NeuroImage 54 (2011) 1385–1394

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Choice modulates the neural dynamics of prediction error processing during rewarded learning

David A. Peterson [a], Daniel T. Lotz [b], Eric Halgren [c,d,e], Terrence J. Sejnowski [f,g], Howard Poizner [a,b,*]

[a] Institute for Neural Computation, UCSD, La Jolla, CA, USA
[b] Department of Cognitive Science, UCSD, La Jolla, CA, USA
[c] Multimodal Imaging Laboratory, UCSD, La Jolla, CA, USA
[d] Department of Neuroscience, UCSD, San Diego, CA, USA
[e] Department of Radiology, UCSD, San Diego, CA, USA
[f] Howard Hughes Medical Institute, Computational Neurobiology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA, USA
[g] Division of Biological Sciences, UCSD, La Jolla, CA, USA

ARTICLE INFO

Article history:
Received 10 March 2010
Revised 30 August 2010
Accepted 20 September 2010
Available online 25 September 2010

Keywords:
Dopamine
Reward
Event related potential
Decision making

ABSTRACT

Our ability to selectively engage with our environment enables us to guide our learning and to take advantage of its benefits. When facing multiple possible actions, our choices are a critical aspect of learning. In the case of learning from rewarding feedback, there has been substantial theoretical and empirical progress in elucidating the associated behavioral and neural processes, predominantly in terms of a reward prediction error, a measure of the discrepancy between actual versus expected reward. Nevertheless, the distinct influence of choice on prediction error processing and its neural dynamics remains relatively unexplored. In this study we used a novel paradigm to determine how choice influences prediction error processing and to examine whether there are correspondingly distinct neural dynamics. We recorded scalp electroencephalogram while healthy adults were administered a rewarded learning task in which choice trials were intermingled with control trials involving the same stimuli, motor responses, and probabilistic rewards. We used a temporal difference learning model of subjects' trial-by-trial choices to infer subjects' image valuations and corresponding prediction errors. As expected, choices were associated with lower overall prediction error magnitudes, most notably over the course of learning the stimulus–reward contingencies. Choices also induced a higher-amplitude relative positivity in the frontocentral event-related potential about 200 ms after reward signal onset that was negatively correlated with the differential effect of choice on the prediction error. Thus choice influences the neural dynamics associated with how reward signals are processed during learning. Behavioral, computational, and neurobiological models of rewarded learning should therefore accommodate a distinct influence for choice during rewarded learning.

Introduction

Many forms of learning are driven by reward. Although we can learn associations between environmental conditions and rewards (as in classical conditioning), our ability to selectively engage with our environment (as in instrumental conditioning) enables us to guide our learning and to take advantage of its benefits. As a result, in the face of multiple possible actions, our choices are a critical aspect of rewarded learning. Yet the neural dynamics of choice's influence in rewarded learning remain a mystery.

There has nevertheless been substantial progress in recent years toward elucidating the neural correlates of rewarded (and non-rewarded) feedback processing. A broad body of theoretical and empirical evidence has accumulated suggesting that trial by trial feedback-based learning is driven by phasic activity of the mesencephalic dopamine system (Abler et al., 2006). The predominant concept is that the phasic dopamine activity signals actual versus expected reward values, or a reward "prediction error" (Fiorillo et al., 2003; Montague et al., 1996; Schultz et al., 1997). Concurrently, this prediction error has gained widespread use in temporal difference models of reinforcement learning (Sutton and Barto, 1998). The mesencephalic dopamine system has been shown to modulate frontocentral feedback-related potentials in monkeys (Vezoli and Procyk, 2009) and humans (Jocham and Ullsperger, 2009). The predominant characterization of this effect measured by subtracting human scalp EEG after positive from that after negative feedback conditions is the feedback-related negativity (FRN; Miltner et al., 1997). The FRN has a frontocentrally dominant topography and is thought to arise from generators in anterior cingulate cortex (ACC;

* Corresponding author. University of California, San Diego, 9500 Gilman Drive, MC 0523, La Jolla, CA 92093, USA. Fax: +1 858 534 2014.
E-mail address: hpoizner@ucsd.edu (H. Poizner).

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.09.051

Gehring and Willoughby, 2002). Holroyd and Coles (2002) suggested that this differential activity in ACC, and the associated frontocentral negativity at the scalp, reflect the reward prediction error's influence in ACC. Subsequent experiments with fMRI in humans (Holroyd et al., 2004) and single unit recordings in monkeys (Amiez et al., 2005) have supported this role for ACC in prediction error processing.

In its simplest form, prediction error is determined not only by the actual reward signal generated by your choice, but also the reward you expected from that choice. In many everyday settings, the relationships between choices and rewards are probabilistic. In this context, and especially given the added temporal dynamics of learning, subject expectations are at best only inferred and at worst completely unpredictable. In an attempt to address this, in a probabilistic reward task Hajcak et al. (2007b) asked subjects to indicate their expectations before receiving feedback on each trial. Subjects were asked before their response in one condition and after (but still before feedback) in the other. Curiously, the FRN varied with expectancies only in the latter condition. The authors suggested that the FRN, and therefore ACC activity, were relatively more sensitive to conditions in which expectations are more closely linked to choices. Another way to measure expectation is to infer it from the experimental conditions. For example, after a period of learning, one can make reasonable assumptions about how subjects value various alternatives, and then use stimulus/reward combinations to determine crude estimates of expectations (Bellebaum and Daum, 2008). Of course, a more direct and finer-grained method is to use not only the stimulus and reward contingency structure of the experiment, but also the trial-by-trial evolution of subjects' actual choices, to infer their relative choice valuations, expected rewards and corresponding prediction error on each trial.

In this study we sought to determine whether and how choice influences the neural dynamics associated with post-feedback reward prediction error processing. To investigate this issue, we used a novel paradigm where we can explicitly investigate the differential influence of choice in rewarded learning. Specifically, we used a rewarded learning task in which subjects are uninformed about the probabilistic relationship between stimulus choices and rewards and the relative merit of various options has to be inferred indirectly through trial-and-error learning. We fit each subject's trial-by-trial choices with a temporal difference reinforcement learning model. We used the model to infer on each trial their choice valuations and, based on the feedback, the corresponding per-trial continuous valued prediction error. Using a similar paradigm in primates, Morris et al. (2006) found that phasic DA cell firing influenced choice policy. Furthermore, we have previously shown that Parkinson's disease patients off dopaminergic medications exhibit deficient performance in this task, especially after a covert reversal of reward contingencies (Peterson et al., 2009). Degeneration of mesencephalic dopamine cells is a classic neuropathology of Parkinson's disease (Dauer and Przedborski, 2003), so our previous results suggest that the feedback-based learning inherent to the task depends on the integrity of the mesencephalic dopamine system. Because of this, and the growing body of evidence for ACC involvement in dopamine-mediated learning (Amiez et al., 2006; Holroyd and Coles, 2002; Jocham and Ullsperger, 2009; Vezoli and Procyk, 2009), we expected that reward prediction errors would evoke neural responses in ACC that have been partly ascribed to the dopamine reward system. We computed reward-onset locked ERPs separately for "choice" trials on which subjects faced a two-alternative forced choice and pseudo-randomly intermingled "reference" trials on which no choice was required but all reward contingencies remained the same. Based on previous studies demonstrating a maximal effect of the FRN with a frontocentral scalp topography, we focused our ERP analysis on activity at the FCz electrode. Because the reference trials involved the same stimuli, motor response execution, and reward contingencies as their choice trial counterparts, comparing reference and choice trial types allowed us to selectively characterize the differential influence of choice in reward feedback processing.

