A Cognitive Science Honors Thesis

The Influence of Goal Setting on Value Learning: Examining how Goal-directed Behavior leverages the Reinforcement Learning System

Nora Harhen

May 2018

The Computational Cognitive Neuroscience Lab

Advisor: Anne Collins

Second Reader: Linda Wilbrecht

Introduction

Reinforcement Learning theories have contributed greatly to our understanding of how our learning and decision-making is shaped by external reward through providing a framework to which neural correlates could mapped on. However, in everyday life, we generally do not receive external rewards like food or money that guide our behavior. Instead, we set goals for our self, and use them as reference points in learning the correct actions to take. A large body of work from behavioral economics and social psychology has found that goal setting is intrinsically motivating, and hence goal achievement is, in itself, valuable. This suggests that it may function as a pseudo-reward, with the same reinforcing properties as external reward. Merging ideas from RL theory and behavioral economics, it is plausible that the same computations that are performed on external reward may be performed on goal achievement if it generates a pseudo-reward.

Reinforcement Learning (RL) Algorithms and their Neural Substrates

Reinforcement Learning is learning the correct actions to take in a certain environment in the pursuit of maximizing reward over time. It is distinguished by its trial-and-error nature: the learner has no examples of desired behavior to learn from, so she predicts the outcomes of the possible actions she can take, selects an action, receives feedback, and uses the difference between the actual outcome of her action and the predicted outcome to drive her learning (Sutton & Barto, 1998). The difference between actual and expected outcome is referred to as the reward prediction error (RPE). There is a wealth of evidence linking computational constructs in RL to neural substrates (Niv, 2009). Specifically, it has been found that during reward-based learning the firing of dopaminergic cells in ventral tegmental area (VTA) and substantia nigra signal RPE (Montague, Dayan & Sejnowski, 1996; Schultz et al., 1997). These cells

synapse on to other cells in striatum, nucleus accumbens, and frontal cortex, structures implicated in motivation and goal-directed behavior.

Hierarchical Reinforcement Learning (HRL)

Classic RL can only describe a small subset of learning problems. Most of our behaviors are complex and hierarchically structured. Lashley (1951) observed that a sequence of primitive actions requires a higher-level task context representation. An extension of RL is Hierarchical Reinforcement Learning (HRL) (Parr & Russell, 1998; Sutton et al., 1999; Barto & Mahadeven, 2003). There are multiple implementations of HRL. Within the MAXQ implementation (Dietterich, 2000), a goal to be completed is divided into internally-defined subgoals. Reaching a subgoal elicits a pseudo-reward, referred to as such to differentiate it from the reward of reaching the overall goal. Learning of subroutines to complete the subtask is driven by a pseudo-reward prediction error (PPE) that functions similarly to a RPE. The subgoal may not lead to an external reward at all. Its value stems from the agent internally defining it as valuable. The HAM implementation (Parr & Russell 1998) does not use pseudo-rewards, instead relying on exogenous rewards and prior knowledge constraining the policies considered. These different implementations make different predictions for dopaminergic function (Botvinik, Niv, & Barto, 2009). A prediction originating from the MAXQ framework, is that a neural substrate of the pseudo-reward prediction error should exist. Results from fMRI studies have been consistent with this. Ribas-Fernandes et al. (2011) found that structures responsive to RPEs also responded to PPEs. In another study, dopaminergic neurons were found to have the ability to signal both a global reward prediction error and local pseudo-reward prediction error that temporally coincided (Diuk et al, 2013). Taken together, these are indications that neural correlates of RPEs are more flexible in their use than previously thought.

Goal Achievement and Intrinsic Motivation

There is a rich body of research in behavioral economics and social psychology on the psychological value of goal achievement. A goal may be or lead to an extrinsic reward. However, for our purposes, when referring to goals, we will be referencing "mere goals." Mere goals differ from extrinsic reward in that they describe a certain level of performance/achievement that does not affect external amounts of reward (Heath, Larrick, & Wu, 1999). Goal setting has been shown to motivate behaviors that lead to those goals (McDougall, 1908; Mitchell, 1982). Tied to this, goal setting has been shown to increase task performance (Locke & Latham, 1991; Mento, Steel & Karen, 1987; Tubbs, 1986; Mossholder 1980), goals that are internally-defined as opposed to assigned are particularly effective (Schmidt & Hunter, 1983).

We can draw parallels between Bandura's theory of goal pursuit (1991) as discrepancy reduction and the role of reward in guiding behavior as prescribed by RL theories. Within Bandura's framework, goals serve as desired end states and also as a reference point for evaluating performance, making the act of goal setting a discrepancy inducing process. Behavior is shaped by feedforward and feedback controls. Through feedforward control, an action is chosen and motivated by its predicted outcome. Once the action is performed, feedback received induces changes in order to reduce the discrepancy between prediction and actual outcome in service of getting closer to the goal (Bandura, 1988). This discrepancy reduction closely mirrors RL algorithms in which the difference between predicted reward and actual reward drives learning.

Reward is defined by the value it has, but research on the psychological value of goal achievement and its relation to the value of reward is a relatively new area of research. Ballard and colleagues (2017) found that the motivating effects of goals may lead to departures from objective reward maximizing behavior, indicating that there is subjective value attached to goal achievement. People exerted unnecessary effort in

order to obtain multiple goals. Interestingly, there were individual differences in the value attached to goal achievement

Intrinsically-motivated Learning

In classic RL, behavior is shaped by external reinforcers and punishers. However, we know that humans have both external and internal sources of motivation. Inspired by how humans learn, there has been a wave of RL research on intrinsically motivated learning, in which an agent engages in learning for their own sake rather than as a step towards an explicit end state (Barto, Singh, & Chentanez, 2004). This is implemented by making exploration rewarding in and of itself. Curiosity is the reinforcer as opposed to an external reward (Singh, 2005). This is biologically plausible as the dopaminergic neurons responding to RPEs also respond to salient novel stimuli. It is reasonable that a motivational system as robust as the RL one would be flexible enough to perform computations on internal rewards as well as external ones.

Merging the RL and Goal Achievement Literatures

We seek to merge frameworks from both the reinforcement learning and goal setting literatures by answering the question: can goal-directed behavior leverage reinforcement learning mechanisms? We propose that goal achievement is treated as a pseudo-reward which allows RL computations to be performed upon it in order to drive learning the behaviors that brings the learner closer to their goals. Accordingly, the neural signature of PPEs should be similar to that of RPEs. However, there may be interesting differences in how people learn from rewards and pseudo-rewards behaviorally, and that could be correlated with neural differences. We aim to quantify the relative value of pseudo-rewards in relation to external rewards, explore how they influence computation, and how individual differences in value are reflected neurally.

Methods and Materials

Experimental Protocol

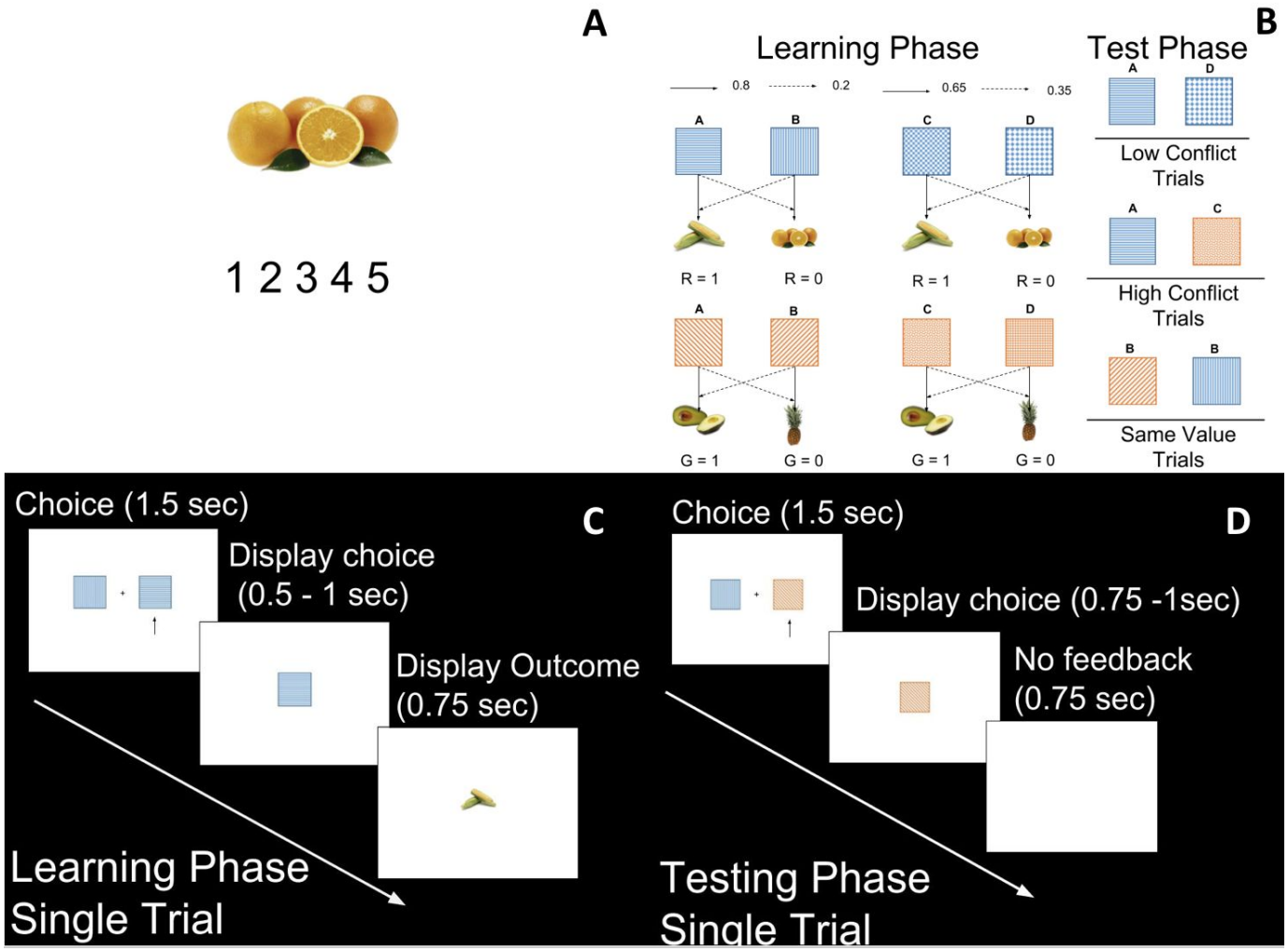


Figure 1: Experimental protocol **a)** Participants were asked to rate 20 fruits and vegetables on a scale of 1 to 5. This allowed us to identify 4 items of equal value that will become the four possible items that can be found in the boxes. 2 can be found in reward boxes, and the other 2 can be found in goal boxes **b)** Structure of the learning phase and the test phase. Each box had a different expected value of reward/pseudo-reward. For both conditions, there were two pairs of boxes. In one pair, there was a box that would lead to a positive outcome (reward/goal) 80% of the time (solid line) paired with a box leading to a positive outcome 20% of the time (dashed line). In the other pair, one box would lead to a positive outcome 65% of the time (solid line) and the other 35% (dashed line). Participants always saw the same boxes paired together and the items identified as reward, goal, non-reward, and non-goal retained their roles throughout the task. During the test phase, all eight boxes were shown paired together. Easy trials included AD (reward contingencies: 0.8/0.35) and CB (0.65/0.2) pairs. Hard trials included AC (0.8/0.65) and DB(0.35/0.2) pairs. Same value trials are ones in which a reward box and a goal box with the same expected value were paired together. **C)** During the learning phase, participants were shown two boxes and were asked to make a choice. The chosen box was displayed in the center of the screen, following which they received feedback in the form of a vegetable. **D)** During the test phase, participants were asked to pick one of two boxes. Their choice was displayed, however, they received no feedback.

We developed a novel variant of the probabilistic selection task (Frank, 2004) to directly compare how participants learned from two types of outcomes: external reward and goal achievement, representing a pseudo-reward. Participants were asked to rate 20 fruits and vegetables on a scale of 1-5 based on how much they liked them. This allowed us to identify 4 fruits and vegetables that the participant valued equally. These four fruits and vegetables became the reward, non-reward, goal, or non-goal for the participant. The participant was told that they would be choosing between boxes to open and finding fruits and vegetables. There were two conditions: reward and goal. In the reward trials, there were two fruits/vegetables that could be found in reward boxes. The experimenter instructed the participant which vegetable would lead to points when found and which would lead to no points. In the EEG experiment, the amount of points accumulated translated to the participant's bonus payment, so the fruit/vegetable became tied to an real-world external reward. In the behavioral experiments, there was no bonus payment offered. In the goal trials, the participant chose their own goal amongst two potential fruits/vegetables. The boxes leading to reward/non-reward and those leading to the goal/non-goal were different colors. This notified participants as to which condition they were in. The mapping between which vegetables represented the reward,non-reward, non-goal was held constant throughout the task. For both types of trials, participants sampled and learned about 4 pairs of boxes of various expected values and with probabilistic feedback Figure 1B). We defined easy trials as pairings with reward contingencies of 0.8/0.35 and 0.65/0.2, while hard trials were defined as pairings with reward contingencies of 0.8/0.65 and 0.35/0.2. The "correct" box in a pair was the higher-valued one. Comparatively, easy trial pairs had a larger value difference between the two boxes, making the decision of which is the box more likely to lead to reward easier. Goal trials sequences and outcome contingencies were yoked to those of the reward trials to ensure identical reward histories.

Following the learning phase, there was a test phase in extinction. Participants continued to select boxes, but they did not receive feedback following their choice. This allowed the values acquired from the learning phase to be fixed. Participants encountered all possible pairings of the 8 boxes. This allowed us to ensure that participants learned the expected values of the boxes rather than an action policy (e.g. "when presented horizontal and vertical striped boxes, pick vertical striped") and to compare preferences between goal achievement boxes and reward boxes. In our analysis of test phase performance, we excluded trials of pairs previously seen in the learning phase and only analyze novel pairings. Choosing correctly on these trials requires having integrated reward histories to compute an expected value for each box. An easy choice is between a previously "correct" and one previously "incorrect" box, but a hard choice is between two previously "correct" or "incorrect" boxes. One pairing of particular interest was a goal achievement box and a reward box of the same expected value (assuming goal achievement produces a pseudo-reward equivalently valued as reward). A systematic preference for one type suggests which type of outcome is more highly-valued.

Temporal Structure within each trial and between trials

After presentation of the box pair, participants had 1.5 s to choose a box using two keys (Q for left box and P for right box) on a standard computer keyboard. Their choice was displayed for an amount of time between 500 to 1000 ms. Feedback was then displayed for 750 ms. The intertrial interval was jittered between 500 ms and 1000 ms. Jittering discourages participants from predicting the temporal onset of the following trial. This is critical as this preparation can be reflected within the EEG (Luck, 2005, Cohen, 2014).

