigra pars compacta in classical conditioning experiments provide empirical support for the predictions of RL theory. Seminal work by Schultz, Dayan, and Montague (1997) demonstrated that, when monkeys are initially given a reward, there is an associated phasic increase in the firing rate of dopaminergic neurons in the substansia nigra pars compacta. However, Schultz and colleagues also observed that when a reward was consistently paired with a predictive stimulus the phasic increase in dopamine firing rate observed at the time of reward delivery diminished over time and instead a phasic increase in dopamine firing rate was observed shortly after the onset of the predictive stimulus. RL theory specifically predicts this pattern of results. First, a prediction error should be computed early in learning for unexpected rewards as the value of the cue state did not predict the value of the reward. Second, the prediction error at the time of reward should diminish with learning as the value of the cue state approaches the value of the reward state—the difference between these states becomes zero and thus there is no error in prediction. Third, a prediction error should be observed at cue onset after learning has occurred as the monkey has moved from a state with no value—the state before the cue—to a state with value—the cue state. In summary, the pattern of changes in the dopaminergic response to the predictive cue and the reward mirrored the predictions of Rescorla and Wagner—prediction errors at the time of reward diminished and prediction errors at stimulus presentation increased with learning.

Studies in human observing the neural response to feedback have demonstrated a pattern of results similar to the theoretical predictions of Rescorla and Wagner (1972) and the results observed in monkey by Schultz and colleagues (1997). Specifically, in a series of experiments, Holroyd and colleagues (Holroyd, Pakzad-Vaezi, & Krigolson, 2008; Holroyd & Krigolson, 2007; Holroyd & Coles, 2002) have demonstrated that the amplitude of the feedback error-related negativity (fERN), a component of the human brain ERP, is sensitive to reward expectancy and further, that it only occurs in situations when participants must rely on feedback to determine response outcome. In other words, a fERN is only observed when one moves from a state with no value into a state with either positive or negative value. Extending this, Krigolson, Pierce, Holroyd, and Tanaka (2009) found that the magnitude of the fERN at the time of reward delivery diminished with learning—a result that mirrored the aforementioned predictions and results. Unifying a large body of empirical work, Holroyd and Coles (2002) proposed a comprehensive theory that suggested that the fERN reflects the impact of a dopaminergic prediction error signal sent by reward evaluation units within BG (O'Doherty et al., 2004) to response selection processes in ACC (Holroyd, Yeung, Coles, & Cohen, 2005; Holroyd & Coles, 2002) to facilitate the optimization of behavior.

In the present experiment, we sought to demonstrate a rapid shift in the occurrence of prediction errors from feedback delivery to choice presentation with learning. To accomplish this, we recorded ERP data while participants played a simple gambling game where they learned which of two response options yielded a reward. In a key manipulation, sets of distinct gambles were repeated within each experimental block, making the gambling

task completely deterministic and thus learnable. In this manner, we sought to demonstrate that in the early stages of learning a fERN would be elicited at the time of reward delivery. However, in line with RL theory, we predicted that the amplitude of the fERN at the time of reward delivery would diminish with learning. Furthermore, we also predicted that, with learning, we would see an increase in an ERP response at the time of choice presentation reflecting the occurrence of a prediction error at this time. We hypothesized that the characteristics of this signal would be similar to that of the fERN—that it would be maximal 200–300 msec post-stimulus onset with a medial-frontal scalp topography. Finally, to verify that the pattern of our ERP results mirrored the predictions of RL theory, we implemented a computational model that utilized a RL algorithm (cf., Sutton & Barto, 1998) to learn and perform the gambling task.

## METHODS

### Participants

Eighteen undergraduate students (8 men, 10 women; aged 18–30 years) with no known neurological impairments and with normal or corrected-to-normal vision participated in the experiment. All of the participants were volunteers who received extra credit in undergraduate psychology courses at the University of British Columbia for their participation and a financial performance-based reward (see below). The participants provided informed consent approved by Research Services at the University of British Columbia, and the study was conducted in accordance with the ethical standards prescribed in the original (1964) and subsequent revisions of the Declaration of Helsinki.

### Apparatus and Procedure

Participants played a simple gambling game while electroencephalographic data were recorded. On each trial of the game, participants first viewed a fixation cross for 400–600 msec. Following this, two uniquely colored squares appeared, one on either side of the fixation cross. The colors of the squares were picked randomly from a set of 24 distinct colors before each block, and no unique color pair was used more than once. After 1000 msec, the fixation cross changed color from a dark to a lighter shade of gray, an event that signaled the participant to select one of the colored squares by depressing either the left or right button of a standard computer USB gamepad. Subsequent to the button press, the squares disappeared, leaving the fixation cross on the screen for another 400–600 msec. After this interval, participants were provided with feedback—either a "0," "1," or "2" presented centrally on screen for 1000 msec, indicating that they had won either 0, 1, or 2 cents. Following feedback presentation, the screen went blank for 500 msec.

After the first gamble of an experimental block (gamble "A"), participants were presented with a second unique gamble (gamble "B") that was distinguishable from the gamble A via the colors of the two squares. Note that the colors of the squares for the second gamble were picked to ensure they were different and unique from the first gamble. As with the gamble A, participants selected a response for gamble B and then were informed as to whether or not they won. Note that the outcomes of gambles A and B were completely deterministic—in other words, participants were able to use the initial outcome of a gamble to learn the correct response for subsequent gambles. More specifically, for each gamble one outcome was always "0" cents and the other was either "1" or "2" cents (50% probability of each). Furthermore, the mappings between square color, choice, and gamble outcome remained the same throughout each experimental block, thus making the gambles "learnable." From a theoretical perspective, a RL prediction error would occur when the feedback was viewed for the first presentations of gambles A and B as participants moved from a state with no value, the choice state, to a state with value, the reward state.