Methods

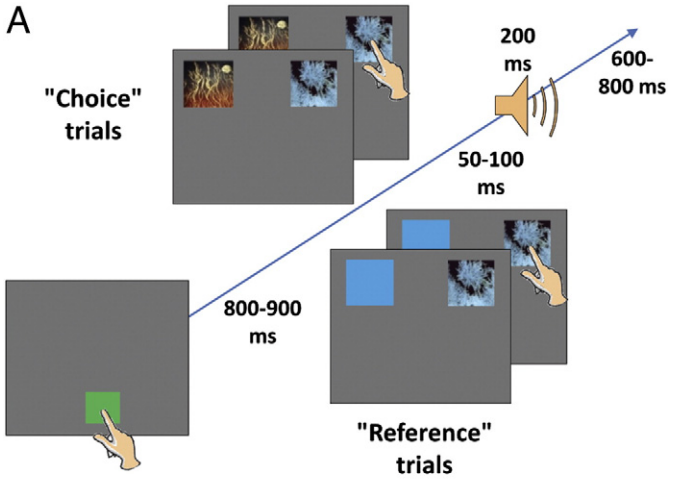
Subjects

Nineteen neurologically intact undergraduate students at the University of California San Diego (UCSD) participated. Subjects were recruited through the UCSD Department of Psychology. After detailed explanation of the procedures, all subjects provided written informed consent consistent with the Declaration of Helsinki. All subjects declared no history of neurological illness or brain surgery, normal hearing, vision correctable to at least 20/40 with corrective lenses, and no current medications for depression. All subjects were right handed according to the Edinburgh handedness inventory (Oldfield, 1971). After a description of the task, subjects were asked if they had prior experience with similar experiments. One subject explicitly asked if there would be a change part way through this task. There were technical problems with EEG acquisition software on an additional six subjects. These seven subjects were omitted from the present analysis leaving twelve subjects with a mean age of 20.3 (SD 1.2; range 19–23). Five of these subjects were female. Subjects received partial research credit toward completion of their Psychology course, in addition to their cash winnings from the task as detailed below. All procedures were approved by the UCSD Institutional Review Board.

Experimental task

We adapted a task originally used to study firing rates of dopamine cells in primate substantia nigra pars compacta (Morris et al., 2006) for use as an instrumental reward-based learning task with humans. The task is a probabilistic rewarded learning task described previously in a study of Parkinson's patients (Peterson et al., 2009). Briefly, subjects were presented with a series of trials on which they chose abstract visual images with a possibility of accruing a small reward on each trial. The images presented on each trial were selected from among four possible images, each with a fixed probability of producing an identical reward value. In order to maximize their earnings, subjects had to learn through trial-and-error which images were more likely to pay off. Half way through the experiment, the reward probabilities of the four images were covertly reversed.

Throughout the task, subjects were seated in front of a 19″ touch monitor (Elo Touchsystems, model number et1925L-7uwa-1) in sufficiently close proximity to allow comfortable reaches to both upper corners. The touch monitor was placed on a table with the top approximately 45° back from vertical. As depicted in Fig. 1A, subjects initiated each trial by pressing the green "go button" square in the lower middle of the touch monitor. After 800–900 ms, a square visual image appeared in each of the two upper corners of the touch monitor. Subjects chose an image by pressing it. Subjects were given an auditory reward feedback signal 50–100 ms after selecting an image. If they won money on that trial, they were presented with a 200 ms "high" tone (600 Hz). If they did not win money on that trial, they were presented with a "low" tone (200 Hz). The two tones were provided free field by standard PC speakers. The tones were identical in amplitude and linear ramp up/down (40 ms each). Prior to starting the experiment, subjects confirmed by verbal report that they could hear and distinguish the two tones. Approximately 600–800 ms after the reward feedback signal, the images disappeared and the go button reappeared in the lower center of the monitor, prompting the subject to begin the next trial. Subjects were required to wait until the two images appeared before releasing the go button. There were no other temporal constraints on their choice or the return to the go button. They were simply instructed to "move to touch the image as soon as you have decided which one to choose". Actual durations of each time



B

Image	j	Probability (reward) Initial Phase	Reversal Phase
	1	0.25	1.00
	2	0.50	0.75
	3	0.75	0.50
	4	1.00	0.25

Fig. 1. Task design. (A) Per-trial timeline. Time intervals in square brackets represent durations drawn randomly from a uniform distribution over the specified range. (B) Visual images, their index *j*, and their phase-contingent reward probabilities.

interval specified above were chosen randomly from a uniform distribution on each trial. Total trial duration averaged about 4 s.

The task consisted of two phases of 256 trials each. Interleaved throughout the task were two trial types: reference and choice trials comprising 62.5% and 37.5% of the trials, respectively. On the reference trials, subjects were given an "instructed" choice. They were presented with a solid blue square and one of four abstract images. They were instructed to always choose the abstract image. On the choice trials, subjects faced a two-alternative forced choice. They were presented with two of the abstract images and were told to "choose the image that is more likely to pay off". If rewarded, they received $0.07. Given the evidence that rewarded learning is particularly sensitive to the use of real versus symbolic monetary rewards (Kunig et al., 2000; Martin-Soelch et al., 2001; Smith, 1991), we gave subjects actual cash for rewards. The abstract images and the probability with which choosing them produced a reward [0.25, 0.50, 0.75 and 1.00] are shown in Fig. 1B. These reward contingencies were flipped in the otherwise identical post-reversal phase of the experiment. There were no choice trials on which the two images were identical. We fully counterbalanced the number of presentations of each image, the side on which they were presented, and the side on which rewards were available. Maximum run lengths were three choice trials, five reference trials, five trials with reward on the same side, three reference trials with the image on the same side, and five trials containing the same image on either side. Both 256-trial phases were divided into 8 blocks of 32 trials each. At the end of each block, subjects were shown their cumulative winnings and the actual monetary amount placed on the table beside them was updated accordingly, rounded up to the nearest $.25.

Subjects were first given a brief practice session, with eight reference and four choice trials. The practice stimuli were four simple geometric shapes that were different from any of the stimuli used in the actual experiment. There were no feedback signals or rewards in this practice session in order to avoid teaching any associations prior to the actual experiment. Subjects were simply familiarized with the mechanics of the trials, and particularly the explicit instruction to not choose the solid blue square on reference trials. Prior to starting the primary experiment, subjects were given an explanation of the feedback signals and rewards. They were told that some images were more likely to pay off than others, and it did not matter which side they appeared on. Finally, they were told that to maximize their winnings, they should try to figure out which images are more likely to pay off than others. The average duration of the overall session, including application, testing, and removal of the EEG cap, was approximately 2.0 h.