Participants

21 behavior only subjects and 40 EEG subjects participated in this experiment. 2 subjects were excluded from analysis for performing below chance during the last 60 trials of the learning phase suggesting that they did not engage in the task.

Modeling Methods

Reinforcement Learning Models Considered

All models considered are either some variant of a reinforcement learning (RL) model or contain an RL component. RL models implement a simple delta learning rule. The expected value, Q, of a stimulus is updated as a function of the difference between the actual reward and the predicted reward, δ. The learning rate, α, scales δ's magnitude of impact on the updated Q value. As α increases, the more weight given to recent outcomes in updating the expected value.

$$Q_{(t+1)} = Q_{(t)} + \alpha\delta_t \quad 0 \leq \alpha \leq 1 \qquad (1)$$

Choices are generated probabilistically from a softmax probability distribution such that actions with higher Q values have a higher likelihood of being chosen (Sutton & Barto, 1998).

$$p(a|s) = e^{\beta Q(a)}/(e^{\beta Q(a1)}+e^{\beta Q(a2)}) \qquad (2)$$

β is an inverse temperature parameter that controls the degree of exploration versus exploitation. An increase in β decreases the degree to which the the higher valued option will be chosen.

We add additional parameters to increase fits to behavior.

Two Alphas: alpha Gain and alpha Loss

We include two alphas, one for positive reward prediction errors and another for negative reward prediction errors to capture individual differences in impacts of positive versus negative outcomes (Frank, Moustafa, Haughey, Curran, & Hutchison, 2007) .

$$Q_{(t+1)} = Q_{(t)} + \alpha_+\delta_t \text{ if } \delta_t > 0 \qquad (3)$$

$$Q_{(t+1)} = Q_{(t)} + \alpha\_\delta_t \text{ if } \delta_t < 0 \qquad (4)$$

Forgetting

To account for potential forgetting, we add a parameter, ϕ,that controls the rate of decay of the unchosen box's integrated value to the initial expected value, 0.5.

$$Q_{(t+1)} = Q_{(t)} + \phi*(Q_0 - Q_{(t)}) \qquad 0 \leq \phi \leq 1 \qquad (5)$$

$$Q_o = 1/n_a \qquad (6)$$

There are two choices on each trial ($n_a = 2$).

Sticky Choice

Irrespective of value, participants have a tendency to pick the box on the same side as their last choice or have a tendency to switch sides as evidenced by results from our logistic regression model on test phase data (see results section). To capture this in our model, we add a "stickiness" parameter, *s*, that increases or decreases the probability the option chosen will be on the same side (left or right) as the choice on the last trial .

If previous choice left,

$$P(\text{choose left}) = 1/(1 + e^{(-\beta(Q(\text{left})*s - Q(\text{right}))}) \qquad (7)$$

If right,

$$P(\text{choose left}) = 1/(1 + e^{(-\beta(Q(\text{left}) - Q(\text{right})*s)}) \qquad (8)$$

Preference for Goal

To quantify the value of pseudo-rewards relative to reward, we add a preference for reward parameter, *r,* that represents the value of receiving a pseudo-reward relative to reward.

When reward is received,

$$\delta = 1 - Q_{(t)} \qquad (9)$$

When pseudo-reward is received,

$$\delta = (r + 1) - Q_{(t)} \qquad -1 \leq r \leq 1 \qquad (10)$$

If $r = 0$, then rewards and pseudo-rewards are treated equivalently. $r > 0$ indicates pseudo-reward is more valued than reward, while $r < 0$ indicates it is less valued than reward.

RLWM Model

There is evidence that the reinforcement learning system and working memory work cooperatively to enable quick and efficient learning (Collins et al., 2012). Thus, we combine two modules, a RL component and a working memory (WM) one, in which each module learns the values of stimuli separately and makes a weighted contribution to the final action choice.

The RL component is the RL model that provided the best fit, RLprefR**.**

The task does not directly manipulate strain on the WM system, so we approximate it through casting it as a RL model with a high learning rate to capture the flexibility and quick updating characteristic of WM. The WM learning rate is constrained to be less than the gain and loss learning rates of the RL component. Defining θ this way, ensures that $\alpha_{\mathrm{RL}} \leq \alpha_{\mathrm{WM}}$.

$$\theta = \alpha_{\mathrm{RL}} / \alpha_{\mathrm{WM}} \quad 0 \leq \theta \leq 1 \qquad (11)$$

The probability of taking an action is determined with weighted contributions from both RL and WM modules according to:

$$\mathrm{p(a)} = \mathrm{p(a)}_{\mathrm{WM}} * w_{WM} + \mathrm{p(a)}_{\mathrm{RL}} * (1 - w_{WM}) \quad 0 \leq w_{WM} \leq 1 \qquad (12)$$

W_{WM} is a free parameter representing the weight of WM's contribution to the final action choice. It is fixed as opposed to dynamically updated throughout the task.

RLBayes Model

The RLBayes model combines an RL module with a Bayesian learner module which infers the expected value of the stimuli. Each module separately learns the expected values of a stimuli and makes a weighted contribution to the final action choice.

The RL module implements the TD learning algorithm and includes additional parameters found in our RLprefR model and RLWM model: α_+ and α_-, ϕ for forgetting, and *r* for preference of reward. We add another parameter unique to this model: .

Win-Stay Lose-Shift

To quickly accumulate reward, one could employ a win-stay lose shift strategy in which an action is repeated if it has been rewarded while if unrewarded, another action is tried. However, this strategy does not help in learning the expected value of the actions. Models combining RL and WSLS models have been effective in capturing behavior in decision-making tasks (Worthy and Maddox, 2013), so we incorporated this strategy into our model through having a parameter, *wsls,* that increases the probability that a choice would be made if rewarded on the last trial with the considered pair.

If previous choice was box A in pair AB and it was rewarded,

$$P(\text{choose A}) = 1/(1 + e^{(-\beta(Q(A)*wsls - Q(B))}) \quad (13)$$

If unrewarded,

$$P(\text{choose A}) = 1/(1 + e^{(-\beta(Q(A) - Q(B)*wsls)}) \quad (14)$$

The Bayes module infers the expected value of the stimuli given the data experienced according to Bayes rule:

$$P(Q_A \mid D) = P(Q_A)*P(D \mid Q_A) / P(D) \quad (15)$$

$$P(D \mid Q_A) = \text{number of times } Q_A \text{ rewarded/number times } Q_A \text{ chosen} \quad (16)$$

$$P(Q_{A\,(t+1)}) = P(Q_{A(t)} \mid D) \quad (17)$$

The priors are the values put into the softmax equation to determine the probability of taking an action. We add an additional parameter to this component in the model to increase fits to behavior:

Decay

Our Bayes module incorporates a decay parameter, γ, that uniformly decays all priors to the initial expected value. Without a decay parameter, we find the values asymptote too quickly for the Bayes component of the model when compared to participant behavioral data.

$$priors_{t+1} = priors_t * \gamma + (1-\gamma) * priors_0 \qquad 0 \leq \gamma \leq 1 \qquad (18)$$

The final probability of an action choice combines input from the RL and Bayes modules according to:

$$p(a) = p(a)_{Bayes} * w_{WM} + p(a)_{RL} * (1-w_{Bayes}) \qquad 0 \leq w_{Bayes} \leq 1 \qquad (19)$$

W_{Bayes} is a free parameter representing the weight of the Bayes module's contribution to the final action choice. It is fixed as opposed to dynamically updated throughout the task.

Parameter Fitting

Parameters for each participant were estimated using individual behavioral data and MATLAB's fmincon function to identify the set of parameters that produced the least negative log likelihood. Parameter estimation was performed 20 times, and parameters with the least negative log likelihood were used in simulations. This process was repeated on fits to three different sets of data: training phase data only, test phase data only, and both data sets.

Model Comparison

Akaike's Information Criterion (AIC) was used to compare goodness of fit between models. AIC penalizes models with more free parameters. The model with the lowest AIC amongst candidate models is the one with the best fit. This is used as opposed to Bayesian Information Criterion (BIC) because of more accurate model prediction (see results section).

$$AIC = -2\ln(L) + 2V \qquad (20)$$

where L is the maximum likelihood of the model considered, and V is the number of free parameters in the model.

Model Simulation

For each subject, fit parameters were used to run 100 simulations which were averaged to represent the behavior of that subject and their contribution to the group average. After running simulations and averaging for each subject, a group average was computed for training phase and test phase performance. This was then compared with group averages from participant's real behavioral data to check if the model captures key aspects of behavior.

EEG Methods

EEG was recorded using a 64 channel BioSemi system. We used previously identified data cleaning and preprocessing methods (Cavanagh, 2009) facilitated by the EEGLab toolbox (Delorme & Makeig, 2004).

Preprocessing EEG

EEG was recorded continuously with hardware filters set to 0.1 and 100 Hz and using a sampling rate of 512 Hz. Data was visually inspected to identify epochs with artifacts for removal and noisy channels for interpolation. Eyeblinks were removed using independent component analysis from EEGLab.

Model-based EEG analysis

For ERP and multivariate regression analysis, data was band-pass filtered between 0.5 and 15 Hz and baselined to the average activity between -300 to 0 ms prior to feedback or stimulus presentation. A regression approach was used to extract the effect of multiple variables of interest. For each subject we performed multiple regression at all electrodes and all time points within -300 and 800 ms of feedback or stimulus presentation (141 time points).

For feedback-locked analysis, the main variables of interest were outcome valence (positive or negative), outcome type (reward or goal) , and trial-by-trial estimates of RPEs extracted from the best-fitting model. For stimulus-locked analysis, the model's predictors were the Q value of the chosen option, whether it was a reward or goal trial, and the number of times the pair of stimuli (boxes) had been presented previously. Only training phase data was analyzed. All regressors both feedback-locked and stimulus-locked analysis were z-scored.

Statistical analysis of GLM weights

We tested the β weights of all regressors against 0 across all subjects from each electrode and time point using a t-test with a threshold set to p=0.001.

ERP

For event-related potentials (ERP), data was band-pass filtered between 0.5 and 15 Hz. ERPs were baseline corrected to the average activity between -300 to 0 ms prior to feedback presentation.

Corrected ERPs

To plot corrected ERPs, the average of the predicted voltage was computed from the multi-regression model when setting all regressors but the one of interest to 0. This was subtracted from the true voltage leaving only the fixed effect, the variance explained by the regressor of interest, and residuals. For binary regressors, we plotted one ERP that was the average of trials with the effect and the other ERP is the average of trials without it. For continuous regressors, we did a median split on all the trials for the regressors and plotted one ERP that was the average of those trials above the median split and an average for those below the split.

Time Frequency Analysis

Time-frequency calculations were computed using custom-written Matlab scripts (Cavanagh et al., 2009, Cohen et al, 2008). Time-frequency measures were calculated by multiplying the fast Fourier

transformed (FFT) power spectrum of single trial EEG data with the FFT power spectrum of a set of complex Morlet wavelets defined between 4 and 8 Hz, and taking the inverse FFT. Each epoch was cut in length (−500 to +1000 ms). The baseline for each frequency consisted of the average power from −300 to −200 ms prior to the onset of the cues. Whereas the ERPs reflect phase-locked amplitude changes, these time-frequency measures reflect total power (phase-locked and phase-varying).

Results

Behavioral

1. *Pseudo-rewards reinforce choices similarly to rewards*

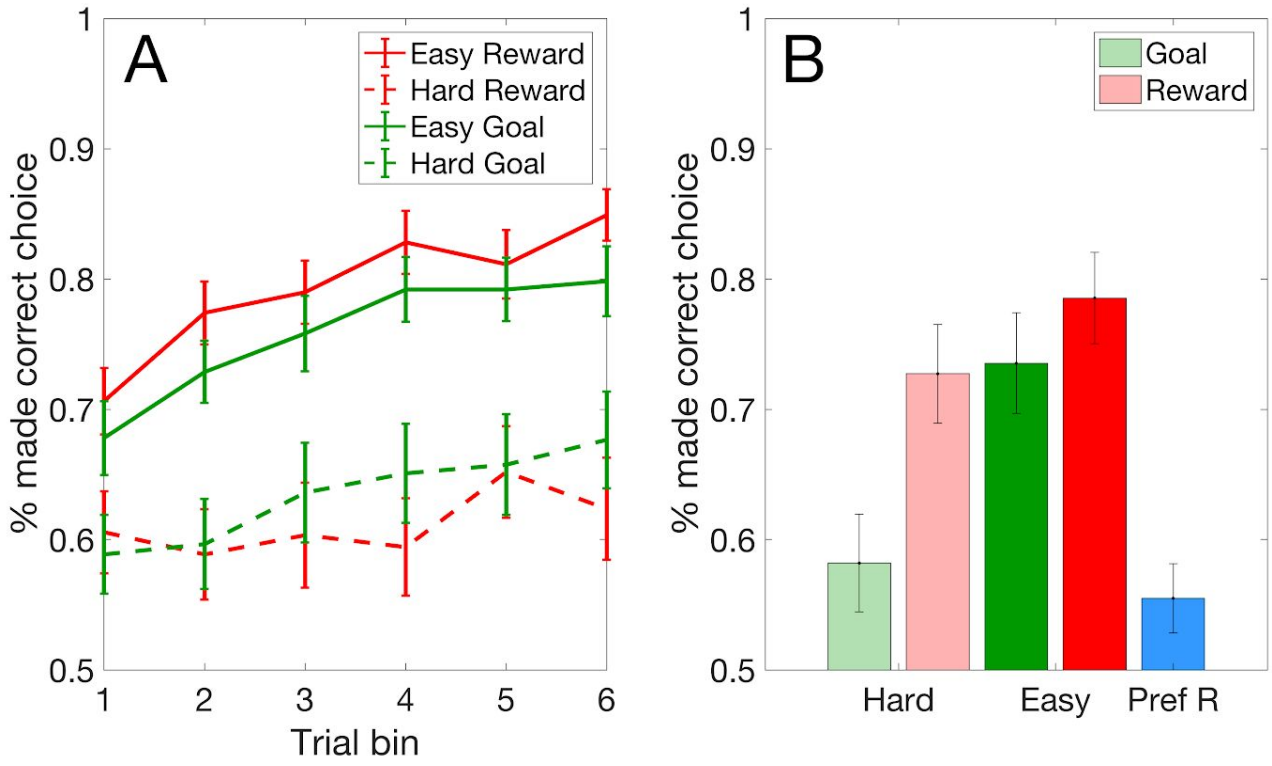


Figure 2: a) Learning curves from the training phase for the four conditions. **b)** Test phase performance for four conditions and preference for reward.