Following the initial two gambles, participants were randomly presented with either gamble A or gamble B again. Importantly, participants knew the correct response for the presented gamble—assuming they learned from the provided feedback—but they did not know which gamble they would see. As such, according to RL theory, a prediction error should occur upon viewing the choice state for gamble A or gamble B a second time as participants moved from a state with no value, the state before the choice, to a state with value, the known choice state. Following the third gamble, participants completed three more gambles randomly chosen as either gamble A or gamble B for a total of six gambles (three trials of each gamble) to complete an experimental block. In total, participants completed 108 experimental blocks, each containing two unique gambles repeated three times each as outlined above. Payout schedules, square colors, and the side of the correct response (left/right) were all randomly counterbalanced across the experimental blocks. Following each block, participants could rest for as long as they wished before commencing another block. On average, participants won just under $8 CDN playing the gambling game.

### Data Acquisition

Accuracy (correct, incorrect) and the RT (msec) were recorded for each trial by the experimental program as behavioral measures of performance. The EEG was recorded from 40 electrode locations using ActiView software (Alpharetta, GA). The electrodes were mounted in a fitted cap with a standard 10–20 layout and were referenced to a two electrode feedback loop (common mode sense to driven right leg). The vertical and horizontal electro-oculograms were recorded from electrodes placed above and below the

right eye and on the outer canthi of the left and right eyes, respectively. Electrode offsets were kept below ±25 mV at all times. The EEG data were sampled at 256 Hz and amplified with an Active Two system (Biosemi B.V., Amsterdam, Netherlands).

### Data Analysis

We calculated mean accuracy (%) and mean RT (msec) for correct and error trials for each experimental condition and participant as measures of task performance.

The EEG analysis was done as follows. First, the EEG data were filtered offline through a (0.1–25 Hz passband) phase shift free Butterworth filter and rereferenced to an averaged mastoid reference. Next, 800-msec epochs of EEG data were extracted from the continuous EEG locked to cue (when the two colored squares appeared) and feedback stimulus onset (200 msec before the event to 600 msec after). Following this, the ocular artifacts in each epoch were corrected using the algorithm described by Gratton, Coles, and Donchin (1983), and each epoch was baseline-corrected using the mean voltage for the 200 msec preceding feedback stimulus onset. Epochs were then examined for artifacts and removed from the data set if there was a change in voltage on any channel that exceeded 35 μV between adjacent sampling points or a difference of more than 150 μV between the maxima and minima of the epoch. On average, less than 10% of the data were discarded per participant, with two participants' data being completely removed from further analysis due an excessive number of artifacts (more than 80% of the trials were lost). Note that we also analyzed the data in the exact manner described here but without the ocular correction. We did this to ensure that there the ocular correction algorithm did not mask any cue or feedback stimulus onset related blink activity. The results from this analysis mirrored the results reported here, albeit with more noise given the reduced number of trials going into the averaged ERP waveforms.

ERP waveforms were created by averaging the EEG epochs for each event of interest (cue, feedback), gamble repetition (one, two, three), and each reward outcome (win 0 cents, win 1 cent, win 2 cents) for each participant. Observation of the grand averaged waveforms (see Figures 1 and 2) led to a quantification of the fERN as the mean voltage 225–275 msec following the onset of the feedback stimulus. We focused our analysis on channel FCz given previous work (Krigolson et al., 2009; Krigolson, Holroyd, Van Gyn, & Heath, 2008; Holroyd et al., 2008; Holroyd & Krigolson, 2007; Holroyd, Yeung, Coles, & Cohen, 2005) and an examination of the fERN topographies that supported our decision (see Figures 1 and 2). We also decided to examine the P300 evoked by the presentation of the feedback stimulus and quantified this component as the mean voltage 300–450 msec post-stimulus onset. Finally, we were interested in the neural response to the presentation of the gambling cue (i.e., the colored squares). An examination of the grand averaged waveforms confirmed our hypotheses, and we observed a difference in the ERP waveforms consistent with accounts of the fERN, albeit a bit later. As such, we quantified this cue locked reward response as the mean voltage 290–340 msec post-cue onset. Note that at this point there is no indication of the outcome of the gamble, so we were only able to quantify this component with regard to gamble repetition (one, two, three) and not reward outcome.

All analyses were done with EEGLAB (Delorme & Makeig, 2004) and custom code written in the Matlab (MathWorks, Natick, MA) programming environment. Repeated-measures ANOVA and paired samples *t* tests

