NeuroImage xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

NeuroImage

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

Temporal dynamics of prediction error processing during reward-based decision making

Marios G. Philiastides [a,b,*,1], Guido Biele [a,b,c,*,1], Niki Vavatzanidis [a], Philipp Kazzer [a], Hauke R. Heekeren [a,b,c]

[a] Max Planck Institute for Human Development, Berlin, 14195, Germany
[b] Max Planck Institute for Human Cognitive & Brain Sciences, Leipzig, 04303, Germany
[c] Department of Education & Psychology, Freie Universität, Berlin, 14195, Germany

ARTICLE INFO

Article history:
Received 26 February 2010
Revised 6 May 2010
Accepted 19 May 2010
Available online xxxx

Keywords:
Decision making
Reward
Reinforcement learning
Prediction error
Single-trial
Model
EEG

ABSTRACT

Adaptive decision making depends on the accurate representation of rewards associated with potential choices. These representations can be acquired with reinforcement learning (RL) mechanisms, which use the prediction error (PE, the difference between expected and received rewards) as a learning signal to update reward expectations. While EEG experiments have highlighted the role of feedback-related potentials during performance monitoring, important questions about the temporal sequence of feedback processing and the specific function of feedback-related potentials during reward-based decision making remain. Here, we hypothesized that feedback processing starts with a qualitative evaluation of outcome-valence, which is subsequently complemented by a quantitative representation of PE magnitude. Results of a model-based single-trial analysis of EEG data collected during a reversal learning task showed that around 220 ms after feedback outcomes are initially evaluated categorically with respect to their valence (positive vs. negative). Around 300 ms, and parallel to the maintained valence-evaluation, the brain also represents quantitative information about PE magnitude, thus providing the complete information needed to update reward expectations and to guide adaptive decision making. Importantly, our single-trial EEG analysis based on PEs from an RL model showed that the feedback-related potentials do not merely reflect error awareness, but rather quantitative information crucial for learning reward contingencies.

Introduction

Adaptive decision making depends on the accurate representation of the rewards associated with potential choices. These representations can be acquired with reinforcement learning mechanisms, which use the prediction error (PE), the difference between expected and actual rewards, as a learning signal to update expectations (Rescorla and Wagner, 1972; Sutton and Barto, 1998). Hence, the understanding of feedback processing, in particular processing of PEs, is a fundamental aspect of any explanation of instrumental learning. Accordingly, a large number of fMRI studies investigated the representation of PEs during instrumental learning, resulting in a precise spatial localization of PE encoding (Dayan and Niv, 2008; Niv and Schoenbaum, 2008). In contrast, the speed and temporal evolution of PE processing are less well understood as within- and across-trial temporal characterization of PEs received less attention.

Animal studies showed that dopaminergic midbrain neurons represent the PE during associative learning (Schultz, 2006). For instrumental learning, neurophysiological experiments showed that in addition to the dorsal striatum (Apicella et al., 1991), PEs are also represented in cortical regions, such as the anterior cingulate cortex (ACC) (Matsumoto et al., 2007), the orbitofrontal cortex (OFC) (Tremblay and Schultz, 2000), and the dorsal premotor cortex (PMd) (Buch et al., 2006). These results are generally supported by human functional magnetic resonance imaging (fMRI) studies, which showed that PEs are represented in projection sites of dopaminergic midbrain regions such as the ventral striatum (O'Doherty et al., 2003; Krugel et al., 2009) but also in other brain regions, including the dorsal striatum, prefrontal regions and the ACC (Nieuwenhuis et al., 2005; Balleine et al., 2007; Behrens et al., 2007; Dayan and Niv, 2008; Niv and Schoenbaum, 2008; Krugel et al., 2009). While these findings provide clear evidence that PE processing takes place in a highly distributed network, the temporal dynamics of this network and how these relate to different aspects of PE encoding remains an open question.

Human electroencephalography (EEG) experiments that take advantage of the high temporal resolution of the acquired signals have the potential to identify these temporal dynamics. Previous EEG studies have mostly relied on event-related potential (ERP) analysis to

* Corresponding authors. Philiastides is to be contacted at Max Planck Institute for Human Development, Berlin, 14195, Germany. Biele, Department of Education & Psychology, Freie Universität, Berlin, 14195, Germany.
E-mail addresses: marios.philiastides@gmail.com (M.G. Philiastides), guido.biele@gmail.com (G. Biele).
[1] These authors contributed equally to this work.

1053-8119/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2010.05.052

provide a characterization of PE processing (Horst et al., 1980; Johnston and Holcomb, 1980; Holroyd and Coles, 2002; Frank et al., 2005; Yeung et al., 2005). Specifically, they have identified an early categorical valence evaluation of the outcome (i.e., positive vs. negative PE) (Gehring and Willoughby, 2002; Hajcak et al., 2005; Yeung and Sanfey, 2004; Yeung et al., 2005; Hajcak et al., 2006) and a later representation of outcome expectancy (i.e., low vs. high PE magnitude) (Horst et al., 1980; Johnston and Holcomb, 1980; Yeung and Sanfey, 2004; Yeung et al., 2005; Hajcak et al., 2005; Eppinger et al., 2008). The ERP components that are thought to represent these two processing stages are the feedback-related negativity (FRN) and the feedback-related positivity (FRP), respectively. Note that here we use the term FRP to refer to both the classic P300 expectancy effects (e.g., Yeung and Sanfey, 2004) and those reported on ERP difference waves in the P300 time window (e.g., Hajcak et al., 2005).