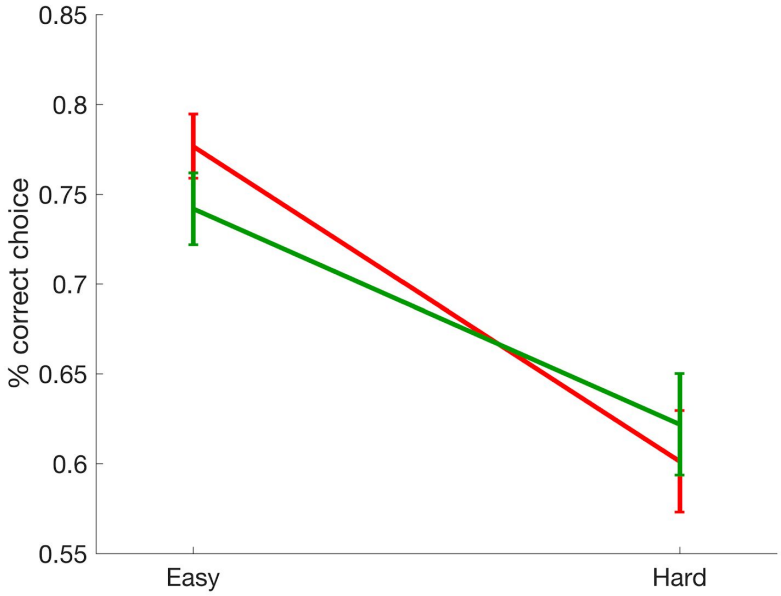


Figure 3: Learning phase performance averaged over time and separated along two dimensions: outcome type and uncertainty.

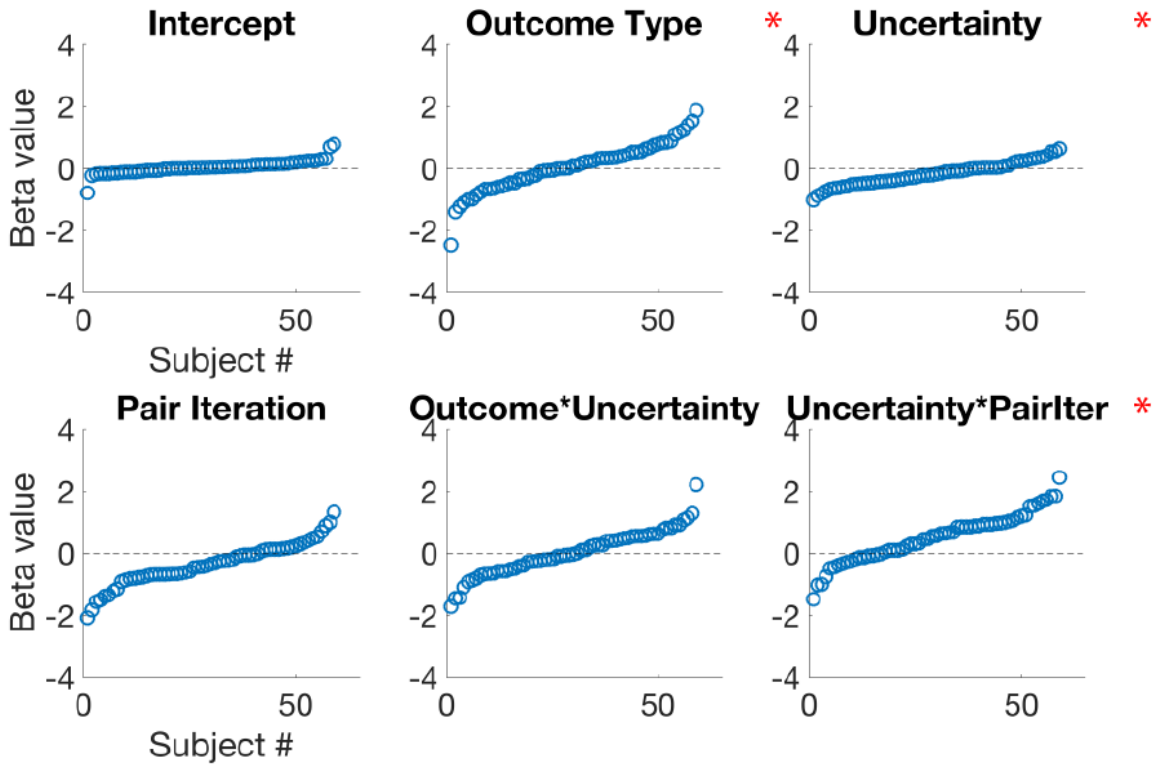


Figure 4: Beta weights sorted in ascending order from a logistic regression model predicting choice based on learning phase data. Regressors which were significant when t-tested against 0 have a red star next to the regressor name.

Results from 59 participants showed that they were able to learn the value of the stimuli from both rewards and pseudo-rewards as outcomes. The 360 trials of the learning phase were split into 6 time bins for analysis. There was an increase in proportion of correct choices from the first time bin to the last with no significant difference between reward and goal conditions at any trial bin (Figure 2; all $p > 0.10$), but a significant effect of uncertainty across the learning phase (Figure 2; all $t(58) > 2.23$ and all $p < 0.05$). Uncertainty is determined by the difference in the true expected value between the presented boxes. There is greater uncertainty for box pairs that have less value difference between each other. We averaged over trial bins in the learning curves to remove the time component, and find that there was no interaction of condition and uncertainty on the overall proportion of correct trials ($t(58) = 1.29$, $p= 0.20$), such that goal trial performance was less sensitive to conflict than reward-based learning (Figure 3). There may have been worse discrimination between easy and hard trials in learning from pseudo-rewards, but it was a non-significant difference. We predicted learning phase choices using a multiple linear regression model including regressors: outcome type (reward or pseudo-reward), uncertainty defined as the absolute value of the difference in expected value of the boxes, the number of iterations the pair of boxes has been presented, the interaction between reward and goal and uncertainty, and the interaction between uncertainty and pair iteration. Testing beta weights for each regressor against 0, the only regressors that had a significant effect were uncertainty, pair iteration, and the interaction between uncertainty and pair iteration (Figure 4; $t(58) = -3.52$, $t(58) = -3.91$, $t(58) = 4.98$, respectively; all $p < 0.001$). Consistent with our previous analysis (Fig. 3), there was no interaction of outcome type and uncertainty.

For test phase analysis, we first split trials into four conditions along two dimensions: outcome type and difficulty. Trials with pairs previously seen during the learning phase were excluded from analysis as the "correct" for that trial could be arrived at from learning an action policy as opposed to integrating reward histories to compute an expected value for each item in the pair. Integration of reward histories is

characteristic of the RL system, so it is critical to test that this is the mechanism generating correct choices. Easy pairs have a larger value difference relative to hard pairs making it easier to identify the higher-valued choice within a pair. Participants performed above chance in all conditions, with a greater proportion of correct choices on easy trials as compared to hard trials (Figure 2). To understand what factors underlie a participant's choice and how much weight these factors have relative to one another, we implemented a logistic regression model predicting the choices of participants during the learning phase. Results from a logistic regression model support that value was an important factor in determining choice (Figure 3; $t(58) = 12.92$, $p < 10\text{-}4$) as well as previous trial choice (left or right) Figure 3; $t(58) = -2.07$, $p = 0.0428$). There was no interaction between value difference and value mean (Figure 3; $t(58) = -0.976$, $p = 0.333$).

2. Individuals learn differently from rewards and pseudo-rewards as outcomes

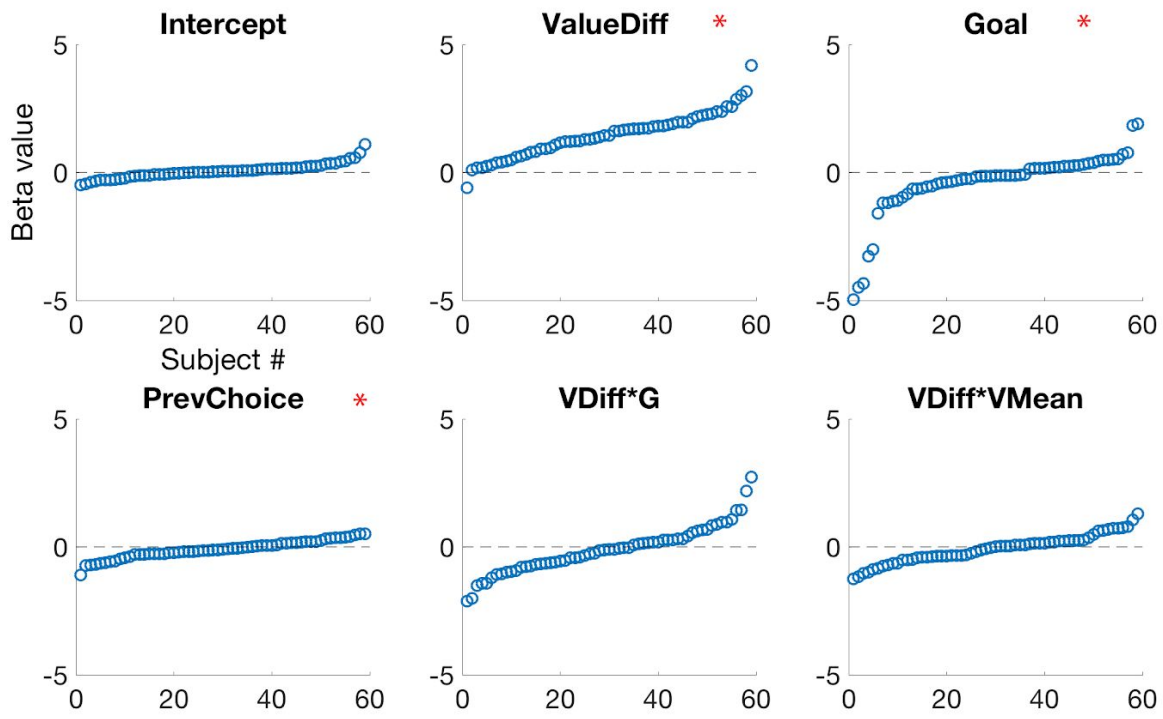


Figure 5: Beta weights sorted in ascending order from a logistic regression model predicting choice based on test phase data. Regressors which were significant when t-tested against 0 have a red star next to the regressor name.

Test phase analysis found no effect of outcome type on easy trial performance, but a significant effect on performance on hard trials (Figure 2; t(58) = -2.75, p = 0.0080). One test phase trial of particular interest was one in which a reward associated box and a goal achievement associated box with the same expected value (assuming the value of pseudo-reward is 1) are paired together. A systematic preference for choosing the reward or goal box would have indicated attaching a higher value to that type of outcome relative to the other. A significant preference for reward was present (Figure 2; t(58) = 2.08, p = 0.042). From beta weights obtained from the test phase logistic regression, participants were found to be less inclined to choose a goal box over a reward box regardless of value difference (Figure 5; t(58) = -2.44, p = 0.018). However, the effect of the interaction between value difference and goal was not significant (Figure 4; t(58) = -0.87, p = 0.39).

3. There are individual differences in preference for rewards over pseudo-rewards

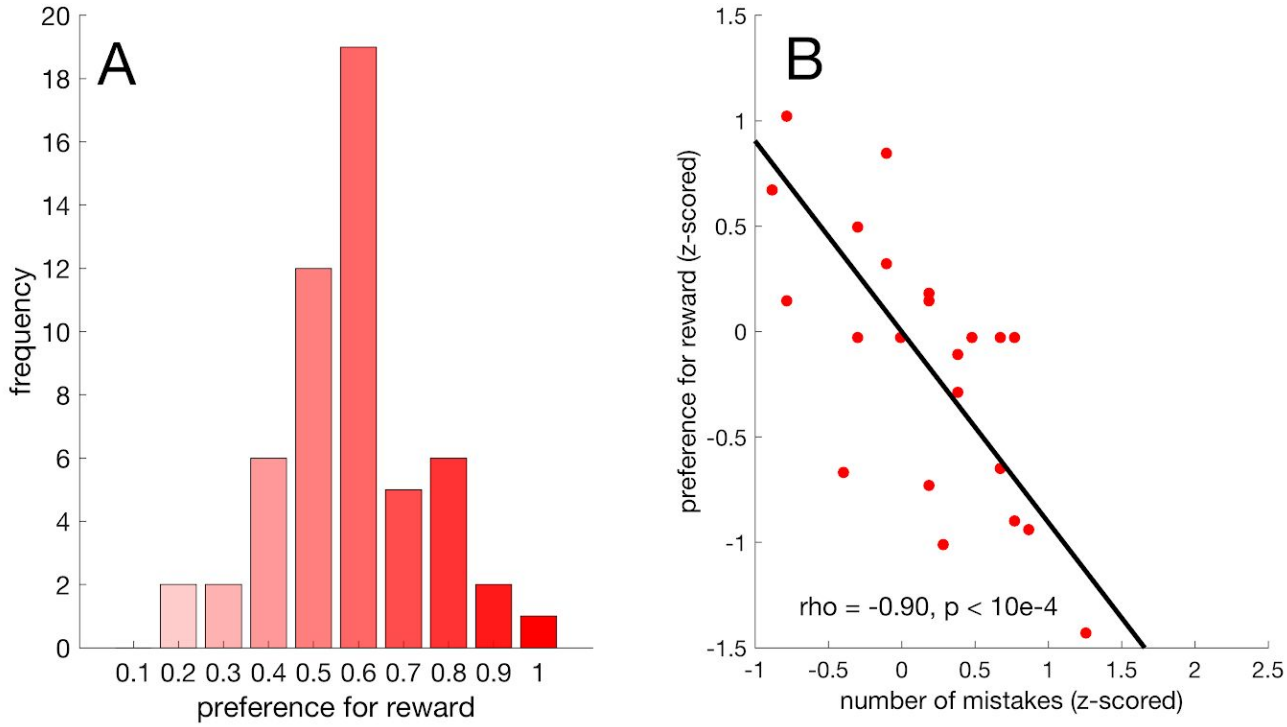
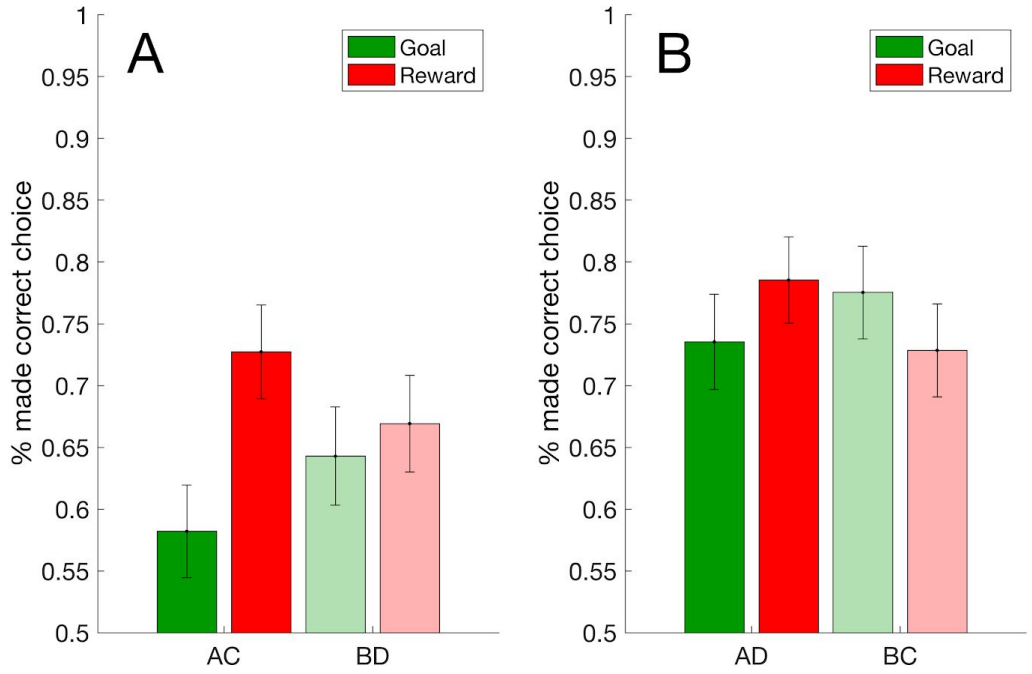


Figure 6: a) Histogram with each bar representing the number of subjects with that behavioral measure of preference for reward over pseudo-reward **b)** Correlation between preference for reward and number of mistakes made on trials with a reward box of higher value than the goal box, but the participant chose the goal box instead. A mistake in this case is considered choosing the lower valued goal box.