Reinforcement learning model

We implemented a computational reinforcement learning model to fit subjects' trial-by-trial behavior. Images $j \in \{1,2,3,4\}$ were assigned values $Q_t(j)$ at each trial t of the experiment. When image k was chosen, its value was incremented as a function of the reward $r_t \in \{0,1\}$ received upon choosing it:

$$Q_{t+1}(j) \leftarrow \begin{cases} Q_t(j) + \alpha[r_t - Q_t(j)] & \begin{cases} j = k \\ \text{o.w.} \end{cases} \\ Q_t(j) \end{cases}$$

The term $[r_t - Q_t(j)]$ was referred to as the prediction error. Note that a prediction error is calculated not only for choice but also for reference trials. The amount by which the prediction error was used to increment the image's value was weighted by the learning rate, or "gain", α. On choice trials where subjects had to choose between two images m and n, we modeled their choice probabilistically with the softmax function:

$$p_t(m) = \frac{e^{\beta Q_t(m)}}{e^{\beta Q_t(m)} + e^{\beta Q_t(n)}}$$

where the parameter β quantified the bias between exploration (low β) and exploitation (high β). We investigated the role of gain α and exploration/exploitation bias β, evaluated over the ranges [0.01 0.70] and [0 10], at uniform intervals of 0.04 and 0.5, respectively. We used a simple grid search of the parameter space to evaluate the model's fit with each subject's actual behavior. The fit at each point in the parameter space was computed as the log likelihood that the model makes the same choices a_t that the subject makes on the (two-alternative forced) choice trials:

$$\text{LLE} = \log \prod_{t \in 2\text{AFC}} p_t(a_t)$$

We used the parameter value combination that best fit each subject's choices to determine the trial-by-trial image valuations $Q_t(j)$ and the reward prediction errors for each subject.

EEG acquisition and preprocessing

Scalp EEG was measured throughout the experiment at 512 Hz using a 70-channel active electrode EEG system (Biosemi, Inc.), including 64 scalp electrodes, one electrode on each of left and right mastoid, electrodes above and below the right eye for vertical electrooculogram (VEOG), and lateral to the outer canthus of each eye for horizontal electrooculogram (HEOG). We used EEGLAB

(Delorme and Makeig, 2004) and custom Matlab routines for EEG analysis. Recordings were digitally band pass filtered offline between 1 and 100 Hz. The average reference was computed using all electrodes (excluding EOG) that did not exceed a threshold of two standard deviations above the mean for both variance and kurtosis. Rereferenced scalp and EOG electrodes were used in all subsequent analysis. All data were zero-phase band stop filtered with equiripple FIR filters in the frequency ranges of 30–34 and 58–66 Hz to attenuate extraneous artifacts from a motion capture system and line sources, respectively. For purposes of a subsequent independent components analysis (ICA) decomposition using extended Infomax (Bell and Sejnowski, 1995), all experimental session blocks were temporarily epoched into contiguous, non-overlapping 1 sec segments. Those segments with iteratively determined improbable distributions were removed. For each independent component (IC), we used DIPFIT (Oostenveld and Oostendorp, 2002) and the canonical Montreal Neurological Institute boundary element head model to generate a single equivalent current dipole. To conservatively mitigate the effects of ocular and myogenic artifacts, we rejected all ICs whose dipoles were localized outside the brain volume or accounted for the IC's scalp topography with greater than 20% residual variance. The remaining ICs (mean 9 per subject, range 6–15) were projected back to the native electrode space without further data rejection for all subsequent analysis. Trials were epoched from 200 ms before to 700 ms after reward signal onset. Based on previous studies demonstrating a maximal effect of the ERN and FRN with a frontocentral scalp topography (as discussed in the Introduction), we focused our analysis on activity at the FCz electrode. Reward-onset locked event related potentials (ERPs) were computed separately for choice and reference trials. We used native ERPs, rather than negative minus positive difference ERPs as is commonly used in FRN studies for three reasons: a) because we balanced for negative and positive feedback signals (as described in the next section), b) to avoid confusion arising from double negatives, and c) because the focus of our analysis was on the differential effect of choice on general prediction error processing.

Data analysis and statistics

Subjects' performance in each block was measured as the percentage of the 12 choice trials on which they chose the favorable image, i.e. the image more likely to pay off. Learning was evaluated using a two-factor repeated measures ANOVA, with PHASE (pre-reversal, post-reversal), and BLOCK as within subjects factors. For the model simulation of the task, we used total winnings over the entire task as a global measure of performance. The combination of model parameters with the highest total winnings was used to generate the simulation's learning curves. Based on the model fit to individual subjects' behaviors, we also calculated the mean reward prediction error magnitude in each block over all trials and separately for choice and reference trials. The mean FCz ERP amplitude was similarly calculated per block per subject for all trials and the choice and reference trials separately. In order to control for the frequency of rewards in the choice and reference trials, all comparisons involving the two conditions' ERPs used a correspondingly "balanced" subset of reference trials. Specifically, for each subject and each block, we searched forward and backward from the middle of the block to find reference trials involving the same distribution of images (and therefore reward contingencies) as those produced by their choices in that block. Because by definition this was dependent upon subjects choices, in general the balancing algorithm had to include in any given block's set of matched reference trials some reference trials outside of but temporally adjacent to that block's 32-trial set. This algorithm also served to balance the number of trials of each condition. Prediction error magnitude and FCz amplitude were evaluated using a three-factor repeated measures ANOVA, with BLOCK, PHASE (pre-reversal, post-reversal), and CONDITION (choice, reference) as within subjects factors. In ANOVAs, we used Geisser–Greenhouse corrections for non-spherical covariances. We investigated choice's dynamic influences on the prediction error and the FCz amplitude by evaluating their reference-corrected relationship using Pearson's correlation and comparing the Fisher *Z*-normalized *r* values to zero with the Student's *t*-test. Throughout the analysis, *p*-values less than 0.05 were considered significant.

Results

Rewarded learning

The experiment took an average of 32 min to complete (SD 4, range 25–38). By the end of the experiment, subjects had won an average of $24.04 (SD $0.43; range $23.38–24.85). As depicted in Fig. 2A, subjects learned to choose more favorable images, with choice performance well above the 50% chance level. This was also borne out by the two-factor ANOVA (see Table 1A), in which there was a main effect of BLOCK demonstrating that subjects' performance increased over time within each phase. By the end of the pre-reversal phase, subjects made the more favorable choice an average of 90% of the time. After the reward contingencies were reversed, choice performance dropped to below-chance levels and did not reach pre-reversal levels until the third post-reversal block. The reversal also produced a strong main effect of PHASE. By the end of the post-reversal phase, subjects chose the more favorable image 82% of the time. The significant BLOCK × PHASE interaction is a result of the immediately post-reversal drop and lower plateau in choice performance during the second phase of the task.

For the space of model parameters explored, a simulation of the task using the model produced a range of winnings from $21.93 to 24.92. Low learning rates and an emphasis on exploitation (high beta) rather than exploration (low beta) produced the highest overall "performance" (see Fig. 2B inset). Specifically, alpha = 0.17 and beta = 9.5 produced the optimal overall winnings. Using these parameter values, the simulation's "learning curve" (in Fig. 2B) shows a grossly similar morphology to that of the subjects. The simulation, however, reached higher plateaus of performance in both phases, and recovered from the reward contingency reversal after only two blocks, notably faster than the subjects. After fitting the model to the subjects' trial-by-trial choices, the corresponding reward prediction error (PE) magnitudes were evaluated and are depicted in Fig. 2C. Mean PE decreased with learning, and transiently increased in response to the reversal.