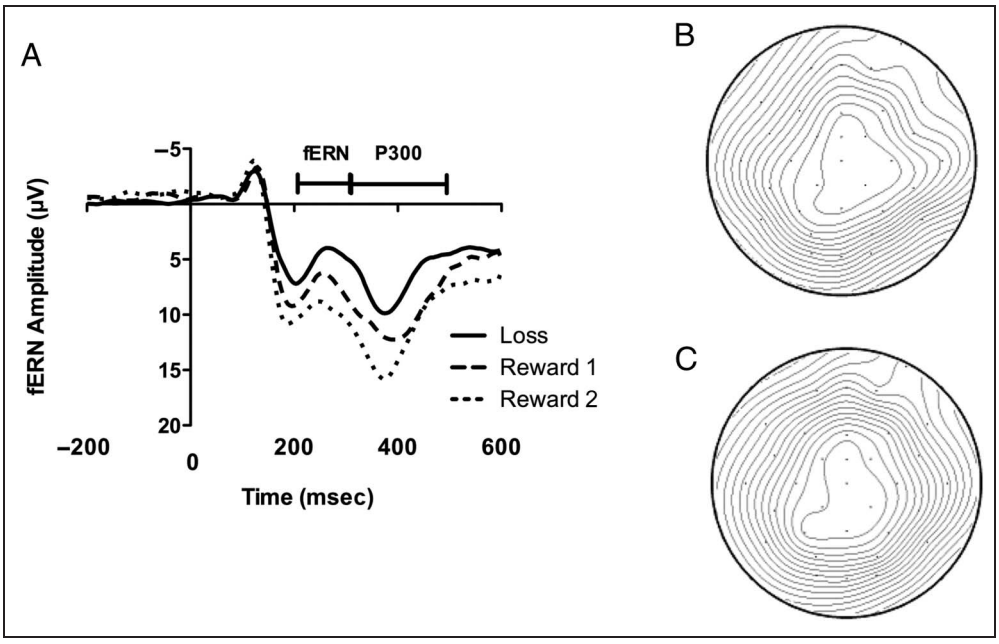


**Figure 1.** (A) Grand averaged ERP waveforms averaged to the time of reward delivery for Trial 1 of the gambling task. (B) The topography of the peak difference between Reward 1 and Reward 0 outcomes in the fERN time range. (C) The topography of the peak difference between Reward 2 and Reward 0 outcomes in the fERN time range.

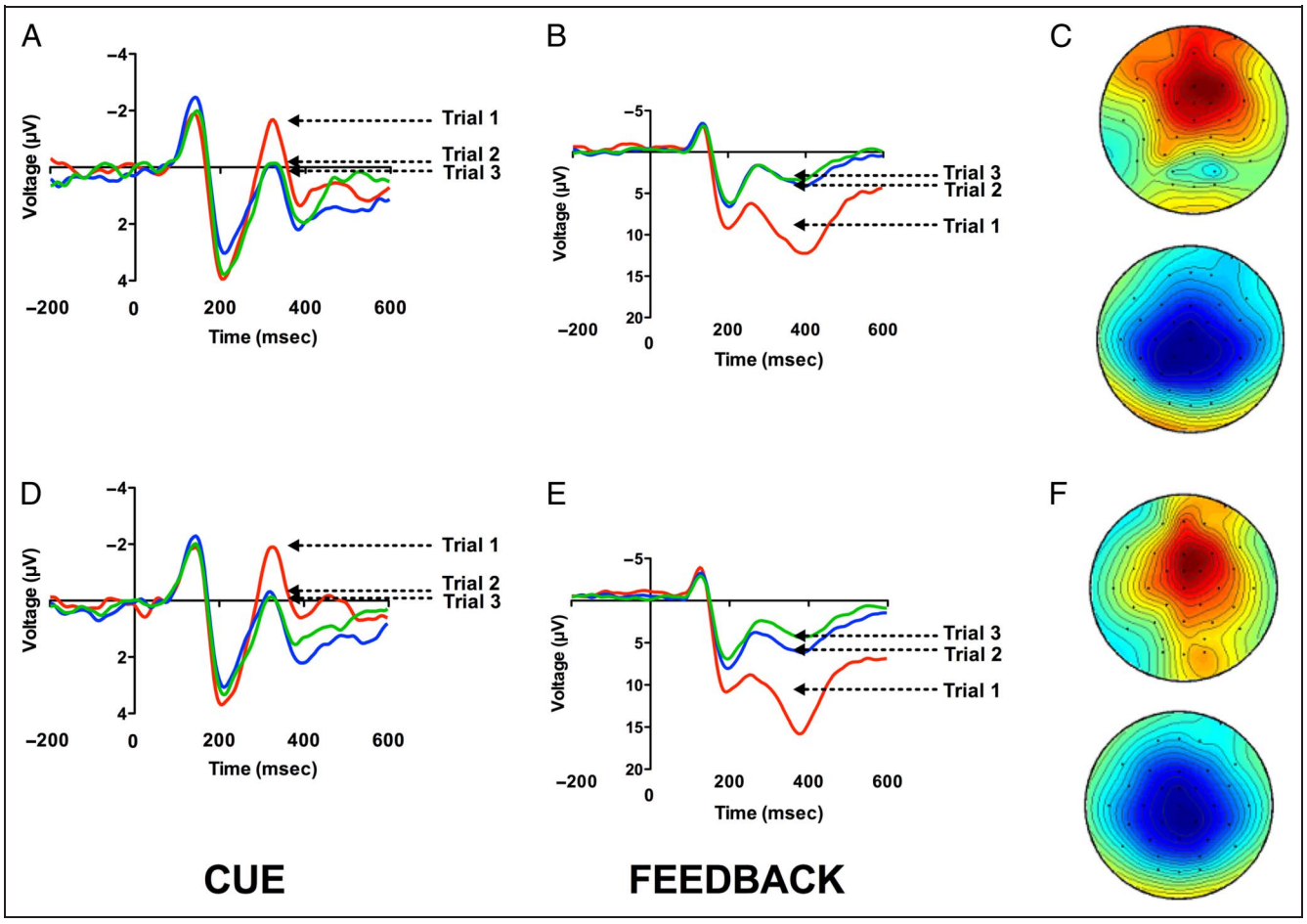


**Figure 2.** (A) Grand averaged ERP waveforms averaged to the time of choice presentation for Reward 1 outcomes for Trials 1, 2, and 3. (B) Grand averaged ERP waveforms averaged to feedback onset for Reward 1 outcomes for Trials 1, 2, and 3. (C) The topography of the peak difference between Trial 3 and Trial 1 Reward 1 outcomes for (top) choice presentation and (bottom) reward delivery. (D, E, F) The same plots as A, B, C, except for Reward 2 outcomes.