While there was a group-level preference for rewards over pseudo-rewards, the preference for reward measures were normally distributed, centered on 0.6, rather than being skewed towards 1 (Figure 6A). This indicates that there are individual differences in the value attached to pseudo-rewards. There was a strong correlation between a subject's preference for reward measure and the number of goal-biased mistakes made on trials with a reward box that was a higher expected value than the goal box it was paired with, after removing subjects who made no mistakes (Figure 6B; Spearman $\boldsymbol{\rho}$ = -0.9, p < 10-4). This suggests that these participants were systematically attaching an internally-defined additional value to goal achievement boxes over reward boxes.

4. Rewards and pseudo-rewards are differentially valued gains, but are equivalent losses



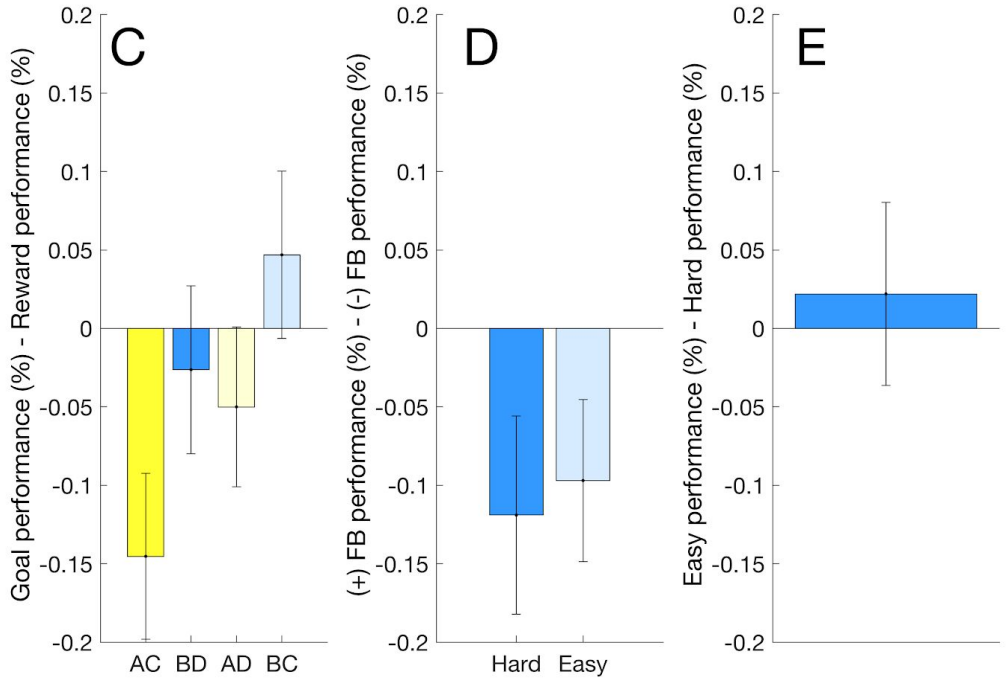


Figure 7: a) Test phase performance for AC and BD trials (hard condition) only, separated by reward and goal condition trials. **b)** Same as a) but for AD and BC trials (easy condition) only. **c)** Effect of outcome type plotted as the difference in percent choices correct for AC, BD, AD, and BC trials. **d)** Effect of valence plotted as the difference in percent choices correct for hard (AC - BD performance) and easy (AD - BC performance) trials. **e)** Effect of difficulty on performance plotted as the difference between hard and easy trials

We further refined our analysis of test phase performance by separating hard trials into AC and BD trials. In AC trials, both boxes presented have been attached to positive outcomes more so than negative while the opposite is true for boxes B and D. Separating these trials gives insight into differences in how participants learn from positive and negative outcomes. They performed significantly better on AC reward trials compared to AC goal trials, but there was not a significant difference in performance on BD trials (Figure 7A,C; $t(58) = -2.75$, $p = 0.008$; $t(58) = -0.49$, $p = 0.625$). The majority of the effect of outcome type in hard trials was accounted for by differences in response to positive outcomes. In other words, the effect was driven by differences in recieving a reward or a pseudo-reward instead of differences in not recieving a reward or not recieving a pseudo-reward. Easy trials further separated into AD and BC trials. AD as a pair had a higher average value relative to BC. There is not an effect of outcome type in either AD or BC trials (Figure 7B; $t(58) = -0.98$, $p = 0.33$; $t(58) = 0.88$, $p = 0.38$), but

there was an interaction between outcome valence and outcome type (Figure 7D; t(58) = -2.17 , p= 0.034). This indicates that rewards and pseudo-rewards are valued differently as gains, but are equivalent losses.

Modeling

To understand the mechanisms producing the behaviors seen during pseudo-reward dependent learning, we use computational modeling. This method allows us to test if the same computations performed by the RL system on rewards are also performed on pseudo-rewards and it allows for comparison between candidate mechanisms for generating the difference in learning from the two types of outcomes.

A preference for reward parameter is necessary to capture differences in learning from rewards and pseudo-rewards.

Model Comparison

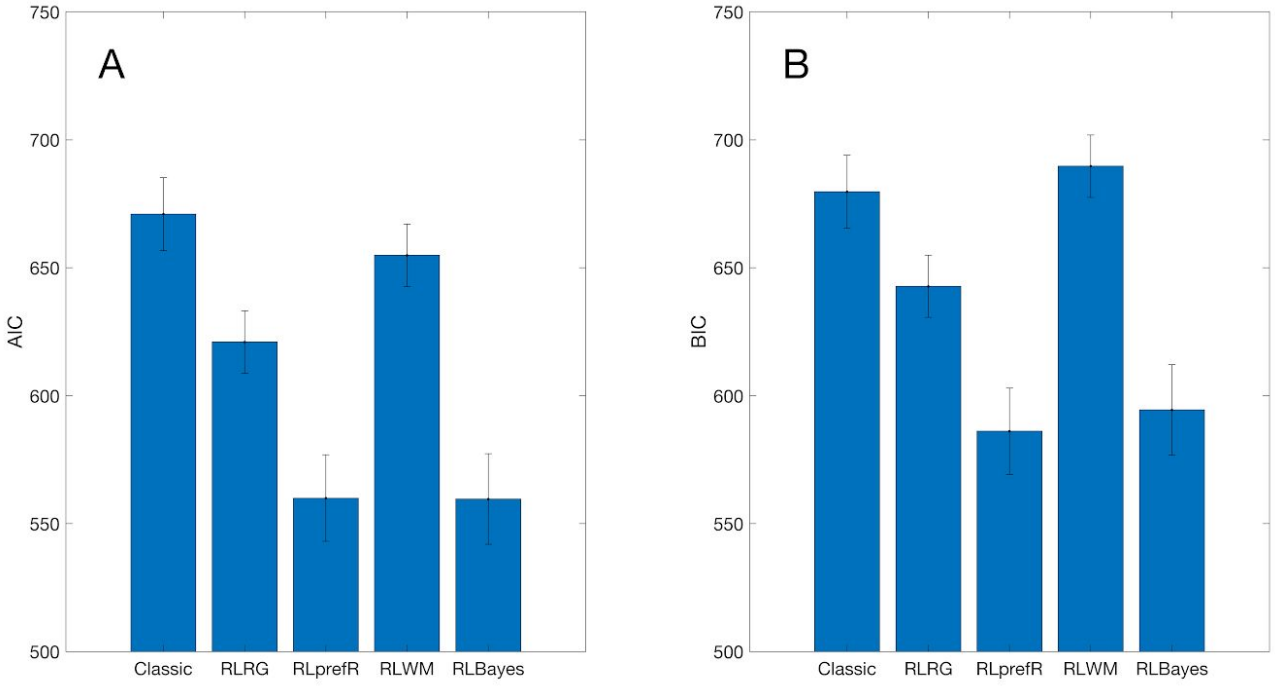


Figure 8: a) average AIC scores for each considered model **b)** average BIC scores

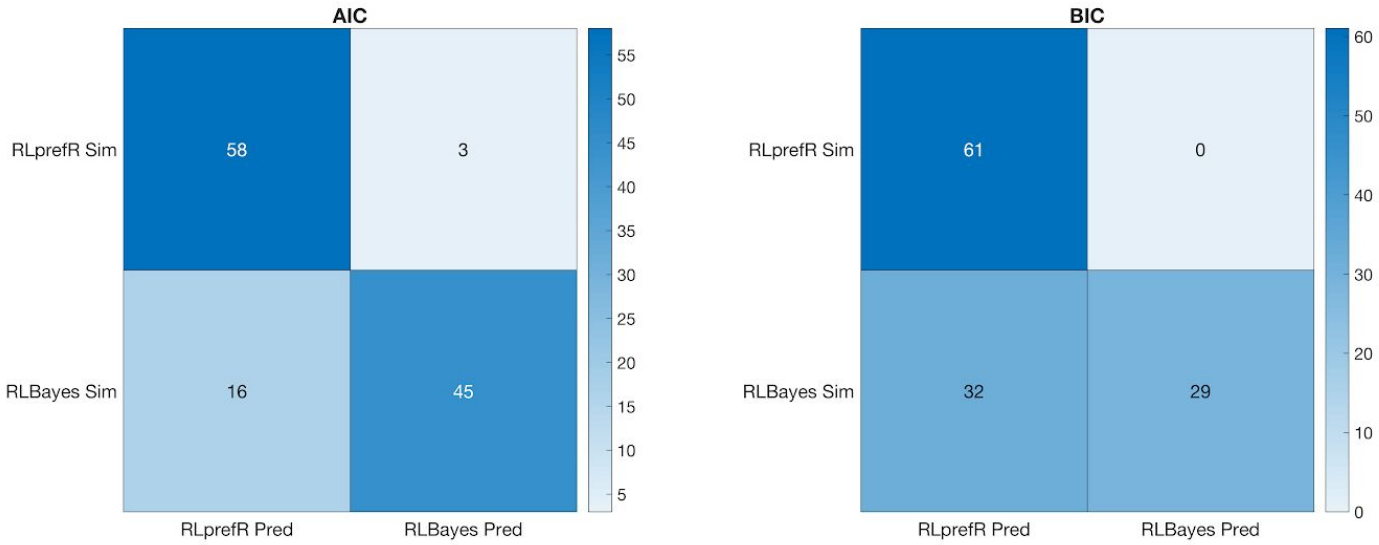


Figure 9: Confusion Matrices for AIC and BIC for 61 subjects. Sim indicates that a data set was generated using that model, and Pred indicates that was the model predicted by AIC or BIC, after fitting the simulated data. The predicted model is the model with the lower of the two scores, meaning it better fit the data. The accurate predictions are along the left-right diagonal. We see that AIC misidentifies the correct model less than BIC for the RLBayes model. BIC is over penalizing RLBayes' extra parameters.

To compare the goodness of fit between models, we use AIC and BIC. The models with the lowest AIC and BIC were the RLprefR and RLBayes models (Figure 8). To determine which measure to rely on, first data is simulated using the best models. Then, the simulated data is fit to both the model it was generated from and the other model. Then, the measures can be compared according to highest number of accurate predictions and lowest number of misidentifications. Concretely, this means it computes a lower score for the model that actually generated the data. AIC produces the fewest mistakes, particularly for the RLBayes model (Figure 9). BIC overpenalizes its extra number of parameters. There is a marginal difference between the average AIC scores of the best models, RLprefR and RLBayes, but a large difference in average AIC scores with the three other candidate models (Figure 10A-B). We identify the model that best fits each individual's behavior by picking the model with the lowest AIC score for that individual and find that the RLprefR and RLBayes models are the only models which are picked out as providing the best fit for individuals (Figure 10C).

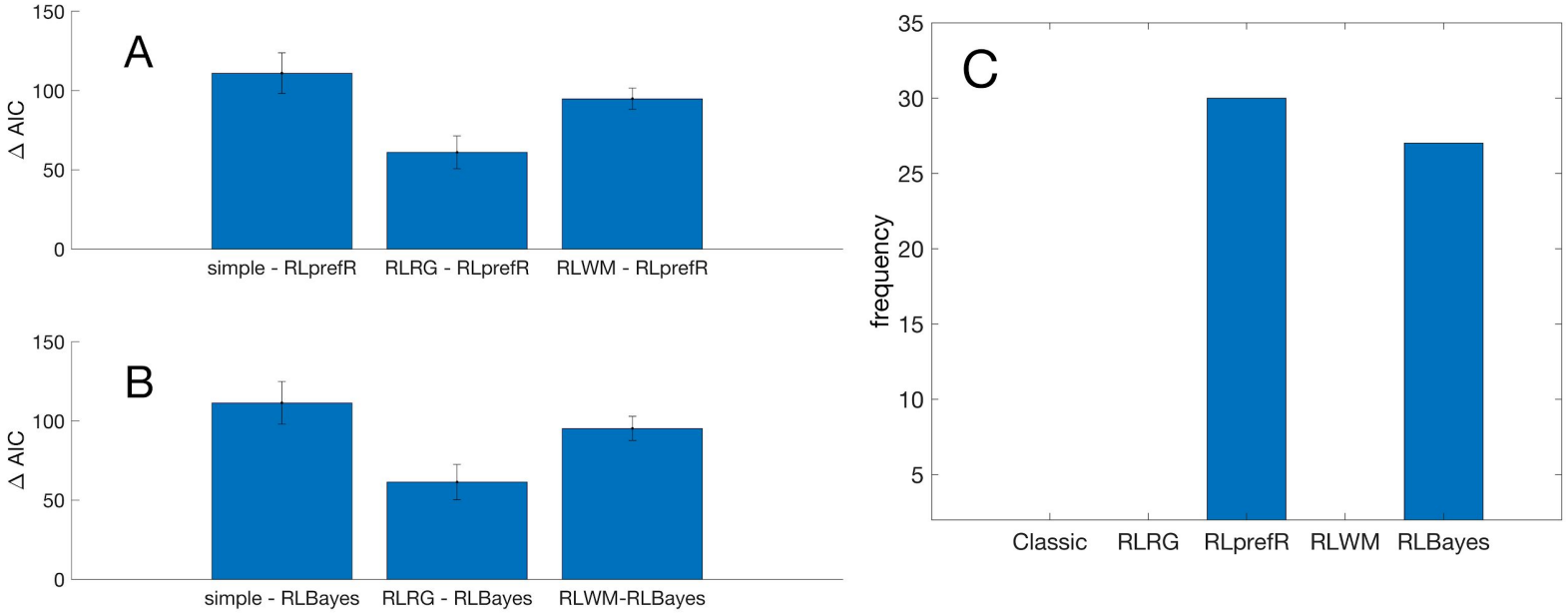


Figure 10: a-b) The difference in AIC score between RLprefR and RLBayes, respectively, and the other considered models. **c)** histogram of the best fitting model based on AIC score for each subject

We performed a "generate and recover" procedure to gauge the efficacy of our parameter estimation process for a particular model, and then compare between models of interest. Parameters estimated from real participants' data were used to simulate a data set, thus the true parameters that produced the data were known. Then, a parameter fitting procedure was performed on the learning and test phase of the simulated data, and recovered parameters are compared to those that generated the data. There was a strong correlation between recovered and actual parameters (Figure 11; RLprefR: all $rho > 0.73$, all $p < 0.0001$; RLBayes: all $rho > 0.56$ except alphaN, $rho = 0.33$, all $p < 0.0001$). Recovery tends to be better for models with fewer parameters, which is consistent with our better recovery with the RLprefR model.