Choice and prediction error processing

To isolate the influence of choice on how the reward prediction error (PE) is processed, we examined the block-by-block temporal dynamics of PE magnitude separately for choice and reference trials. Recall that the PE is computed similarly for choice and reference trials: both conditions involve similar stimulus–reward contingencies and, as inferred from the model, image valuations. As seen in Fig. 3A and indicated by the main effects of BLOCK and PHASE in Table 1B, prediction error magnitude decreases with learning and increases in response to the covert reversal of reward contingencies. There was also a main effect of CONDITION, whereby the mean PE is lower for choice than for reference trials. This effect was most marked in the first block of both phases in Fig. 3A and noted by the BLOCK × CONDITION interaction in Table 1B.

Choice also modulated the frontocentral ERP response during each trial's post reward feedback period (see Fig. 3B). Specifically, choice trials involved a stronger positivity than reference trials over a period of 150–500 ms after the reward feedback signal onset. This effect first peaked at a latency of 190 ms. At that latency, there is a broad

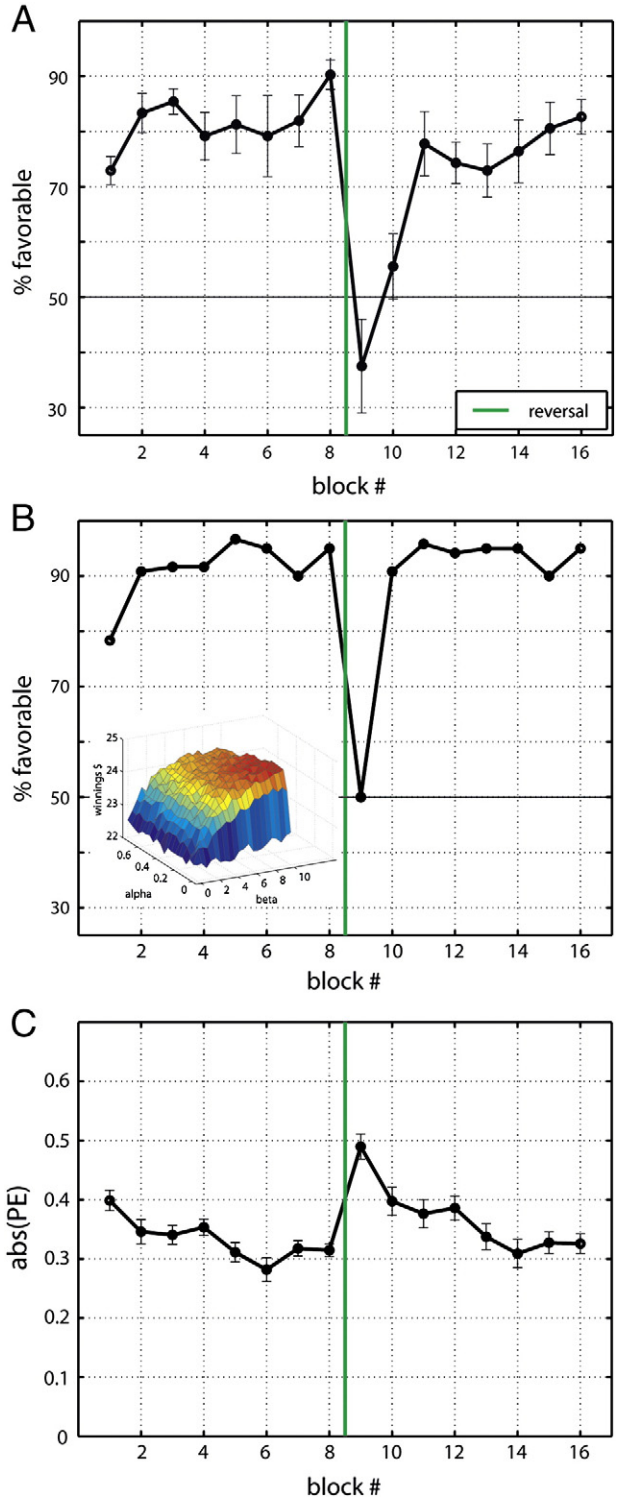


Fig. 2. Rewarded learning. (A) Learning curves, % of "favorable" choices at each block of 12 choice trials. Mean and ±standard error across subjects. Chance performance is 50%. Vertical line after block 8 denotes reward contingency reversal. (B) As in (A), but from average over 30 runs of simulations with the "optimal" model. Inset: overall performance (total winnings) as a function of model parameters. (C) Prediction error magnitude, averaged over all trials in each block, mean ± standard error across subjects.

frontocentral relative positivity for choice compared to reference ERPs, as depicted in the 2-D scalp topographies shown in Fig. 3D. The 190 ms latency ERP at FCz did not exhibit systematic dynamics with learning or reversal, as depicted in Fig. 3C and indicated by non-significant main effects of BLOCK and PHASE in Table 1C. As suggested by the overall ERP effect in Fig. 3B, there was a significant main effect of CONDITION, whereby the choice trials exhibited higher amplitude 190-ms latency FCz ERPs than the reference trials throughout most of the experiment, with the exception of blocks 5 and 9 at which time the mean choice and reference ERPs were not statistically significantly different. There was a significant BLOCK × CONDITION interaction, likely due to the differential response of the choice vs. reference trials' ERPs in response to the reversal. The trial-by-trial evolution of subjects' image valuations $Q(j)$ are depicted in Fig. 3E, averaged across subjects at each trial.

Choice's dynamic, joint influence on FCz amplitude and the prediction error is depicted in Fig. 3F. The reference-corrected measures exhibit an inverse relationship, whereby relatively higher 190-ms latency FCz ERP amplitudes are associated with lower prediction errors (mean $r = -0.24$, SD 0.28, $t = -2.96$, $p = 0.013$, Fisher's z-normalized r).

Discussion

In this study we set out to evaluate the neural dynamics of choice during rewarded learning. We used a probabilistic rewarded learning task, a temporal difference model of reinforcement learning, and event-related potentials (ERPs) to determine whether and how choice influenced reward processing. In the task, subjects were uninformed about the probabilistic relationship between stimulus choices and rewards, so the relative merit of various options had to be learned through trial-and-error. By fitting each subject's trial-by-trial choices with the model, we inferred on each trial their choice valuations and, in conjunction with the feedback they received, the corresponding per-trial prediction error. We evaluated the differential influence of choice on how the prediction error is processed by comparing the reward-locked ERPs in choice and otherwise identical reference trials. Choice evoked a stronger positivity than non-choice reward processing. This difference exhibited a frontocentral scalp distribution and peaked at about 200 ms after feedback onset. This differential brain response was also negatively correlated with the differential prediction error magnitudes.

Table 1
ANOVA summaries.