were used to examine all effects of interest. An alpha level of .05 was assumed for all statistical tests; only significant statistical effects are reported. All error measures reflect 95% within-participant confidence intervals (Masson & Loftus, 2003; Loftus & Masson, 1994).

To verify whether our ERP and behavioral results reflect the output of a RL system, we implemented a computational model based on previous work (Chase, Swainson, Durham, Benham, & Cools, 2010; Holroyd & Coles, 2002, 2008; Cohen & Ranganath, 2007). Our model learned to play the same gambling game as our human participants by computing a prediction error following reward delivery. The prediction error (δ) was then used to update weights ($w_1$ and $w_2$) representing the values of each of the response options. Initially random numbers between 0 and 0.01 (Holroyd & Coles, 2008), these weights determined the probability that an action was selected based on a softmax decision function (Sutton & Barto, 1998):

$$P(\textit{action } i \textit{ is selected}) = \frac{e^{\frac{w_1}{\tau}}}{e^{\frac{w_2}{\tau}} + e^{\frac{w_2}{\tau}}}$$

Here, τ represents temperature, a model parameter that determines the degree to which lower-weighted options are selected. For example, a high temperature makes all options equally likely, whereas a low temperature biases the resulting probabilities toward the higher-weighted option (Sutton & Barto, 1998). Following feedback, a prediction error was calculated by comparing the actual reward ($R$) to the predicted reward (the weight of the action that was taken): $\delta = R - w_i$. As in the actual task, we used positive reward values of 1 and 2; however, to match the performance of our participants, it was necessary to use negative reward values for zero reward outcomes. We chose to use the negative of the win value for a particular gamble: −1 or −2, if the alternative reward (1 or 2) was known, and −1.5 if it was not (i.e., for losses in the first trial). Following the prediction error computation, the weight (value) of the selected action was updated according to the following learning rule: $w_i = w_i + \eta\delta$. Here, η is another model parameter called the learning rate, which determines the degree to which the value of a particular action is updated by the prediction error.

## RESULTS

### Behavioral Data

An analysis of participants' accuracy revealed an effect for Trial, $F(2, 51) = 205.55$, $p < .001$, that demonstrated accuracy increased from Trial 1 (50.3%) to Trial 2 (92.3%), $t(17) = 17.48$, $p < .001$, but did not differ between Trials 2 and 3 (93.1%), $t(17) = 0.94$, $p > .050$. Examination of our RT data showed that RT did not differ on Trial 1 between reward and no reward trials, $F(2, 51) = 0.03$, $p = .969$ (No Reward: 620 msec, Reward 1: 644 msec, Reward 2: 630 msec [*SE* = 9 msec]). However, RT did decrease between the first and second trials in a sequence (637 msec vs. 491 msec, SE = 9 msec, $t(35) = 5.55$, $p < .001$), but not between the second and third trials in a sequence (491 msec vs. 493 msec, SE = 9 msec), $t(35) = 0.37$, $p = .072$. To ensure the task was not "learnable" between blocks, we also analyzed our accuracy and RT data to see if there was a "block" effect over the time course of the experiment. The results of these ANOVAs revealed that there were no differences in Accuracy or RT between experimental blocks ($p$s > .500).

### Electroencephalographic Data

#### *Feedback Presentation: The Feedback Error-related Negativity*

Our analysis of the ERP waveforms averaged to feedback presentation in the fERN time range (200–300 msec) for the first trial of each sequence revealed that fERN amplitude differed between No Reward, Reward 1, and Reward 2 trials, $F(2, 51) = 5.04$, $p = .010$. Specifically, we found that fERN amplitude scaled with increasing reward magnitude—the fERN was more positive for Reward 1 (6.53 ± 1.09 μV) than for No Reward trials (4.23 ± 1.09 μV), $t(17) = 4.22$, $p < .001$, and subsequently was more positive for Reward 2 (9.15 ± 1.09 μV) than for Reward 1 trials, $t(17) = 3.61$, $p = .002$.

Furthermore, we sought to see the change in the fERN amplitude with learning, in other words, how the amplitude of the component changed between Trials 1, 2, and 3 and how that interacted with reward magnitude. Here, we observed a main effect for Trial independent of reward magnitude, the fERN for Reward 1 and Reward 2 trials decreased across trials, $F(2, 102) = 22.67$, $p < .001$. Specifically, the fERN was larger on Trial 1 (7.84 ± 0.85 μV) than on Trial 2 (3.13 ± 0.85 μV), $t(35) = 7.04$, $p < .001$, but the amplitude of the fERN on Trial 2 did not differ from the amplitude of the fERN on Trial 3 (2.71 ± 0.85 μV), $t(35) = 1.36$, $p = .183$. It is also worth noting that we observed a main effect for Reward Magnitude, with the fERN for Reward 2 trials (5.41 ± 0.85 μV) being larger than the fERN for Reward 1 trials (3.71 ± 0.85 μV), $F(1, 102) = 6.11$, $p = .020$.