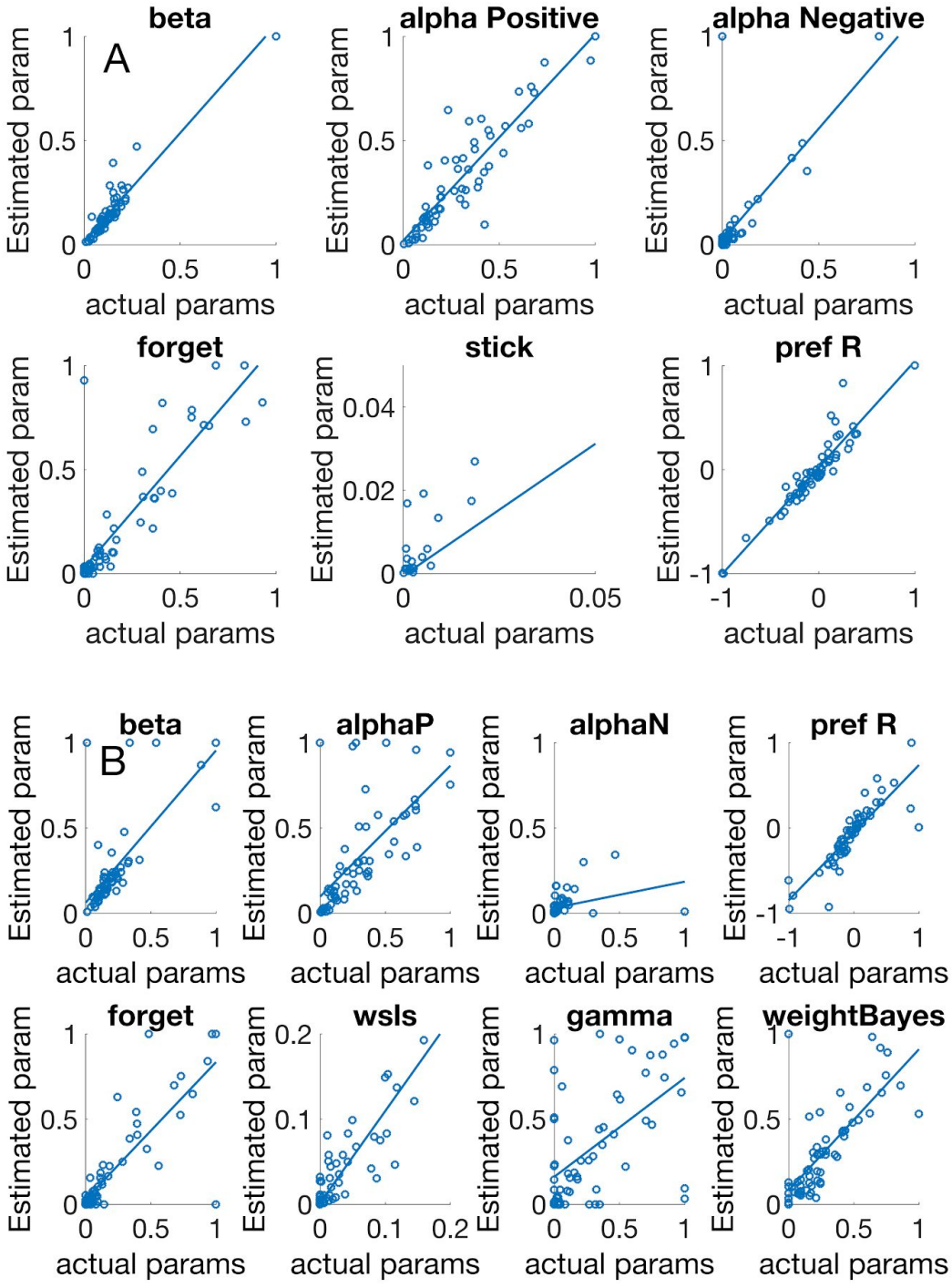


Figure 11: Correlation between parameters used to generate simulated data and the parameter estimated from our fitting procedure. **a)** RLprefR model **b)** RLBayes model

Model Validation

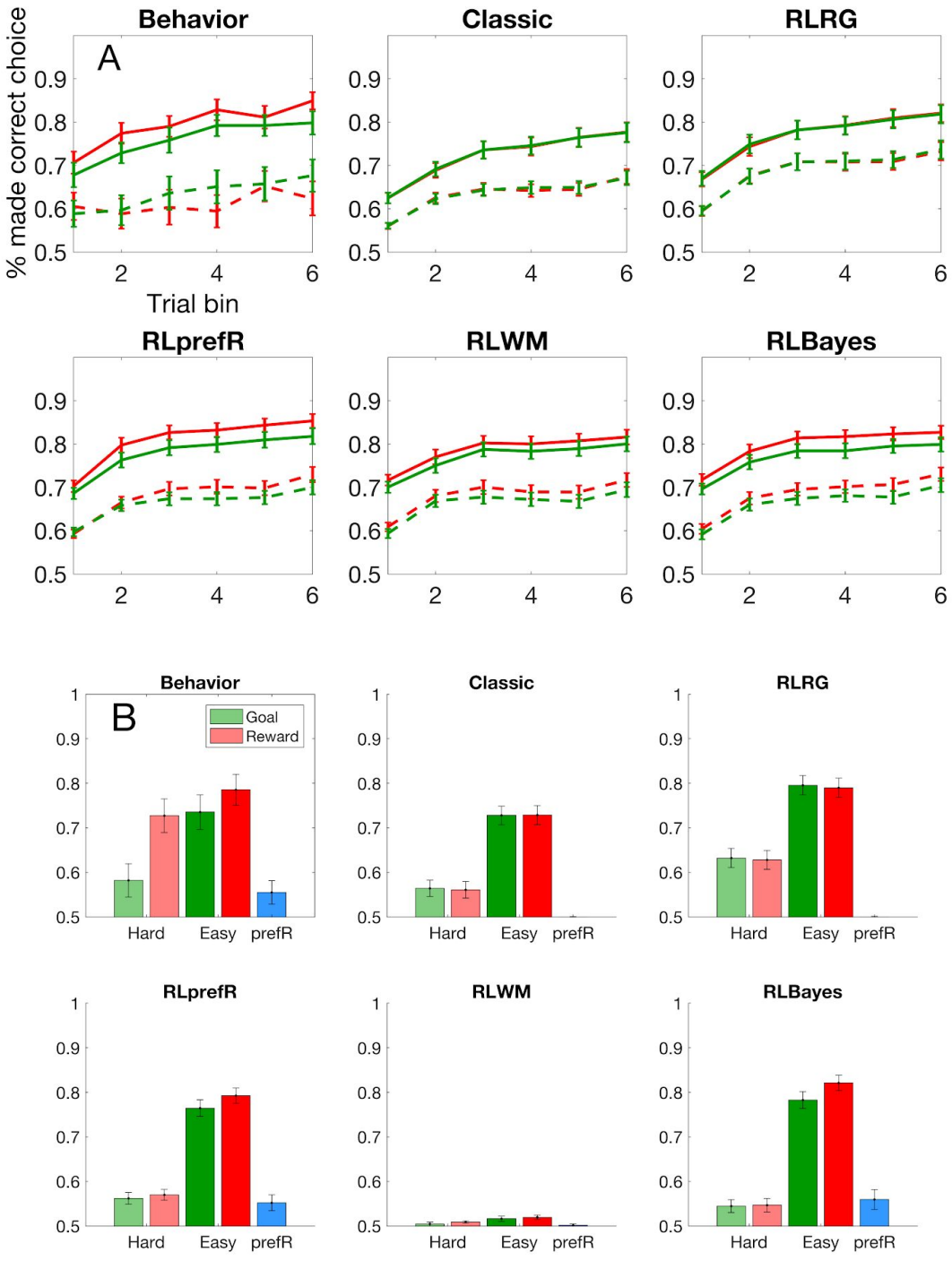


Figure 12: Comparison between model simulation performance and behavior from participants **a)** Learning Phase performance **b)** Test Phase performance

With model validation, we check to see if the best models are able to capture important features of behavior through simulating data using parameters fit to participants' behavioral data (Figure 12). Separate learning rates for learning from rewards or pseudo-rewards do not replicate key aspects of behavior. The models with the lowest AIC, RLprefR and RLBayes, are the only models able to produce the reward and goal learning curve separation and a preference for reward over pseudo-rewards. This implicates that the mechanism generating these differences in learning from the two outcomes is that they are differentially valued. To further validate our models, we compare estimated parameters with their corresponding behavioral measures. Of models with a preference for reward parameter both the RLprefR and RLBayes models produce parameters that are strongly correlated with the behavioral measure from the test phase (Figure 13; Spearman, ρ = 0.66, p < 10e-4; ρ =0.58, p < 10e-4, respectively).

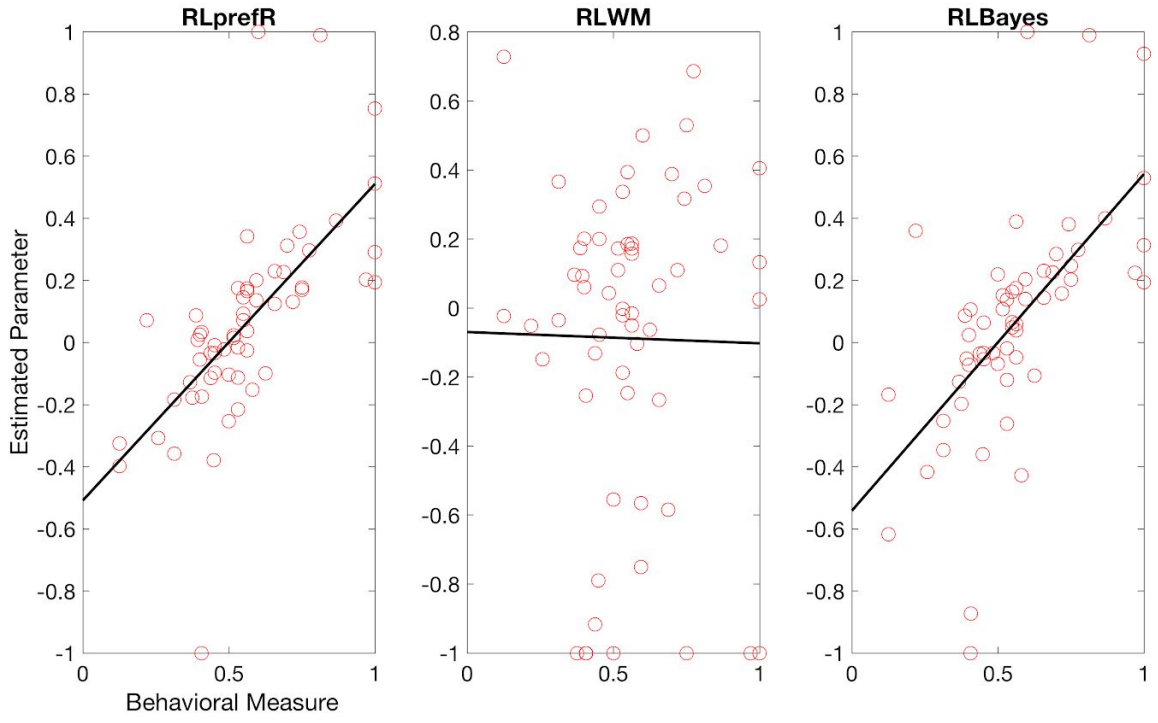


Figure 13: Correlation between behavioral preference for reward measure and the model extracted parameter for each model with a preference for reward parameter. RLprefR (rho = 0.66, p < 0.001), RLWM (rho = -0.02, p = 0. 89), RLBayes (rho = 0.58, p < 0.001)

It is important to note that there were certain aspects of behavior that no model was able to capture. These include the reduced effect of uncertainty for the goal condition in the learning phase, the significantly better performance on reward trials for the hard condition of the test phase, an effect of uncertainty for win-stay lose-shift behavior, and the difference in AC trial performance between reward and goal only trials (Figure 12AB, 14AB). For future work, it will be important to identify this additional mechanism that is currently missing in our models.