Factor(s)	*df*	*F*	*p*
(A) Rewarded learning			
BLOCK	$F(7,77)$	9.08	0.0002
PHASE	$F(1,11)$	11.88	0.0060
BLOCK × PHASE	$F(7,77)$	5.39	0.004
(B) Prediction error			
BLOCK	$F(7,77)$	32.77	<0.0001
PHASE	$F(1,11)$	49.05	<0.0001
CONDITION	$F(1,11)$	70.16	<0.0001
BLOCK × PHASE	$F(7,77)$	4.70	0.0090
BLOCK × CONDITION	$F(7,77)$	34.08	<0.0001
PHASE × CONDITION	$F(1,11)$	2.69	0.13
BLOCK × PHASE × COND	$F(7,77)$	3.74	0.0015
(C) FCz 190-ms latency ERP amplitude			
BLOCK (1)	$F(7,77)$	2.51	0.074
PHASE	$F(1,11)$	2.27	0.16
CONDITION	$F(1,11)$	10.55	0.008
BLOCK × PHASE	$F(7,77)$	0.27	0.97
BLOCK × CONDITION (2)	$F(7,77)$	4.24	0.0026
PHASE × CONDITION	$F(1,11)$	0.56	0.47
BLOCK × PHASE × COND	$F(7,77)$	1.58	0.15

(1) Original $p = 0.0225$, Geisser–Greenhouse epsilon = 0.4384.
(2) Original $p = 0.0005$, Geisser–Greenhouse epsilon = 0.7045.

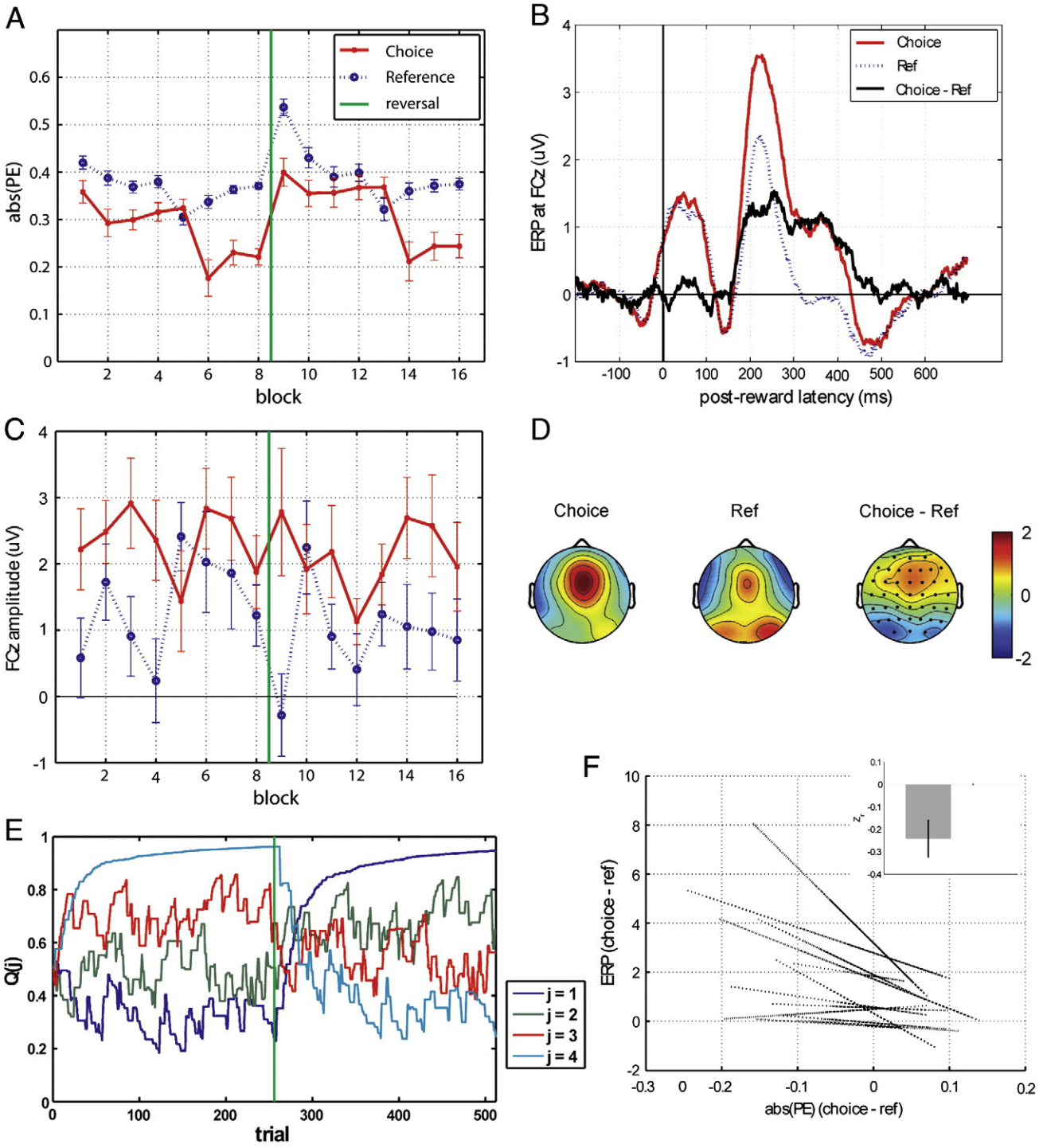


Fig. 3. Choice and prediction error. (A) Prediction error magnitude, averaged separately for choice and reference trials in each block, mean ± standard error across subjects. (B) Event related potential (ERP) responses at electrode FCz to reward signal onset (vertical line at time = 0 ms), averaged separately for choice and reference trials. Solid black line shows difference wave, choice minus reference. (C) Mean 190-ms latency ERP amplitude, averaged separately for choice and reference trials over each block. Error bars are ±standard error across subjects. (D) Scalp topography of 190-ms latency ERP, interpolated across 64 scalp channels, showing grand average amplitude separately for choice and reference trials. Difference topography includes approximate electrode locations in 2-d projection. Color bar indicates ERP amplitude in microvolts. (E) Image valuations, mean across subjects (*j* indexes images in order least to most favorable pre-reversal). (F) Relationship between mean difference FCz ERP amplitude at 190 ms latency and mean choice-reference prediction error. Linear regressions for each subject, across 16 blocks. Inset; Fisher's normalized *r*, mean ± SE across subjects.

Rewarded learning

Subjects demonstrated rapid albeit imperfect learning of the relative value of the four images in the task. Subjects chose the more favorable image an average of 72% of the time over the choice trials in the first block, already significantly higher than the 50% chance level. This was likely due to a combination of two factors: subjects could learn about stimulus–reward contingencies from the reference trials and block-wise performance was determined by collective performance over 12 choice trials in each block. By the second block, choice

performance was at 83%, within the range around which performance persisted throughout the rest of the first phase of the experiment, until it rose to 90% in the last block. The model simulation exhibited a similar profile, with performance well-above chance in the initial block and an increase to "steady state" performance by the second block. During this steady state period, however, the model achieved on average 10% better performance than the human subjects. That the model did not rise to 100% favorable choices illustrates the difficult, probabilistic nature of the task. It also suggests that the relative reward contingencies among all four stimuli might only be robustly learned if a sufficiently conservative learning rate and emphasis on exploitation was used. This would, however, severely hamper subjects' ability to adapt to the reversal in reward contingencies. Alternatively, one could bypass this trade-off and improve subjects' steady state performance by increasing information about the contingencies without increasing choice demands, as could be implemented experimentally by increasing the ratio of reference to choice trials. In other words, non-choice sampling of a noisy probabilistic environment would benefit subsequent choices in the same context.