Given the somewhat unusual nature of the feedback averaged waveforms, we also conducted a wavelet analysis on these data to confirm our ERP results. The results of this analysis demonstrated that frontal midline theta at electrode FCz between 200 and 300 msec—the location, time, and frequency band associated with reward evaluation within medial frontal cortex (e.g., Hajihosseini & Holroyd, 2013; Christie & Tata, 2009)—mirrored our ERP results, $F(2, 34) = 12.22$, $p < .001$. Specifically, theta power at electrode FCz between 200 and 300 msec post-feedback onset increased between the No Reward and Reward 1 conditions and again between the Reward 1 and Reward 2 conditions ($p$s < .05). We note here that we did not observe any change in the latency of the theta activity with learning or differences in theta activity in different time windows.

#### *Feedback Presentation: The P300*

Here, we observed a pattern of results similar to that observed for the fERN with regard to the effect of trial order—the P300 diminished between subsequent trials, $F(2, 102) = 35.08$, $p < .001$. Specifically, we found that the P300 decreased in amplitude from Trial 1 (13.36 ± 1.21 μV) to Trial 2 (4.99 ± 1.21 μV), $t(35) = 7.89$, $p < .001$, and then again between Trials 2 and 3 (3.78 ± 1.21 μV), $t(35) = 3.19$, $p = .003$. Interestingly, and counter to previous work (i.e., Wu & Zhou, 2009; Bellebaum & Daum, 2008; Yeung & Sanfey, 2004), we did not observe the amplitude of the P300 scaling to reward magnitude, $F(1, 102) = 2.66$, $p = .111$.

#### *Prediction Error Propagation: A fERN at Stimulus Onset?*

Recall that we predicted, in line with RL theory, that we would see an ERP response similar to the fERN in response to the onset of the gambling cue with learning. In other words, after participants learned the correct gambling response following feedback on Trial 1, we predicted we would observe a prediction error (i.e., a fERN) to the onset of the gambling cue itself. Interestingly, and in line with our hypothesis, we observed an increased medial-frontal positivity on Trials 2 and 3 of each gamble relative to Trial 1 that had a scalp topography and timing consistent with previous accounts of the fERN. Specifically, we observed an effect for trial order, $F(2, 102) = 8.83$, $p < .001$, that indicated that the medial-frontal positivity we observed scaled to trial order independent of feedback valence and reward magnitude. Specifically, we found that the magnitude of this positivity was greater on Trial 2 (0.10 ± 0.66 μV) than on Trial 1 (−1.70 ± 0.66 μV: $t(35) = 3.69$, $p < .001$), but did not differ between Trials 2 and 3 (0.13 ± 0.66 μV: $t(35) = 0.27$, $p = .792$).

#### *RL Model*

We modeled responses from 18 participants, with each simulated participant completing 108 blocks of the same gambling task as our human participants. Our computational

model was tuned to match the human behavioral performance by adjusting both the learning rate and the temperature and observing the resulting output. Our final model had a learning rate of $\eta = 0.95$ and a temperature that varied randomly and uniformly from $\tau = 0.1$ to $\tau = 0.3$. A unique $\tau$ was chosen for each participant to create variability in the model output. Gaussian noise was added to each prediction error calculation to simulate variability because of neural noise. Model accuracy closely matched human performance (Trial 1: 48.9% ± 0.8%; Trial 2: 90.6% ± 0.1%; Trial 3: 94.3% ± 0.9%). For Reward 1, the model produced prediction errors that shifted from feedback (Trial 1: 0.97 ± 0.07; Trial 2: 0.07 ± 0.03; Trial 3: 0.01 ± 0.03) to cue (Trial 1: 0.01 ± 0.03; Trial 2: 0.99 ± 0.03; Trial 3: 0.98 ± 0.02). Similar results were observed for Reward 1 for both feedback (Trial 1: 1.97 ± 0.02; Trial 2: 0.16 ± 0.03; Trial 3: −0.01 ± 0.03)) and cue (Trial 1: −0.01 ± 0.02; Trial 2: 1.84 ± 0.04; Trial 3: 2.03 ± 0.03).

### *Relationship between Human and Model Prediction Errors*

To ensure that the changes in the timing of the reward prediction error signal observed in ERP waveforms followed the predictions of our computational model, we conducted a regression analysis. First, we standardized the cue and feedback ERP peaks (i.e., the medial-frontal cue and reward positivities) for Reward 1 and Reward 2 trials to compensate for between- and within-subject variability. Next, we computed the difference between the standardized medial-frontal response at the time of choice presentation and the standardized medial-frontal response at the time of feedback delivery for both reward levels for each trial for each participant. The logic here was simple, on Trial 1 this difference should be negative as the standardized reward positivity following feedback onset should be greater than the reward positivity following choice presentation. Conversely, for Trials 2 and 3, this difference should be positive as the standardized response should be greater at the time of stimulus cue onset than at feedback delivery. We repeated this process for the prediction errors computed by our model at choice presentation and feedback delivery for each trial for each simulated participant. Finally, we ran a regression between the human and model data and found a strong relationship, $r = 0.473$, $p < .001$. In other words, a statistical confirmation that the pattern of results observed in the ERP data mirrored the output of our computational model (see Figure 4).