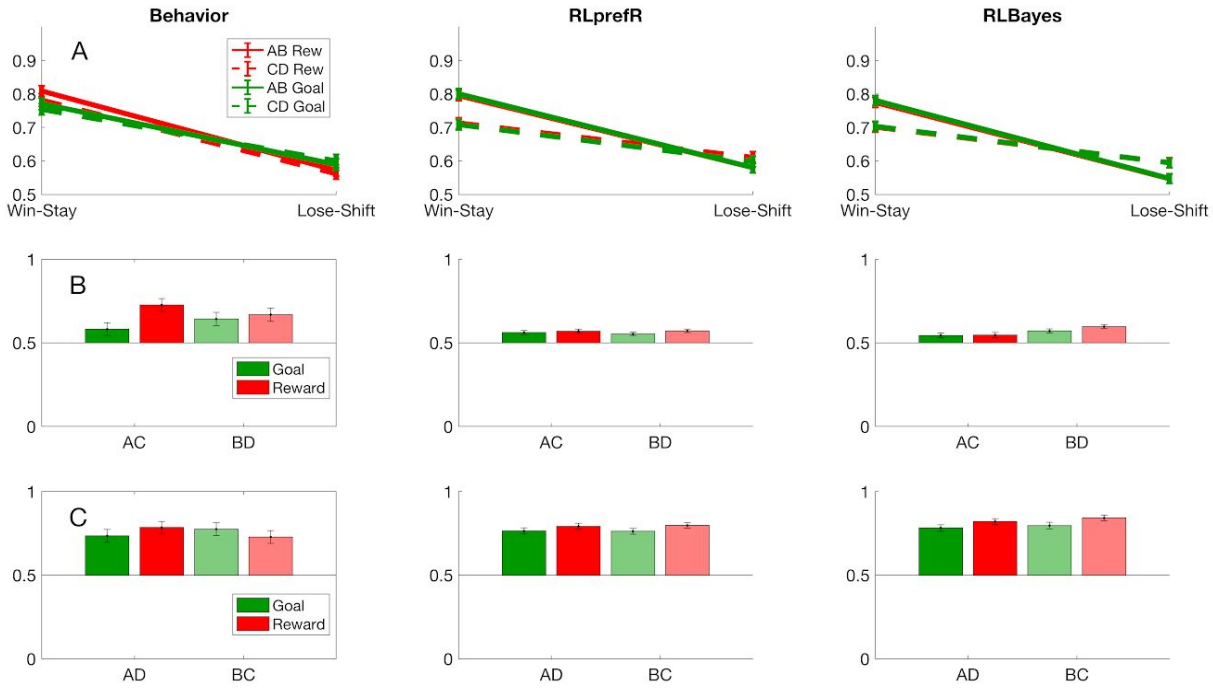
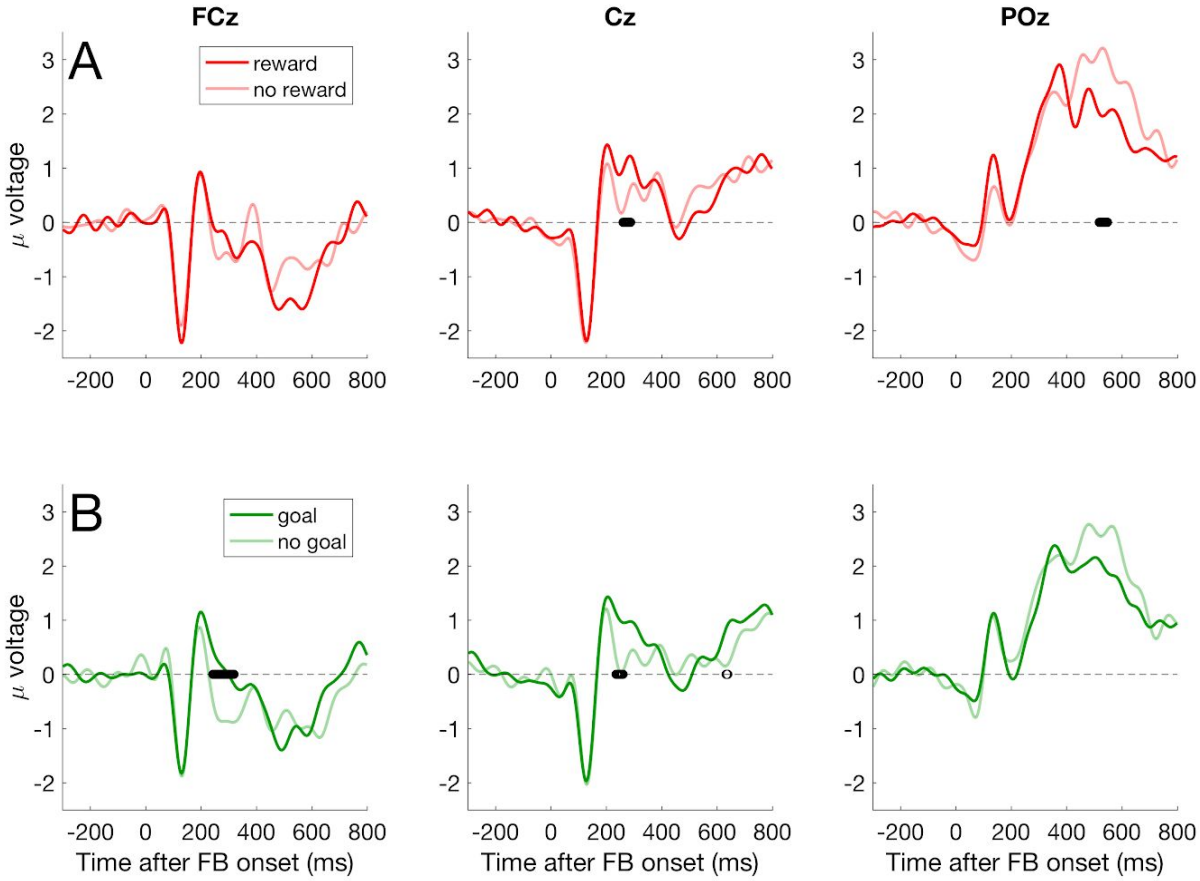


Figure 14: Comparison between models' simulated performance and participants' performance **a)** win-stay lose-shift behavior **b)** test phase performance for AC and BD trials separated by reward and goal only box pairs **c)** same as b) except with AD and BC trials

EEG

From modeling, we have a better understanding of the computational mechanisms underlying learning from these two types of outcomes. But, to understand the neural mechanisms that generate this behavior, we use EEG to test if the same neural mechanisms responsible for learning from reward underlie learning from pseudo-rewards.

1. *There is a robust effect of outcome valence but not of outcome type on neural signal*



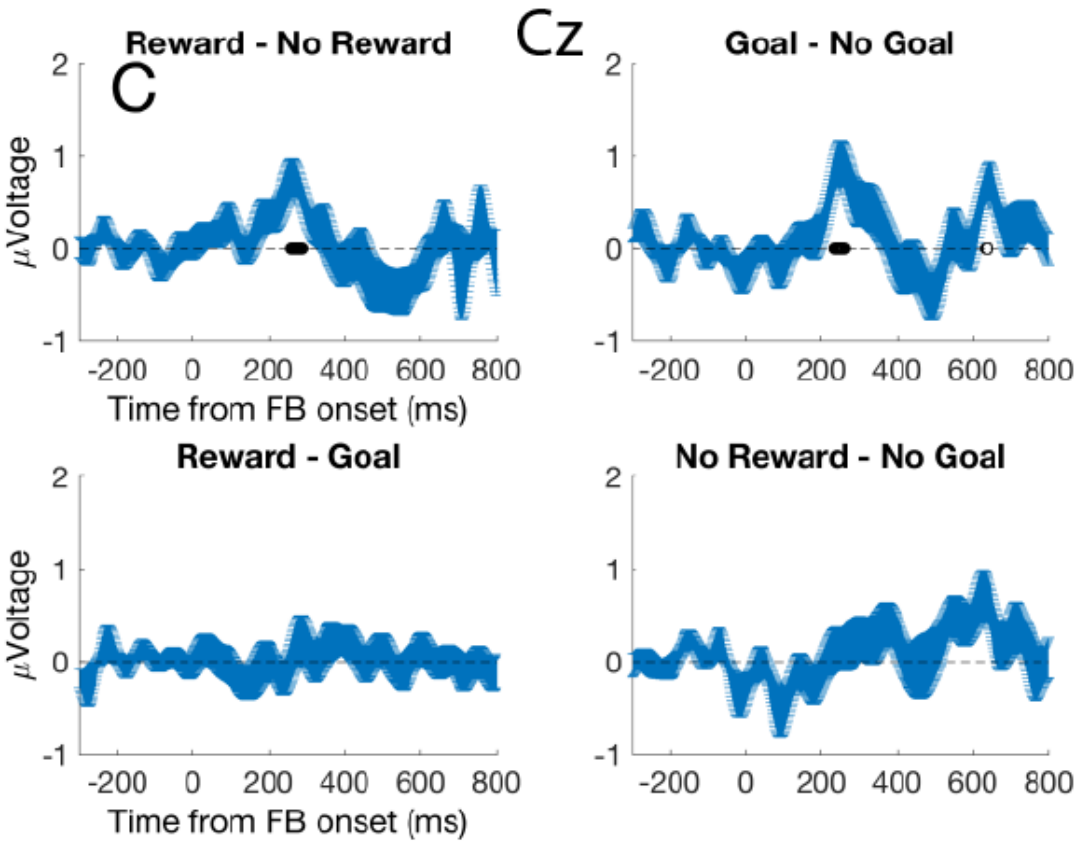


Figure 15: ERPs **a)** ERPs in locked to events of outcome valence reward and no reward plotted together. Time points of significant difference (threshold = 0.001) are plotted in black **b)** same as a) but ERPs locked to events of outcome type goal and no goal. **c)** Difference waves from electrode Cz comparing each of the four events to each other. Time points of significant difference (threshold = 0.001) from 0 are plotted in black.

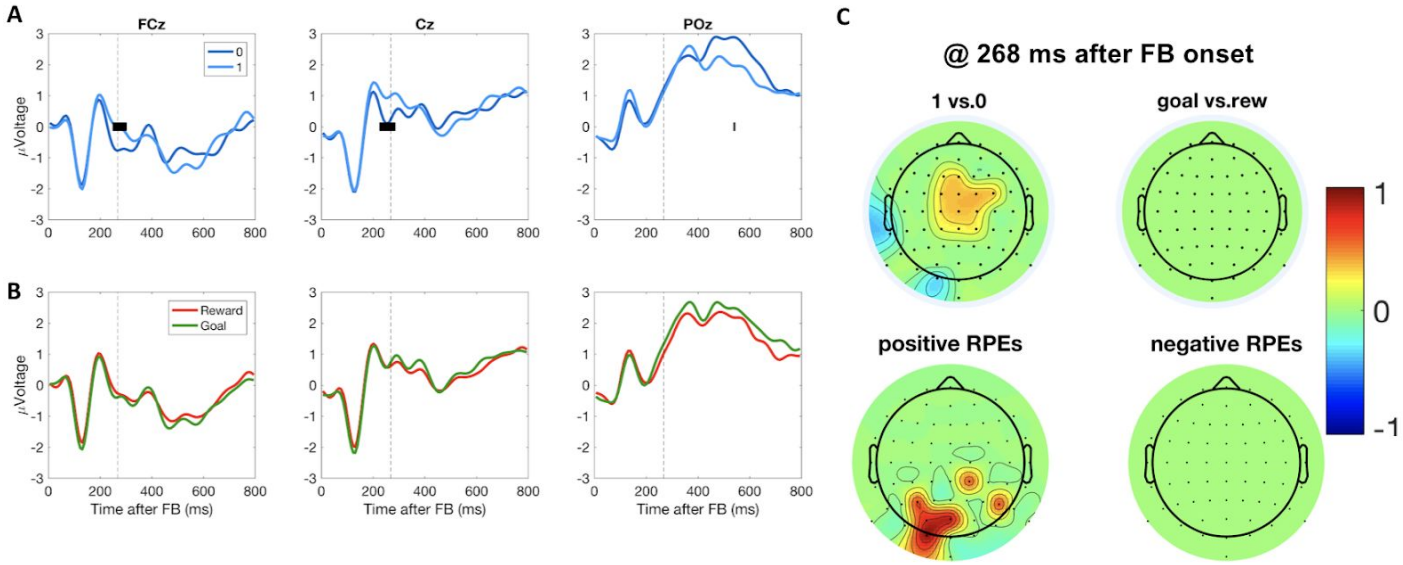


Figure 16: Corrected ERPs plotted to visualize the effects of two regressors of interest. The dashed line is at the time point the scalp topography is observed at. . In black, we plot time points of significant difference (threshold = 0.001) . **A)** outcome valence

(positive or negative) **B)** outcome type (reward or goal) for three electrodes (FCz,Cz,POz) **C)** Scalp topography at 268 ms after feedback onset, plotting regions with significant thresholded regression weights at that time point.

For feedback-locked analysis, we replicate findings that fronto-central electrodes are sensitive to the distinction between positive and negative outcomes (Frank et al., 2005; Holroyd & Coles, 2002; Figure 15). We see this effect in both the reward and goal condition. However, there are no electrodes are sensitive to the distinction between reward and pseudo-reward (Figure 15). We use a multiple regression approach to gauge the strength and spread of effect of variables of interest. Using a multi-regressor linear model to predict voltage allows for measurement of continuous effects that influence the signal such as RPEs which evolve trial-by-trial. Variables of interest included valence of outcome (positive or negative), type of outcome (reward or pseudo-reward/goal achievement), and positive and negative RPEs. From our RLprefR model, the RPE is extracted trial-by-trial for each subject. There is a strong and widespread effect of valence on fronto-central electrodes at the same time point of 260 ms (Figure 16) reinforcing our findings from basic ERP analysis and replicating previous findings using this multiple regression approach (Collins et al., 2016; Collins et al., 2018). Positive RPEs have a weaker and more localized effect. The area sensitive to this effect is more posterior than areas previously found (Collins et al., 2016; Collins et al., 2018). It is most likely noise and would be unable to withstand correction for multiple comparisons. No region was sensitive to negative RPEs.

Taken together, we were unable to identify a region differentially sensitive to rewards and pseudo-rewards similar to results from basic ERP analysis nor from the multiple regression approach.