When the reward contingencies were covertly reversed, subjects continued to make choices that were previously more rewarding. This perseveration was evident from the below-chance performance in the first post-reversal block. However, the subjects learned the new reward contingencies, with above-chance performance by the third post-reversal block and for the remainder of the experiment. Interestingly, this adaptation occurred faster in the model, suggesting that human subjects are slower at "letting go" of previously learned associations. Human subjects also demonstrated a lower mean "steady state" performance in the post-reversal than in the pre-reversal phase, consistent with previous studies using different types of reversals (Cools et al., 2002; Frank and Claus, 2006). In contrast, the model achieved a mean steady state performance level in the post-reversal phase that was approximately the same as in the pre-reversal phase. Thus, subjects' slower adaptation and lower steady state performance post-reversal may indicate a higher level of exploration and/or increased expectation of dynamics in the world. This would represent a meta level strategy not captured by simple static learning rate and exploration bias parameters in the current model.

In the present task the highest performance model coincided with a relatively low learning rate and a strong bias to exploit existing knowledge. Both of these suggest that a relatively conservative approach leads to better overall performance in this task. Conversely, high learning rates would give undue influence to each individual trial's feedback. Likewise, too much exploration would not allow one to take advantage of reward contingency knowledge built up over learning. These characteristics are natural outcomes of a task with probabilistic reward contingencies and a covert reversal in those contingencies. It is worth noting, however, that of course *some* learning is needed, because if the learning rate is too low (e.g. 0.01), reward contingencies are not learned fast enough in this task, and overall performance (as measured by total winnings) is correspondingly reduced. One might expect a similarly convex function for the exploration bias: too much leads to nearly random choices, but too little leads to inflexible choice policies that may not appropriately sample the space of stimulus–reward contingencies. This latter issue is particularly important in dynamic environments, as represented by the reward contingency reversal in the present experiment. However, even if there was no exploration in the choice trials, reference trials in the present experiment allow subjects to still sample the space of stimulus–reward contingencies. This might be why overall performance of our model simulation suggests monotonic improvement with a bias toward exploitation. This question could be partly addressed by expanding the space of exploration bias parameter to higher values, i.e. beyond 10 used in the present experiment. A more principled way to investigate this would be to systematically vary the nature of information provided in the reference trials or the ratio of reference to choice trials.

Choice and prediction errors

The TD model we used allowed us to infer from each subject's trial-by-trial choices their choice valuations and the corresponding per-trial prediction error. This is clearly less obtrusive and less subjective than explicitly asking subjects about their expectations on every trial (Hajcak et al., 2007a). It is also more fine grained than estimating prediction errors based on expectations dichotomized from early and late phases of learning (Bellebaum and Daum, 2008). Block-wise dynamics of the reward prediction errors, collapsed across both trial types, illustrated the subjects' gradual learning of the relative reward contingencies and adapting to the unannounced reward contingency reversal: the mean amplitudes of the prediction errors computed from the model generally decrease with blocks, with the exception of a transient dramatic increase after the change in reward contingencies.

When prediction error magnitudes are evaluated separately for choice and reference trials, two interesting patterns emerged. The first pattern to note is that the choice trial prediction error magnitudes are generally lower. This is expected from the simple fact that, over the course of learning the relative reward contingencies, choice trials allow one to more frequently choose the image more likely to pay off. On average this leads to lower prediction error magnitudes. It may be that your expectations are more in line with actual feedback when you have volitional control over the actions on which the expectations are based. This could be investigated in future studies by dynamically matching the frequency of image presentations in the reference trials to the images selected by, for example, the previous subject. The second pattern of interest is how the mean prediction error magnitudes for the choice trials become substantially lower than the non-choice trials in the last three blocks of each phase. This roughly corresponded to the periods during which stimulus valuations (Q-values) were "well separated" not only between the most favorable and second most favorable stimuli, but also between the second and third most favorable stimuli. Given the rational bias toward choosing more favorable images, this would lead to markedly lower prediction error magnitudes on choice trials. Earlier periods of commensurate "separation" between the third and fourth best stimuli (but not the second and third) would not produce commensurately lower prediction error magnitudes, because the mean valuations for those stimuli are lower (closer to 0.5). In summary, choice trials correspond to lower overall prediction error magnitudes and a greater prediction error magnitude reduction with learning. Prediction error magnitudes for non-choice reference trials did not decrease as dramatically as for choice trials toward the end of each phase simply because the lack of volitional control precludes subjects from "choosing" an image whose valuation is higher, more likely to produce a reward, and more likely to meet their expectations.

The prediction error magnitudes for both choice and reference trials follow temporal dynamics that are very similar between the pre- and post-reversal phases. We expect that this was due to our use of identical trial sequences in the two phases (notwithstanding the reward contingency reversal, of course). We intentionally implemented identical trial sequences in the two phases to eliminate that as a causal factor for comparing the two phases. One could, of course, use different trial sequences between subjects, while retaining the same counterbalancing and run-length limitations. Although this would likely wash out trial sequence-specific effects in measures derived from across-subject averages, it could introduce additional subtle variables that would have to be carefully controlled for in future experiment designs. Likewise, one could hypothesize that prediction errors could be differentially weighted in choice versus reference trials. This could be tested by fitting subjects' behavior with separate learning rates for the two trial types. Although we deemed there to be

insufficient data to justify use of an additional model parameter in the present study, this would nevertheless represent an interesting line of inquiry for future studies.

Neural dynamics associated with choice

Importantly, choice not only modulated the prediction error magnitude associated with rewarded learning, but also differentially influenced the associated neural dynamics. Specifically, choice evoked greater frontocentral positivity in the scalp ERP than non-choice reward processing, peaking about 190 ms after feedback onset. It is unlikely that this *post*-feedback ERP difference was an artifactual result of baseline-subtracting different *pre*-choice activity, because the ERP waveforms associated with the choice and reference conditions start and finish the post-feedback period at approximately the same amplitudes. Furthermore, the ERP difference is not due to a different proportion of rewarding versus non-rewarding trials in the choice compared to the reference trials, because for the ERP analyses we used for each subject a "balanced" subset of reference trials involving the exact same distribution of images (and therefore reward contingencies) as those produced by their choices in that block.

The temporal morphology of our ERP waveforms was strikingly similar for choice and non-choice post-reward processing, with the primary distinction being one of amplitude over the approximately 170–400 ms window after reward onset. It may be that reward prediction error processing is recruiting similar neural substrates in both cases, but choice places greater "demands" on the same substrate. However, we cannot exclude the possibility that choice recruits a different neural substrate for reward processing. This might be suggested by early evidence that choice is associated with post-choice changes in single unit firing in human hippocampal gyrus (Halgren et al., 1978) and more recent evidence of dorsolateral prefrontal cortical influence on the choice valuation putatively instantiated in ACC (Hare et al., 2009). The topographic distribution of the differential ERP associated with choice that we found is similar to the frontocentrally dominant topography found in previous studies of the feedback-related negativity. Furthermore, although the purpose of the present study was not to compare ERP responses to rewarding versus non-rewarding feedback, the higher proportion of rewarding feedback in our experiment is consistent with the higher amplitude positive ERPs Yeung et al. (2005) found in "gain" versus "loss" trials. In both cases, our results are consistent with previous research suggesting a role for ACC in prediction error processing, as shown with fMRI in humans (Holroyd et al., 2004) and single unit recordings in monkeys (Amiez et al., 2006). Although our results suggest that choice influences prediction error processing, the results do not preclude the likelihood that the ACC is involved in an array of more general "pertinence monitoring" (Fujiwara et al., 2009; Procyk and Josephy, 2001; Quilodran et al., 2008) or "salience" signaling functions, as with one notable interpretation of the phasic mesencephalic DA signal (Redgrave and Gurney, 2006). It may be that post-choice processing is inherently treated as more salient as a simple consequence of having been actively engaged by the choice, compared to similar post-reward processing in the absence of a preceding choice. Another not mutually exclusive possibility is that the ERPs reflect an element of surprise correlated with the magnitude of the prediction error and also associated with the occurrence of low probability events (Mars et al., 2008).