## DISCUSSION

In the present experiment we provide novel ERP evidence demonstrating that actions rapidly acquire value with learning. Specifically, our results demonstrate that participants (a) processed novel rewards and punishments, (b) the neural response at the time of reward diminished with learning, and importantly (c) there was a concomitant increase in a novel neural response at choice presentation, before action selection, that we propose reflects an RL prediction error. Together, these data provide strong evidence that the computations made by a neural learning system within medial-frontal cortex (Holroyd & Coles, 2002) follow RL principles.

In line with a large body of existing research, an analysis of the neural response to reward delivery on Trial 1 of the gambling task revealed an ERP component with a timing and scalp topography consistent with previous accounts of the fERN (Krigolson et al., 2009; Holroyd & Krigolson, 2007; Hajcak, Moser, Yeung, & Simons, 2005; Yeung & Sanfey, 2004; Gehring & Willoughby, 2002; Miltner, Braun, & Coles, 1997). Furthermore, the feedback-averaged waveforms for Trial 1 bore a similar resemblance to the results of Ferdinand, Mecklinger, Kray, and Gehring (2012)—with particular similarities seen between the P200–N200–P300 complex in both studies. Recall that the fERN has been proposed to reflect an RL prediction error (Holroyd & Coles, 2002), and supporting that contention here we saw the component elicited by novel feedback. Previous accounts of the fERN have suggested that it reflects a binary judgment of task outcomes and thus does not scale to reward magnitude (Hajcak, Moser, Holroyd, & Simons, 2006). However, here we show that the fERN scaled to reward magnitude—the component scaled linearly between Reward 0, Reward 1, and Reward 2 trials—a result in line with studies in monkey demonstrating that the amplitude of dopaminergic responses to reward delivery scale to reward magnitude (Tobler, Fiorillo, & Schultz, 2005). Furthermore, studies using the ERP technique have shown that the fERN scales to reward expectancy, and this is what we propose drives the effect observed here. Recall that participants were aware that they could either win 1 or 2 cents. As such, a reduced fERN was elicited when 1 cent was won as participants were expecting a 2-cent reward—a possibility in line with previous work that has shown a similar effect (Wu & Zhou, 2009).

Furthermore, in the present experiment, we did not see the P300 scale to reward magnitude as in some previous studies (e.g., Yeung & Sanfey, 2004). However, more recent research suggests that the P300 is not sensitive to reward magnitude but instead is sensitive to the riskiness of a gamble (Oberg, Christie, & Tata, 2011; Christie & Tata, 2009). Our results are in line with these more recent reports as there was no manipulation of riskiness in this study. Alternatively, if the P300 does indeed scale to reward magnitude (Christie et al., Oberg et al., notwithstanding), the lack of the effect observed here may be because reward magnitude was encoded by an earlier process (i.e., the fERN). Thus, the process underlying the P300 did not encode this information in our experiment. Indeed, recent work has shown that the timing of the fERN is sensitive to cognitive load—the component occurs later on high cognitive load trials (Krigolson, Heinekey, Kent, & Handy, 2012)—thus, it is not unreasonable to assume that the timing of reward processing may vary from experiment

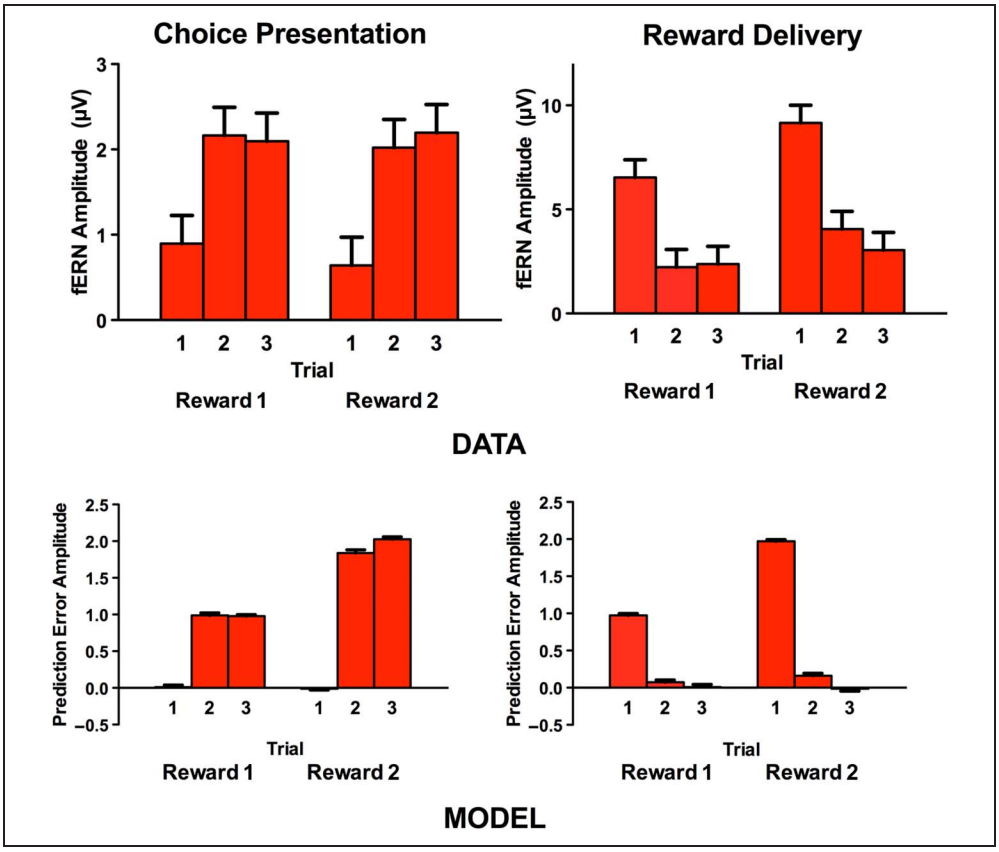


**Figure 3.** Mean fERN amplitudes (top) for choice presentation and reward delivery for Reward 1 and Reward 2 outcomes. Mean prediction errors (bottom) calculated by the computational model of the task. Note the similarity between the outputs with the exception that the fERN at choice presentation did not seem to scale to value magnitude.

to experiment because of task complexity or other factors. It is worth noting that we did not have a sufficient number of Reward 0 outcomes on Trials 2 and 3 to analyze the fERN at these instances.