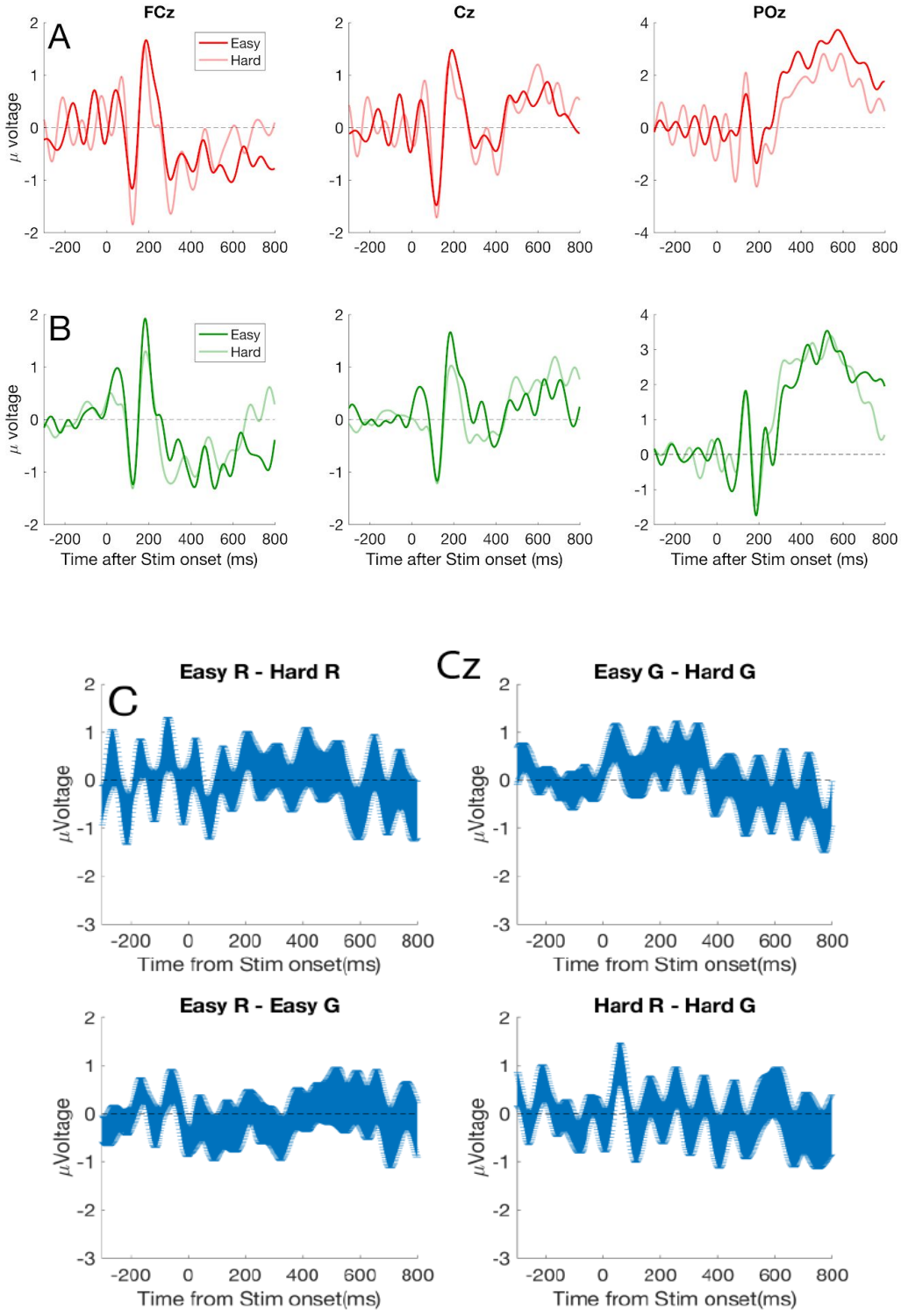


Figure 17: ERPs locked to stimulus presentation **a)** We plot reward easy pairs and hard pairs together. Time points of significant difference (threshold = 0.001) are plotted in black **b)** same as a) but goal easy and hard pairs **c)** Difference waves

from electrode Cz comparing each of the four events to each other. Time points of significant difference (threshold = 0.001) from 0 are plotted in black.

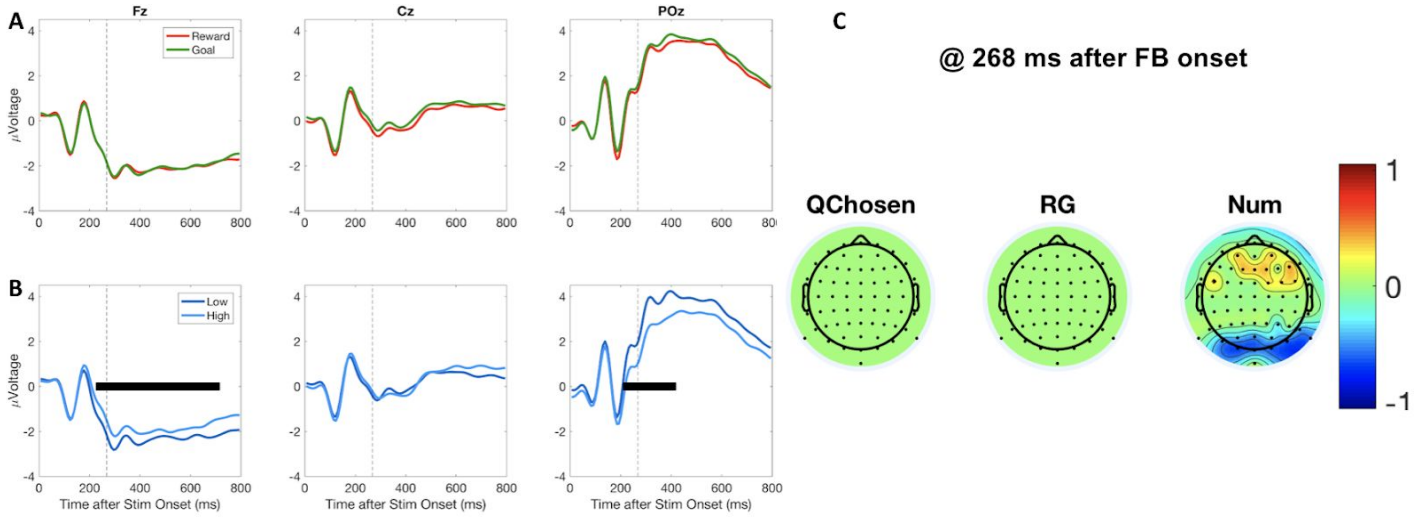


Figure 18: Results from multiple regression. The dashed line is at the time point the scalp topography is observed at. **A-B)** Corrected ERPs b) Effect of type of outcome tied to stimuli c) effect of number of times pair presented. In black, we plot time points of significant difference between the two ERPs (threshold = 0.001) **C)** Scalp topography at 268 ms for three variables of interest: Q-value of chosen option, type of outcome tied to stimuli (reward or pseudo-reward/goal achievement), and number of times pair has been presented.

2. Number of times choices were presented produces a robust effect on neural signal while the value of the chosen option does not

For stimulus-locked analysis, we compare the ERPs for when easy box pairs (defined as having a large difference in expected value) are presented versus the response to hard box pair presentation. Surprisingly, we did not find an effect of difficulty nor of type of box pair (reward or goal associated) in our three electrodes of interest (Figure 17). A multiple regression approach was used as was done feedback-locked analysis to extract the contributions of Q-values, type of outcome tied to stimuli (reward vs. pseudo-reward), and number of iterations a pair has been presented to the signal. Our Q-values are the value of the chosen stimuli which is estimated trial by trial for each subject using our RLprefR model. There is an effect of number of presentations on frontal and posterior electrodes at approximately 224 ms after stimulus onset (Figure 18, threshold = 0.001). This effect reflects a familiarity with the stimuli. The Q-value of the chosen option did not yield significant effects contrary to results from Collins and Frank

(2018) in which they found robust markers of Q values. The type of outcome tied to stimuli also did not produce a significant effect (Figure 17).

3. Rewards or goal achievement as outcomes produce differences in theta power

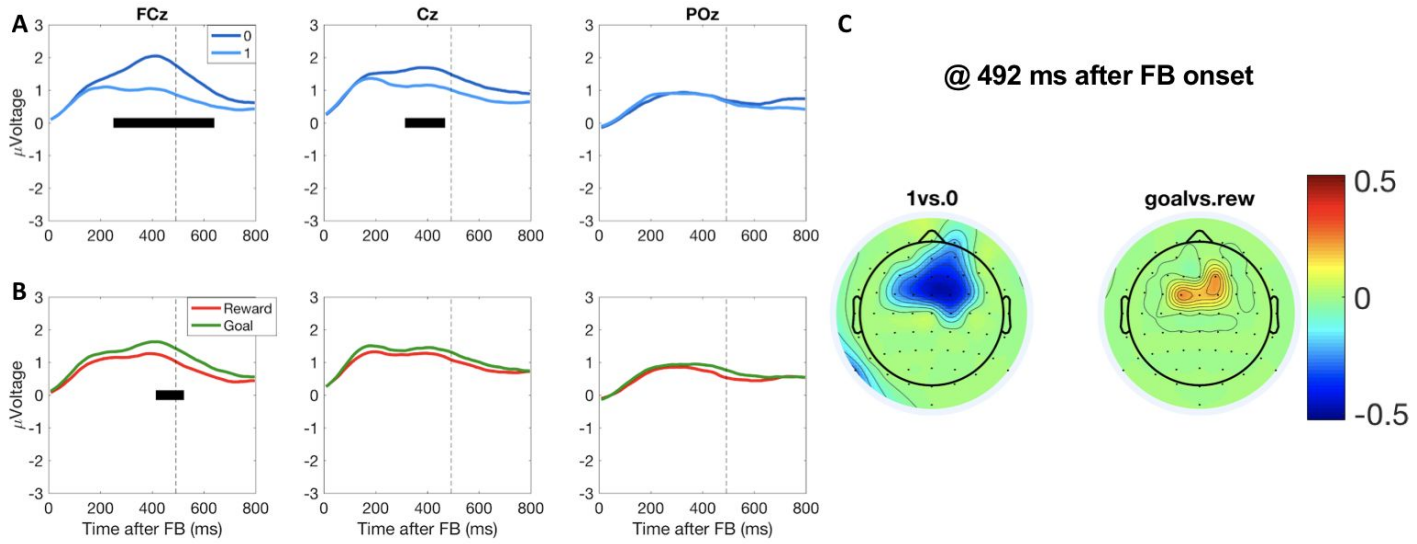


Figure 19: A-B) Corrected Theta Power over time. The dashed line is at the time point the scalp topography is observed at. **a)** outcome valence. Time points of significant difference between positive and negative outcomes are plotted in black (threshold = 0.001). **b)** Outcome type (reward or goal)**.** Time points of significant difference are plotted in black. **c)** Scalp topography for outcome valence and type of outcome (reward or pseudo-reward/goal achievement) at 492 ms.

Frontal midline theta (FMθ) has been suggested to reflect a common computation used to realize the need for control across several contexts (Cavanagh et al., 2011, Cavanagh and Frank, 2014) . It serves dual roles as a call for further control and as a teaching signal. Potentially, FMθ may reflect a common feature amongst several ERP components including ERN, FRN, and CRN, all of which are mapped on to cognitive processes that include the need for control. Cavanagh and colleagues (2011) found theta power to be more sensitive to between condition differences than ERPs, producing larger effect sizes. Because we did not find differences between the reward and goal conditions in our ERP analysis, we additionally analyze theta power to see if we can uncover differences that may have been obscured in those analyses. The same multiple regression approach was used for both feedback-locked and stimulus-locked analysis, however, instead predicting voltage the GLM predicts theta power.

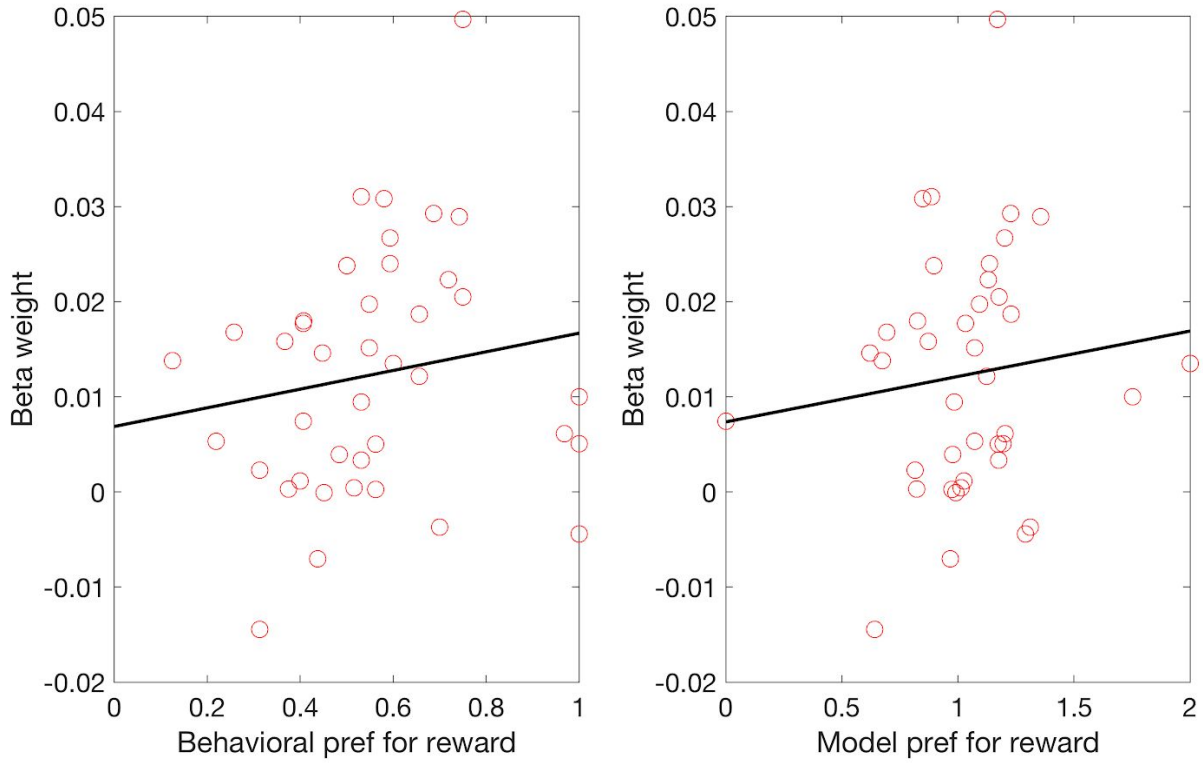


Figure 20: Pearson's correlations between behavioral measure for preference for reward, preference for reward parameter from model RLprefR, and averaged beta weights from group-level ROI. $\boldsymbol{\rho} = 0.16$, $p = 0.32$; $\boldsymbol{\rho} = 0.12$, $p = 0.47$, respectively.

For feedback-locked analysis, variables of interest include outcome valence and outcome type (reward or pseudo-reward). There was a robust and widespread effect of outcome valence (positive versus negative) across fronto-central electrodes around 485 ms (Figure 19) replicating previous findings (Cavanagh et al., 2010; Cavanagh et al., 2011). Those fronto-central electrodes at the same time point were also found to be sensitive to the distinction between reward and pseudo-rewards as outcomes (Figure 19). This region has been found, previously, to be sensitive to the valence of outcomes(Holroyd & Coles, 2002; Botvinick et al., 2004; Yeung et al., 2004; Frank et al., 2005), converging with our results from computational modeling indicating that the behavioral differences in learning from rewards and pseudo-rewards arise from the different valuations of the outcomes. To explore links between results from behavior, modeling, and EEG, we correlated the weighted average of beta weights for the electrodes in our region of interest

for the outcome type regressor with the behavioral measure for preference for reward and with the preference for reward parameter extracted from the RLprefR model. There was not a strong correlation between EEG beta weights and either the behavioral measure or preference for reward parameter (Figure 20; Pearson, $\boldsymbol{\rho} = 0.16$, p = 0.32; $\boldsymbol{\rho} = 0.12$, p = 0.47).