The choice and reference conditions also differed in their posterior ERP amplitudes at the 190 ms latency, with choice inducing *lower* amplitude than the non-choice condition. Because the stimuli remained viewable for 600–800 ms after reward signal onset, it is tempting to speculate that this difference reflects distinct post-reward visual processing of the stimuli depending on whether or not the subject just made a choice. This possibility could be tested experimentally in future studies with for example eye tracking. Alternatively, it may be that in the case of the reference trials, subjects did only the pre-response visual processing of the stimuli necessary to detect which one to choose. Because the images remained visible for 600–800 ms after response, the subjects could have delayed image value retrieval on the reference trials, possibly spilling over into the post-feedback period. This would induce a confound when interpreting the difference in choice- versus reference-ERPs post feedback. However, if this were the case, the subsequent processing, putatively involving comparison with the feedback signal to form a prediction error used to update image valuations, should be temporally delayed in the reference compared to the choice trials. Yet the choice- and reference-ERPs as depicted in Fig. 3B are precisely time locked until about 150 ms after the feedback signal onset, making this interpretation unlikely.

Previous studies have indirectly sought to investigate how prediction error processing is influenced by choice. For example, in an experiment without reward or learning per se, Gentsch et al. (2009) dichotomized the sources of feedback into "internal" and "external" by modifying the Eriksen flanker task to include two types of error conditions. The "internal" errors arose when subjects realized they made an incorrect response. A novel "external" error condition was produced by omitting the confirmatory feedback usually provided on correct trials. Subjects were informed prior to the experiment that this would occasionally occur due to malfunctions in the response equipment. The internal errors were associated with a much earlier latency response negativity (~60 ms), whereas the external errors evoked a negativity consistent with the typical FRN (~200–300 ms). Thus the internal and external "sources" of errors evoke differential processing of the feedback signal. Internal and external sources of actions and feedback are homologous to conditions in which someone may or may not get to choose between available options. Thus one would predict that conditions with and without choice would likewise evoke differential processing of rewarding feedback. This is also warranted by recent demonstrations of the effect of "personal control" in gambling games (Clark, 2010). In the present study, however, the choice- and non-choice ERP waveforms did not substantially diverge until almost 200 ms after the feedback signal onset, and as such might both be considered physiological metrics of "external" origin.

Yeung et al. (2005) demonstrated that choice increases the amplitude of the FRN. However, the choice versus non-choice comparison was done with two separate experiments, neither of which included a contingent relationship between images and rewards, and both of which included the confounding task demands associated with updating and maintaining a cumulative sum of winnings. By not including other task demands temporally coincident with the immediate post-reward period and interleaving choice and non-choice trials within a single experiment, we were able to more directly investigate the differential influence of choice on reward processing.

We used the same rewarded learning paradigm with which we have previously shown that learning performance depends on the integrity of the human dopamine system (Peterson et al., 2009). A growing body of convergent evidence (Amiez et al., 2005, 2006; Holroyd and Coles, 2002; 2008; Jocham and Ullsperger, 2009; Vezoli and Procyk, 2009) suggests that ascending projections from mesencephalic dopaminergic nuclei modulate areas such as ACC. If the frontocentrally focal ERP effects we observed are modulated by ACC activity, then our results showing a dependency on choice are consistent with Morris et al.'s (2006) finding using the same experimental paradigm in primates that phasic DA cell firing influenced choice policy. Moreover, as reviewed in the Introduction, phasic dopamine activity is thought to encode the reward prediction error posited in our model (Schultz, 1997). Collectively, our evidence lends further support to Holroyd and Coles (2002) conceptual model linking dopamine, reinforcement learning, and feedback-processing

assayed with scalp potentials. Our results also suggest a possible extension to Holroyd's framework to accommodate differential reward processing for learning that does or does not involve choice.

It is worth noting that important differences between the present study and that of Morris et al. (2006). Specifically, Morris et al. were interested in whether dopamine cell responses predicted future actions. Their monkeys were highly overtrained, having undergone extensive behavioral training on the task before the physiological recordings. Thus, Morris et al. were studying the monkeys after learning had already taken place—they were not studying brain processes that occurred during learning. In contrast, we focused in the present study on brain events associated with post-feedback event processing during learning.

We also posed the question as to whether choice's differential influence on post-reward neural dynamics bore a relationship to its differential influence on prediction error magnitude. We found that the differential brain response was negatively correlated with the differential prediction error magnitudes. Although perhaps related to the steep drops in choice prediction error midway through each phase, the two classes of events do not perfectly align in a block-wise fashion: the choice prediction errors drop in the sixth block in each phase, whereas the choice ERPs drop transiently in blocks four and five. Thus, the ERP dynamics do not appear to be a result of using the identical trial sequence in each phase. Furthermore, they temporally precede the drops in prediction error. One possibility is that relative stimulus valuations are crystallizing but not fully exploited until some number of trials later, at which time exploitative choices begin to produce decreased overall prediction error magnitudes. Relatedly, perhaps the transient decreases in choice ERP amplitudes are associated with transient changes in relative certainty, in which subjects make a transition from an uncertain to a relatively more certain state. Uncertainty has been postulated to have a significant influence in learning, choice, and their neural correlates (Behrens et al., 2007; Doya, 2008). Nevertheless, these notions are strictly speculative and future experiments explicitly designed to test them are needed. Importantly, because the differential ERP is relatively high both early and late in each phase, whereas the differential prediction error magnitude is much higher in late than in early parts of each phase, the differential ERP results appear to be due to the differential contribution of choice, and not differences in reward prediction error magnitude induced by choice.

The time window of interest in the present experiment was the post-feedback period when the putative prediction error is being used to update image valuations. The image valuations are, in turn, used to help inform choices on subsequent trials. Thus the post-choice processing is likely an important part of the intersection between rewarded learning and choice's more broadly defined domain of decision making.

Conclusions

In the face of multiple possible actions, choice plays a critical role in rewarded learning. We found that the reward prediction error and its neural correlates measured with scalp ERPs were differentially modulated by choice. Rewarded learning that required choice involved different dynamics, both in terms of the learning process and the neural activity involved in processing the reward signal. These effects were specific to choice, because the intermingled choice and non-choice learning conditions were otherwise identical in terms of stimuli, motor responses, and stimulus–reward contingencies. The results support the notion that behavioral, computational, and neurobiological accounts of reinforcement learning should carefully consider the differential influence of choice on reward processing during learning. Choice allows us to actively sense relationships between stimuli, actions, and rewards. This ability to selectively engage with our environment in the form of choice enables us to guide our learning and to take advantage of its benefits.

Acknowledgments

We thank Genela Morris and Hagai Bergman for helpful discussions on the paradigm, Julie Onton and Klaus Gramann for helpful discussions about the EEG analysis, Andrey Vankov for assistance with the custom acquisition software and Alice Ahn for help with data collection. This work was supported by the National Science Foundation grant SBE-0542013 to the Temporal Dynamics of Learning Center, an NSF Science of Learning Center, Office of Naval Research MURI grant N00014-10-1-0072, and the National Institutes of Health grant 2 R01 NS036449-11.