Our results also demonstrated that the amplitude of the positive component of the fERN diminished in amplitude with learning—a result in line with the predictions of RL theory and mirroring the output of our computational model that predicted a similar pattern of results for the amplitude of predictions errors calculated at the time of reward delivery (see Figures 3 and 4). Recall that a prediction error occurs when there is a discrepancy between the actual and predicted values of a reward. In this study, on Trial 1 participants were unaware which action would result in a reward—pressing the left or the right gamepad button. The value of each action would have initially been zero when a reward was first won; thus, a prediction error would be computed by the reward system. In line with this, as mentioned above, we observed a fERN on Trial 1 of the gambling task when feedback was novel. However, given the ease of the task, it stands to reason that the learning rate would be high—a contention supported by the rapid improvement in response accuracy between Trials 1 and 2 and by our computational model, which needed a learning rate of 0.95 to mirror participants' accuracy results. With a high learning rate, the value of the correct action would increase substantially between Trials 1 and 2. As a result of this rapid increment in value, one would expect a small fERN on Trials 2 and 3—and this is what we observed in our ERP data. It is worth noting that our results are similar to those of Krigolson et al. (2009), who also found that the amplitude of the fERN diminished with learning for participants who acquired a measure of perceptual expertise though a trial and error shaping process. Finally, what about No Reward outcomes on Trial 1? The same logic applies as the task was completely deterministic—if the outcome of response selection on Trial 1 did not result in a reward, then participants knew the other action would lead to a reward, and thus, it would make

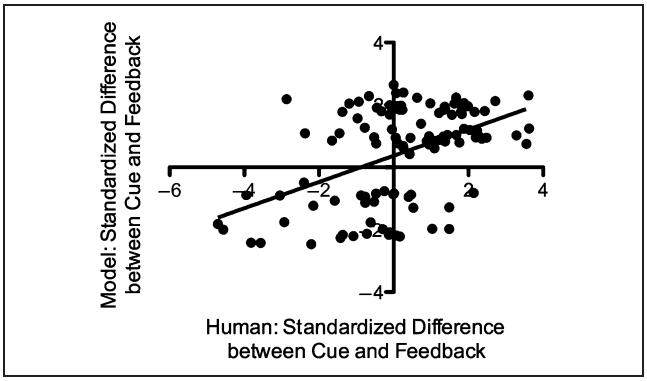


**Figure 4.** Standardized differences between cue and reward prediction errors for human participants plotted against matched simulation participants. The relationship suggests that the pattern of human results mirrored the predictions of our computational model.

sense to increase the value of the action that was not chosen.

One may wonder why the positive component of the fERN is of interest here given that we are discussing a fERN. Indeed, the fERN is typically associated with the negative waveform. However, recent research suggests that perhaps what we are observing in feedback evaluation tasks is a correct-related positivity (Holroyd, Pakzad-Vaezi, & Krigolson, 2008). Our results support this hypothesis, and further, bring the literature in line with studies in monkey that demonstrate phasic dopaminergic activity carries positive but not negative prediction errors (Waelti, Dickinson, & Schultz, 2001). In other words, what our results suggest—at least in line with an interpretation of the Holroyd and Coles (2002) theory—is that a phasic increase in dopamine at the time of reward feedback drives a reward positivity in the fERN time range early in learning and at the time of choice presentation after learning has occurred. It is worth noting that other studies have also reported that unexpected positive feedback resulted in a positive deflection of the N200 component, albeit with a different interpretation of the result (e.g., Ferdinand et al., 2012).

Interestingly, we also observed an ERP component at the time of choice presentation with a timing and scalp topography consistent with the fERN. Furthermore, the amplitude of this component increased with learning concomitantly with the decrease in the fERN at the time of reward delivery—a result that mirrored the predictions of formal RL theory and the output of our computational model (see Figures 3 and 4). Typically, the fERN is defined as the difference between win and loss trials in a gambling task. However, at the time of choice presentation in the present experiment, there was no difference between win and loss trials, that is, a response had yet to be selected and thus the outcome was unknown even if the values of the two response options were known. Furthermore, given our task structure, participants were not able to predetermine their response until they viewed which of the two gambles they were asked to play. As such, the change we observed is novel in that it is not a fERN and we can only speculate as to what cognitive process is reflected by the increase in component amplitude that we observed. With that said, we propose the increase in component amplitude reflects the processing of a prediction error brought about by the learning-driven increase in value of the action selection state for two principle reasons. One, the timing and topography of the component we observed are consistent with the fERN. Two, an examination of Figures 2, 3, and 4 reveals that the changes in the ERP waveforms associated with the onset of choice presentation and reward delivery almost exactly mirror the output of our computational model. Our logic is speculative, and there are potentially other explanations for our findings—more research is needed to clarify the processes underlying the cue-evoked ERP component observed here. It is worth noting that the component evoked by choice presentation did not scale to reward value—a result that differs from the predictions of formal RL theory. Again, more research is needed to clarify why this is—but it is perhaps related to the contention by Hajcak et al. (2006), proposing that the fERN reflects a binary evaluation of the potentials reward outcomes—in this case a binary evaluation that a reward may be attained by selecting the correct response.