Despite the growing number of EEG studies trying to identify the temporal characteristics of PE processing the results remain inconclusive. For instance, some authors have argued that unexpected losses are reflected in a lower amplitude FRN (Holroyd and Coles, 2002; Holroyd et al., 2009) while others have reported that the FRN is larger for unexpected negative outcomes (Bellebaum and Daum, 2008). With regard to the FRP, some authors reported that it is modulated only by PE magnitude (Yeung and Sanfey, 2004; Hajcak et al., 2005), whereas others reported that it is modulated by outcome valence (Johnson and Donchin, 1978; Mathewson et al., 2008) or have suggested that it might implement positive PEs (Eppinger et al., 2008).

The highly distributed nature of the PE network as identified with fMRI and the presence of an early and a late feedback related potential in most EEG studies suggests that PE processing proceeds in multiple stages, which evolve gradually over time with potentially overlapping representations of valence and magnitude. We therefore hypothesized that PE processing might evolve in two stages, whereby an early quick evaluation of outcome valence is followed by a later and more deliberative process representing both PE valence and magnitude. This hypothesis goes beyond previous reports, in postulating that while PE valence is represented only in the earlier stage, PE valence and magnitude are concurrently represented in the later stage in order for the brain to simultaneously have access to the complete information needed to update reward expectations.

To test this hypothesis a model-based single-trial analysis of the acquired data is required. To date, only few studies used reinforcement learning models to estimate PEs, and when so they either did not apply the model on individual participants or did not examine a trial-by-trial correspondence of PEs and EEG activity (Holroyd and Coles, 2002, 2008; Cohen et al., 2007). That is, all of the aforementioned EEG studies used conventional trial-average ERP analysis, which inevitably concealed inter-trial and intersubject response variability. This in turn precluded a complete temporal characterization of PE processing through a quantitative trial-to-trial association between PEs and the underlying neuronal activity. Only such trial-by-trial association can provide unequivocal support that EEG signals carry information instrumental for a trial-by-trial learning of reward contingencies. In contrast, when comparing feedback related potentials before, during and after learning (Yasuda et al., 2004; Hajcak et al., 2006; Eppinger et al., 2008) it is unclear if the observed differences relate to the representation of quantitative PEs or whether they are effects of global learning and time dependent processes such as uncertainty reduction and task engagement. Indeed applying different measures of uncertainty, like the variance of a beta distribution (Behrens et al., 2007) or the average of the most recent PE magnitudes (Krugel et al., 2009), to reinforcement learning situations shows that experience and winning probability are correlated with uncertainty, but PEs from a reinforcement learning model are not (see Fig. S1 for details).

To test our hypothesis and provide a comprehensive account of the temporal dynamics of PE processing we used a reversal learning task, designed to elicit a broad range of PEs during reinforcement learning, in combination with a model-based, single-trial EEG analysis. Specifically, we used a machine learning approach to identify linear spatial weightings of the EEG sensors for specific temporal windows, which optimally discriminated between trials conditioned along different PE valence and magnitude dimensions (Parra et al., 2002, 2005; Philiastides et al., 2006; Philiastides and Sajda, 2006, 2007; Ratcliff et al., 2009). We then used the single-trial information extracted from these discriminating components to relate trial-to-trial variability in neuronal activity with trial-to-trial variability in PEs estimated with a reinforcement learning model.

Methods

Participants

Eighteen young adults participated in the study (10 females, mean age: 26 years, age range: 22–34 years). They were all right-handed, reported no history of neurological diseases, and had normal or corrected to normal vision. Informed consent was obtained from all participants according to procedures approved by the local Ethics Committee of the Max Planck Institute for Human Development.

Stimuli

We used a set of fifteen abstract symbols that were adapted from the Rey Visual Design Learning Test (Rey, 1964) and were provided by the Cognitive Neuroscience Lab at the University of Michigan. In addition to these symbols we used cartoon-like illustrations of a happy and a sad face (henceforth, a smiley and a frowney) to provide positive and negative feedback, respectively. The feedback stimuli were equally face-like and the only difference was the orientation of the line drawing of the mouth (curve up vs. curve down), similar to previous studies (Frank et al., 2005; Yeung et al., 2005). All images were equated for size, luminance and contrast. A Dell (Round Rock, TX) Precision 360 Workstation with nVidia (Santa Clara, CA) Quadro FX500/FX600 graphics card and Presentation software (Neurobehavioral Systems Inc., Albany, CA) controlled the stimulus display. Images were presented on a Dell 2001FP TFT monitor (resolution: 1024×768 pixels, refresh rate: 60 Hz) at a distance of 1 m from the subject. Each image subtended $4^\circ \times 4^\circ$ of visual angle.