References

Abler, B., Walter, H., Erk, S., Kammerer, H., Spitzer, M., 2006. Prediction error as a linear function of reward probability is coded in human nucleus accumbens. Neuroimage 31, 790–795.
Amiez, C., Joseph, J.P., Procyk, E., 2005. Anterior cingulate error-related activity is modulated by predicted reward. Eur. J. Neurosci. 21, 3447–3452.
Amiez, C., Joseph, J.P., Procyk, E., 2006. Reward encoding in the monkey anterior cingulate cortex. Cereb. Cortex 16, 1040–1055.
Behrens, T.E.J., Woolrich, M.W., Walton, M.E., Rushworth, M.F.S., 2007. Learning the value of information in an uncertain world. Nat. Neurosci. 10, 1214–1221.
Bell, A.J., Sejnowski, T.J., 1995. An information-maximization approach to blind separation and blind deconvolution. Neural Comput. 7, 1129–1159.
Bellebaum, C., Daum, I., 2008. Learning-related changes in reward expectancy are reflected in the feedback-related negativity. Eur. J. Neurosci. 27, 1823–1835.
Clark, L., 2010. Decision-making during gambling: an integration of cognitive and psychobiological approaches. Philos. Trans. R. Soc. B-Biol. Sci. 365, 319–330.
Cools, R., Clark, L., Owen, A.M., Robbins, T.W., 2002. Defining the neural mechanisms of probabilistic reversal learning using event-related functional magnetic resonance imaging. J. Neurosci. 22, 4563–4567.
Dauer, W., Przedborski, S., 2003. Parkinson's disease: mechanisms and models. Neuron 39, 889–909.
Delorme, A., Makeig, S., 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J. Neurosci. Meth. 134, 9–21.
Doya, K., 2008. Modulators of decision making. Nat. Neurosci. 11, 410–416.
Fiorillo, C.D., Tobler, P.N., Schultz, W., 2003. Discrete coding of reward probability and uncertainty by dopamine neurons. Science 299, 1898–1902.
Frank, M.J., Claus, E.D., 2006. Anatomy of a decision: striato-orbitofrontal interactions in reinforcement learning, decision making, and reversal. Psychol. Rev. 113, 300–326.
Fujiwara, J., Tobler, P.N., Taira, M., Iijima, T., Tsutsui, K.I., 2009. Segregated and Integrated Coding of Reward and Punishment in the Cingulate Cortex. J. Neurophysiol. 101, 3284–3293.
Gehring, W.J., Willoughby, A.R., 2002. The medial frontal cortex and the rapid processing of monetary gains and losses. Science 295, 2279–2282.
Gentsch, A., Ullsperger, P., Ullsperger, M., 2009. Dissociable medial frontal negativities from a common monitoring system for self- and externally caused failure of goal achievement. Neuroimage 47, 2023–2030.
Hajcak, G., Dunning, J.P., Foti, D., 2007a. Dismantling reappraisal: emotion, cognition, and the late positive potential. Psychophysiology 44, S10–S11.
Hajcak, G., Moser, J.S., Holroyd, C.B., Simons, R.F., 2007b. It's worse than you thought: the feedback negativity and violations of reward prediction in gambling tasks. Psychophysiology 44, 905–912.
Halgren, E., Babb, T.L., Crandall, P.H., 1978. Activity of human hippocampal formation and amygdala neurons during memory testing. Electroencephalogr. Clin. Neurophysiol. 45, 585–601.
Hare, T.A., Camerer, C.F., Rangel, A., 2009. Self-control in decision-making involves modulation of the vmPFC valuation system. Science 324, 646–648.
Holroyd, C.B., Coles, M.G., 2002. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol. Rev. 109, 679–709.
Holroyd, C.B., Coles, M.G., 2008. Dorsal anterior cingulate cortex integrates reinforcement history to guide voluntary behavior. Cortex 44, 548–559.
Holroyd, C.B., Nieuwenhuis, S., Yeung, N., Nystrom, L., Mars, R.B., Coles, M.G.H., Cohen, J.D., 2004. Dorsal anterior cingulate cortex shows fMRI response to internal and external error signals. Nat. Neurosci. 7, 497–498.
Jocham, G., Ullsperger, M., 2009. Neuropharmacology of performance monitoring. Neurosci. Biobehav. Rev. 33, 48–60.
Kunig, G., Leenders, K.L., Martin-Solch, C., Missimer, J., Magyar, S., Schultz, W., 2000. Reduced reward processing in the brains of Parkinsonian patients. NeuroReport 11, 3681–3687.
Mars, R.B., Debener, S., Gladwin, T.E., Harrison, L.M., Haggard, P., Rothwell, J.C., Bestmann, S., 2008. Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. J. Neurosci. 28, 12539–12545.

Martin-Soelch, C., Leenders, K.L., Chevalley, A.F., Missimer, J., Kunig, G., Magyar, S., Mino, A., Schultz, W., 2001. Reward mechanisms in the brain and their role in dependence: evidence from neurophysiological and neuroimaging studies. Brain Res. Brain Res. Rev. 36, 139–149.

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. J. Cogn. Neurosci. 9, 788–798.

Montague, P.R., Dayan, P., Sejnowski, T.J., 1996. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. 16, 1936–1947.

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., Bergman, H., 2006. Midbrain dopamine neurons encode decisions for future action. Nat. Neurosci. 9, 1057–1063.

Oldfield, R.C., 1971. Assessment and analysis of handedness - Edinburgh inventory. Neuropsychologia 9, 97–113.

Oostenveld, R., Oostendorp, T.F., 2002. Validating the boundary element method for forward and inverse EEG computations in the presence of a hole in the skull. Hum. Brain Mapp. 17, 179–192.

Peterson, D.A., Elliott, C., Song, D.D., Makeig, S., Sejnowski, T.J., Poizner, H., 2009. Probabilistic reversal learning is impaired in Parkinson's disease. Neuroscience 163, 1092–1101.

Procyk, E., Josephy, J.P., 2001. Characterization of serial order encoding in the monkey anterior cingulate sulcus. Eur. J. Neurosci. 14, 1041–1046.

Quilodran, R., Rothe, M., Procyk, E., 2008. Behavioral shifts and action valuation in the anterior cingulate cortex. Neuron 57, 314–325.

Redgrave, P., Gurney, K., 2006. The short-latency dopamine signal: a role in discovering novel actions? Nat. Rev. Neurosci. 7, 967–975.

Schultz, W., 1997. Dopamine neurons and their role in reward mechanisms. Curr. Opin. Neurobiol. 7, 191–197.

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

Smith, V.L., 1991. Papers in experimental economics. Cambridge University Press, Cambridge England ; New York.

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction. MIT Press, Cambridge, Massachusetts.

Vezoli, J., Procyk, E., 2009. Frontal feedback-related potentials in nonhuman primates: modulation during learning and under haloperidol. J. Neurosci. 29, 15675–15683.

Yeung, N., Holroyd, C.B., Cohen, J.D., 2005. ERP correlates of feedback and reward processing in the presence and absence of response choice. Cereb. Cortex 15, 535–544.