In conclusion, our results provide novel evidence that the fERN component of the human brain ERP is the scalp signature of the impact of RL prediction error signals sent from the BG to the ACC as posited by Holroyd and Coles (2002). Furthermore, our data provide the first ERP evidence that choice states acquire value with learning and that these increases in value follow the computations predicted by RL theory. Importantly, as actions acquire value with learning, we gain the ability to make effective decisions, and thus, we can follow our inherent desire to maximize utility (Mill, 1879).


## Acknowledgments

The authors would like to thank the National Sciences and Engineering Research Council of Canada for funding for this research. The first author would also like to thank Jeffrey Chang and Hayley May Neimy for their help with data collection.



Reprint requests should be sent to Olav E. Krigolson, Department of Psychology and Neuroscience, Dalhousie University, Life Sciences Centre, Halifax, Nova Scotia, Canada, B3H 4J1, or via e-mail: krigolson@dal.ca.


## REFERENCES


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

Bellebaum, C., & Daum, I. (2008). Learning-related changes in reward expectancy are reflected in the feedback-related negativity. *European Journal of Neuroscience, 27,* 1823–1835.

Bernoulli, J. (1713). *Ars conjectandi*. Basel: Impensis Thurnisiorum, Fratrum.

Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2010). Feedback-related negativity codes prediction error but not behavioral adjustment during probabilistic reversal learning. *Journal of Cognitive Neuroscience, 23,* 936–946.

Christie, G. J., & Tata, M. S. (2009). Right frontal cortex generates reward-related theta-band oscillatory activity. *Neuroimage, 48,* 415–422.

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

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134,* 9–21.

Ferdinand, N. K., Mecklinger, A., Kray, J., & Gehring, W. J. (2012). The processing of unexpected positive response outcomes in the mediofrontal cortex. *Journal of Neuroscience, 32,* 12087–12092.

Gehring, W. J., & Willoughby, A. R. (2002). The medial frontal cortex and the rapid processing of monetary gains and losses. *Science, 295,* 2279–2282.
Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology, 55,* 468–484.
Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biological Psychology, 71,* 148–154.
Hajcak, G., Moser, J. S., Yeung, N., & Simons, R. F. (2005). On the ERN and the significance of errors. *Psychophysiology, 42,* 151–160.
Hajihosseini, A., & Holroyd, C. B. (2013). Frontal midline theta and N200 amplitude reflect complementary information about expectancy and outcome evaluation. *Psychophysiology, 50,* 550–562.
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,* 679–709.
Holroyd, C. B., & Coles, M. G. H. (2008). Dorsal anterior cingulate cortex integrates reinforcement history to guide voluntary behavior. *Cortex, 44,* 548–559.
Holroyd, C. B., & Krigolson, O. E. (2007). Reward prediction error signals associated with a modified time estimation task. *Psychophysiology, 44,* 913–917.
Holroyd, C. B., Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2003). Errors in reward prediction are reflected in the event-related brain potential. *NeuroReport, 14,* 2481.
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,* 688–697.
Holroyd, C. B., Yeung, N., Coles, M. G. H., & Cohen, J. D. (2005). A mechanism for error detection in speeded response time tasks. *Journal of Experimental Psychology: General, 134,* 163–191.
Krigolson, O. E., Heinekey, H., Kent, C. M., & Handy, T. C. (2012). Cognitive load impacts error evaluation within medial-frontal cortex. *Brain Research, 1430,* 62–67.
Krigolson, O., Holroyd, C., Van Gyn, G., & Heath, M. (2008). Electroencephalographic correlates of target and outcome errors. *Experimental Brain Research, 190,* 401–411.
Krigolson, O. E., Pierce, L. J., Holroyd, C. B., & Tanaka, J. W. (2009). Learning to become an expert: Reinforcement learning and the acquisition of perceptual expertise. *Journal of Cognitive Neuroscience, 21,* 1833–1840.
Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin and Review, 1,* 476–490.
Masson, J. E., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57,* 203–220.
Mill, J. S. (1879). *Socialism*. Chicago: Belfords, Clarke & Company.
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.
Oberg, S. A. K., Christie, G. J., & Tata, M. S. (2011). Problem gamblers exhibit reward hypersensitivity in medial-frontal cortex during gambling. *Neuropsychologia, 49,* 3768–3775.
O'Doherty, J. P. (2004). Reward representations and reward-related learning in the human brain: Insights from neuroimaging. *Current Opinion in Neurobiology, 14,* 769–776.
Rescorla, R. A., & Wagner, A. R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. New York: Appleton Century Crofts.
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.
Waelti, P., Dickinson, A., & Schultz, W. (2001). Dopamine responses comply with basic assumptions of formal learning theory. *Nature, 412,* 43–48.
Wu, Y., & Zhou, X. (2009). The P300 and reward valence, magnitude, and expectancy in outcome evaluation. *Brain Research, 1286,* 114–122.
Yeung, N., & Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *Journal of Neuroscience, 24,* 6258–6264.