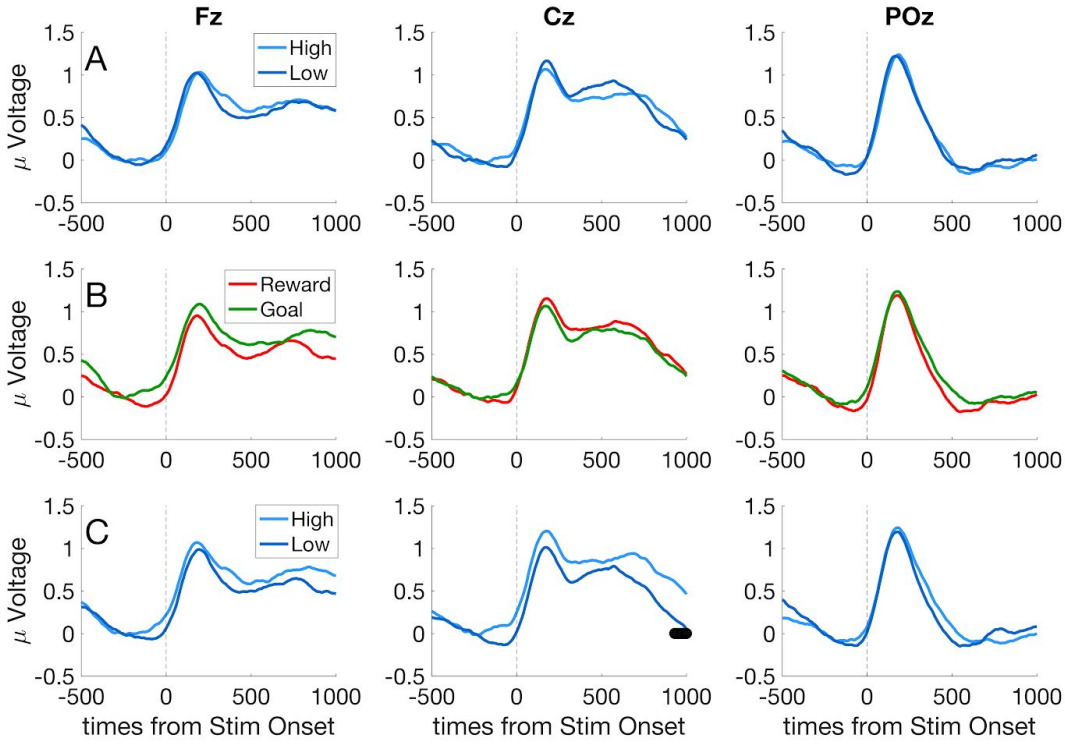


Figure 21: Theta power over time with time points of significant difference plotted in black **a)** Q-Chosen regressor **b)** RG regressor **c)** Num regressor. Time points of significant difference are plotted in black (threshold = 0.001).

For stimulus-locked analysis, variables of interest included the Q-value of the chosen option (Q chosen) , trial type (reward or goal, RG), and number of times the pair of stimuli had been presented previously (Num). There was no significant or widespread effect on theta power for any of our regressors (Figure 21).

Discussion

Previous studies were unable to directly compare how people learn from pseudo-rewards to rewards as the tasks used were hierarchically structured meaning that pseudo-rewards were elicited for only completion of lower level tasks while completing the overall task yielded a reward . Thus, differences in learning from pseudo-rewards and rewards in those tasks could also be attributed to learning at different levels within a hierarchy. We were interested in establishing 1) Does goal achievement generate a pseudo-reward that reinforces behavior using the same reinforcement learning system as extrinsic rewards? 2) how does learning differ from rewards and pseudo-rewards? 3) If there are behavioral differences, do they correlate with neural differences?

To answer these questions, we developed a variant of the probabilistic selection task that enables direct comparison between learning from different types of outcomes. Specifically, it allows for probing the of preference for one type of outcome on test phase on trials in which participants had to choose between boxes of the same expected value, but one is tied to reward and the other pseudo-reward. If participants attached additional value to pseudo-rewards potentially because they are internally defining its value, then we would see participants systematically choosing the pseudo-reward yielding box over the reward box. This would be interesting because it would be strong indication achieving a goal is generating a pseudo-reward to reinforce behavior. While we found that goal achievement has the ability to support learning as previous findings suggest (Ribas-Fernandes et al., 2011, Diuk et al., 2013) and learning progresses similarly from these two outcomes, overall, participants preferred rewards to achieving a goal. This could still be indicative of the generation of a pseudo-reward, but these results are consistent with other accounts. The similarity in learning may be due to participants being directed to choose a goal by the experimenter, and then seek it out. Thus, it is the experimenter who is defining that item as valuable to the participant rather than its value being internally defined by the participant. However, there are some

subjects who do exhibit a preference for pseudo-rewards over rewards, and the additional value is consistently attached to boxes leading to pseudo-rewards.

Differences do exist between learning from rewards and pseudo-rewards. There was a reduced effect of difficulty for learning phase performance which is consistent with pseudo-rewards being less valued than rewards or different learning rates for the types of outcomes. We replicate these findings that there are individual differences in the relative rates people learn from positive or negative outcomes (Frank & O'Reilly, 2004; Frank et al., 2007), and find there is an interaction with type of outcome, however, only for positive outcomes. Gains are differentially valued, while losses are not. This is reasonable because in both instances the outcome received is nothing.

For modeling, a parameter that modulates the value of pseudo-reward relative to reward is the most effective mechanism for capturing the difference seen in behavioral results. Different learning rates for pseudo-reward and reward as outcomes were unable to replicate this effect. The RLprefR and RLBayes were the models that provided the best fits based on AIC scores. The scores alone do not pick out a better model between the two, but the RLprefR model is preferred because parameters were better recovered that the RLBayes model, likely because of its lower number of parameters. Our key parameter, preference for reward, is strongly correlated in both models with the behavioral measure from the test phase. It captures some of the difference in learning from the two outcomes. Even the best models, however, are unable to recreate certain aspects of behavior like decreased discriminability in learning from pseudo-rewards and worse performance on difficult reward only pairs test phase trials than participants. These models are flexible enough to capture these aspects given the right set of parameters, but the parameters we estimate from participant's data do not produce it. Because these models are robust enough to produce the behavioral results, this suggests there is an another mechanism not accounted for by our

model that is producing these differences. There is a problem in hierarchical RL: if PPEs are conveyed by the same RL mechanisms as RPEs, how is credit assigned properly if they temporally coincide? (Botvinick, et al., 2009). Another mechanism that interacts with the RL system and produces a difference in the valuation of rewards and pseudo-rewards would eliminate this credit assignment problem. Identifying this additional mechanism, and incorporating that into new models is a future step to be taken.

We find a similar neural signature for rewards and pseudo-rewards as predicted and previously found in hierarchical reinforcement learning tasks (Ribas-Fernandes et al., 2011; Diuk et al., 2013). However, rewards and pseudo-rewards produce different strengths of theta power. There is greater theta power in fronto-central electrodes in response to pseudo-rewards as feedback when compared to rewards. This is consistent with the pseudo-rewards being relatively lesser valued. There is a wealth of evidence that theta power increases in response to errors as opposed to correct answers (Cavanagh et al., 2011, Cavanagh & Frank, 2014; Cohen, 2014), thus, theta power would be expected to increase for receiving an outcome that is relatively lower valued. These neural response differences to rewards and pseudo-rewards are only present in our analysis of theta power but not in ERP analysis. This supports Cavanagh's (2011) proposal that theta power is more sensitive to between condition differences than basic ERP components. Yet, we are unable to find strong links between behavior and neural signatures when we look at individuals. There is only a weak correlation between behavioral measures of preference for reward and the average theta power in the region sensitive to the difference between rewards and pseudo-rewards as outcomes.

While we found differences in response to feedback, we were unable to find differences in response to the stimuli tied to different types of outcomes. Hence, we were unable to replicate findings from Collins (2014) that fronto-central electrodes are sensitive to the Q-values of options. However, Q-values are extracted from a model. The model was unable to capture all important features of behavior, so perhaps

the Q-values computed on a trial-by-trial basis do not align with the "actual" Q-values that are proposed to be computed by the RL system (Bayer & Glimcher, 2005; Samejima et al., 2005;Fitzgerald et al., 2012). With an improved model that produces more accurate Q-values, it is possible that there could be an effect of the Q-value of the chosen option in these areas, further emphasizing the need to identify other possible mechanisms to inform the development of improved models.

Conclusion

Pseudo-rewards as feedback appear to support learning similar to rewards; however, they are relatively lesser-valued to reward. EEG results suggest that a similar neural mechanism underlies learning from rewards and pseudo-rewards replicating previous findings, but results from modeling suggest that there is another mechanism underlying learning from pseudo-rewards that we were unable to capture. Differences in theta power in the reward and goal condition implicate a mechanism related to the error-processing computations that take place in frontal regions. In further iterations, we hope to explore alternative mechanisms to explain the behavioral differences we see in learning from rewards and pseudo-rewards which can then inform further analysis of the neural data.

Acknowledgements

I would like to thank my advisor, Professor Anne Collins, for spending an inordinate amount of time mentoring me, ensuring I understood each and every step of the research process. Thanks to the Summer Undergraduate Research Fellowship program for funding my work over the summer. Thanks to Professor Linda Wilbrecht for being my second reader. Thanks to Christina Merrick and Assaf Breska for training me on EEG. Thanks to Sarah Master for help collecting EEG data. Finally, thanks to Lucy Whitmore and Priyam Das for bearing with me when it still took me a full hour to prep subjects for EEG. I still owe you both a drink, but hopefully mention in this section will suffice until then.

References

A. Delorme, S. Makeig (2004). EEGLAB: an open source toolbox for analysis of single- trial EEG dynamics including independent component analysis. *J. Neurosci. Methods*. **134**, 9–2.

Ballard, T., Farrell, S., & Neal, A. (in press). Quantifying the psychological value of goal achievement. *Psychonomic Bulletin & Review*.

Bandura, A. (1991). Social cognitive theory of self-regulation. *Organizational Behavior and Human Decision Processes, 50*(2), 248-287.

Barto A, Mahadevan S. (2003) Recent advances in hierarchical reinforcement learning. Disc. Event Dyn. Sys;13:341–379.

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

Botvinick, M. M., Cohen, J. D. & Carter, C. S. (2004) Conflict monitoring and anterior cingulate cortex: an update. *Trends Cogn. Sci.* 8, 539–546.

Botvinick MM, Niv Y, Barto AC. (2009) Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. Cognition.;113:262–280.

Cavanagh JF, Cohen MX, Allen, JJB (2009), Prelude to and resolution of an error: EEG phase synchrony reveals cognitive control dynamics during action monitoring. *J. Neurosci.* **29**, 98–105.

Cavanagh JF, et al. Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. Neuroimage. 2010;49:3198–209.

Cavanagh JF, Zambrano-Vazquez L, Allen JJ., (2012). Theta lingua franca: a common mid-frontal substrate for action monitoring processes. Psychophysiology.

Cavanagh J. F., Frank M. J. (2014). Frontal theta as a mechanism for cognitive control. Trends Cogn. Sci.

Cohen MX. (2014) A neural microcircuit for cognitive conflict detection and signaling. Trends in Neurosciences.;37:480–490.

Collins, A. G. E., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. European Journal of Neuroscience, 35, 1024–1035.

Collins, A. G. E., Frank M. J. (2016).Neural signature of hierarchically structured expectations predicts clustering and transfer of rule sets in reinforcement learning. Cognition. 152, 160–169

Collins, A. G. E., & Frank, M. J. (2018). Within and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. Doi.org, 184812.

Dietterich TG. (1998). The MAXQ method for hierarchical reinforcement learning; Proc. Int. Conf. on Machine Learning; pp. 118–126.

Diuk C, Tsai K, Wallis J, Botvinick M, Niv Y (2013). Hierarchical learning induces two simultaneous, but separable, prediction errors in human basal ganglia. J. Neurosci ;33:5797–5805.

FitzGerald TH, Friston KJ, Dolan RJ. Action-specific value signals in reward-related regions of the human brain. Journal of Neuroscience. 2012;32:16417–16423.

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

Frank, M. J., Woroch, B. S., & Curran, T.(2005). Error‐related negativity predicts reinforcement learning and conflict biases.*Neuron*, 47, 495–501.

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

Heath C. ,Larrick, R.P., Wu, G. (1999). Goals as reference points. Cognitive Psychology, 38, pp. 79-109

Latham, G.P. , E.A Locke, E.A. (1991) *A theory of goal setting and task performance*. Prentice-Hall, Englewood Cliffs, NJ

Lashley, K.S. (1951) The problem of serial order in behavior. In: Jeffress LA, editor. Cerebral mechanisms in behavior: The Hixon symposium; New York, NY: Wiley; pp. 112–136.

McDougall, (1908). *An Introduction to Social Psychology.* Boston: John W. Luce & Co.

Mento, A. J., Steel, R. P., & Karren, R. J. (1987). A meta-analytic study of the effects of goal setting on task performance: 1966–1984. *Organizational Behavior and Human Decision Processes, 39*(1), 52-83.

Mitchell, T. (1982). Motivation: New Directions for Theory, Research, and Practice. *The Academy of Management Review,7*(1), 80-88.

Montague, P.R., Dayan, P., Sejnowski, T.J., (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–4.

Mossholder, K. W. (1980). Effects of externally mediated goal setting on intrinsic motivation: A laboratory experiment. *Journal of Applied Psychology, 65*(2), 202-210.

Niv Y. (2009) Reinforcement learning in the brain. J. Math. Psychol; 53:139–154.

Parr R, Russell S. (1998) Reinforcement learning with hierarchies of machines. Adv. Neu. Inf. Proc. Sys.;10:1043–1049.

Ribas-Fernandes JJ, et al. (2011) A neural signature of hierarchical reinforcement learning. Neuron. 2011;71:370–379.

Samejima, K., Ueda, Y., Doya, K. & Kimura, M. (2005) Representation of action-specific reward values in the striatum. *Science* 310, 1337–1340.

Schmidt, F. L., & Hunter, J. E. (1983). Individual differences in productivity: An empirical test of estimates derived from studies of selection procedure utility. *Journal of Applied Psychology, 68*(3), 407-414.

Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997;275:1593–1599.

Singh S, Barto AG, Chentanez N. Intrinsically motivated reinforcement learning. In: Saul LK, Weiss Y, Bottou L, editors. Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference; MIT Press; Cambridge. 2005. pp. 1281–1288.

Sutton, R.S., Barto, A.G., (1998) *Reinforcement Learning* (MIT Press), vol. 9.

Tubbs, M. E. (1986). Goal setting: A meta-analytic examination of the empirical evidence. *Journal of Applied Psychology, 71*(3), 474-483.

Worthy D. A., Maddox W. T. (2014). A comparison model of reinforcement-learning and win-stay-lose-shift decision-making processes: a tribute to W.K. Estes. *J. Math. Psychol.* 59 41–49. 10.1016/j.jmp.2013.10.001

Yeung, N., Cohen, J. D. & Botvinick, M. M. (2004) The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychol. Rev.* 111, 931–959