Reversal learning task

The experiment consisted of three blocks of 350 trials each (1050 trials total). Blocks were separated by self-terminated breaks. At the beginning of each block subjects were shown a screen with three symbols. For each block a different triple of stimuli was chosen randomly from the set of fifteen symbols. The subjects initiated each block of trials with a button press after they had familiarized themselves with the symbols. Prior to the main experiment, all subjects completed a practice block, which ended either after they had reached the learning criterion (see below) twice or after they had completed 100 trials.

Subjects were told that their goal was to identify the symbol with the highest reward probability. They were also informed that in the course of each block, the highest reward probability might shift to one of the other two symbols and that they would have to adjust their choices accordingly. Each rewarded trial earned them 1 point, while unrewarded trials earned them zero points. Subjects were also told that they would receive a fixed payment for participation (15 €) and an additional amount based on the total amount of points earned. No further details regarding the mapping between earned points and the final payout were given to the subjects.

Each trial began with the presentation of a fixation cross for 500 ms. Two of the three symbols were then placed to the left and to

Please cite this article as: Philiastides, M.G., et al., Temporal dynamics of prediction error processing during reward-based decision making, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.05.052

the right of the fixation cross for a maximum of 1000 ms. During this time, subjects had to choose one of the symbols by pressing the left or right button on a response device using their right index or middle finger, respectively. As soon as subjects reported their choice the two symbols and fixation cross disappeared and the selected symbol appeared in the centre of the screen for 500 ms. Feedback was then provided by placing a smiley or a frowney below the selected symbol for another 500 ms. Trials, in which subjects failed to respond within the 1000 ms presentation time, were followed by a "Too Slow" message and were excluded from further analysis. Fig. 1 summarizes the sequence of these events.

At any point in time one of the three symbols was associated with a "high" reward probability r (i.e., good symbol) compared to the remaining two symbols (i.e., bad symbols), each of which had a reward probability of $1-r$. Through trial and error subjects had to learn to choose the good symbol. Note that this design implies a strong negative correlation between payoffs for the good option and payoffs for the bad options. We introduced a total of four "high" reward probabilities $r=[0.9, 0.85, 0.75, 0.65]$. The choice of r determined the difficulty level of each block. For each participant, the practice block and one of the three main blocks were always associated with $r=0.9$, while for the other two main blocks r was chosen randomly among the remaining three options. The order of difficulty levels during the experiment was randomized. Participants were naïve about the exact reward probabilities or the corresponding block-wise changes.

To detect when subjects learned to chose the symbols with the higher reward probability we defined a learning criterion. Specifically, subjects were thought to have learned the task when they chose the good symbol in 7 out of the last 8 trials. This criterion was slightly relaxed–7 out of 9 trials–for the most difficult reward level (i.e., $r=0.65$). Every time the learning criterion was reached, a reversal was introduced, that is the "high" reward probability was re-assigned to a different symbol. To make reversals less predictable, we included additional trials with the learned reward contingency (ranging randomly from 5 to 15) after the learning criterion was reached and before participants entered a new learning phase.

To prevent subjects from searching for non-existent patterns and to reduce cognitive load we presented the three possible pair combinations of the three symbols in a fixed order—though the presentation side of the symbols (left or right) was randomized. Subjects were explicitly informed about this manipulation. Another key component of this paradigm was that we presented stimulus pairs chosen from a pool of three symbols. This manipulation served two important purposes. First, it encouraged subjects to engage in an exploration phase to identify the most rewarding symbol after reversals occurred. Second, it forced the subjects to choose between the two least rewarding symbols (in every third trial, when the two were presented together), even when they had learned the task. This ensured that on some trials subjects could be positively surprised by a reward (i.e., when one of the least rewarding symbols was actually rewarded). This in turn generated a greater number of positive PEs.

EEG data acquisition

We acquired EEG data in a dark, electrically and acoustically shielded cabin (Industrial Acoustics Company, Niederkrüchten, Germany) using BrainAmp DC amplifiers (Brain Products, Gilching, Germany) from 74 Ag/AgCl scalp electrodes placed according to the 10-10 system (EasyCap, Herrsching-Breitbrunn, Germany). In addition, we recorded data from three periocular electrodes placed below the left eye and at the left and right outer canthi. All electrodes were referenced to the left mastoid with a chin ground. Impedances were kept below 15 kOhm.

Data were sampled at 1000 Hz and filtered online with an analog band-pass filter of 0.1 Hz to 250 Hz. A software-based 0.5 Hz high-pass filter was later used to remove DC drifts and 50 and 100 Hz notch filters were used to minimize line noise artifacts. These filters were designed to be linear-phase to minimize delay distortions. Data were also re-referenced to the average of all channels. To obtain accurate event onset times we placed a photodiode on the monitor to detect the onset of the stimuli. An external response device was used to collect response times. Both signals were collected on two external channels on the EEG amplifiers to ensure synchronization with the EEG data.

Movement artifact removal

Prior to the main experiment, we asked subjects to complete an eye movement calibration experiment during which they were instructed to blink repeatedly upon the appearance of a fixation cross in the center of the screen and then to make several horizontal and vertical saccades according to the position of the fixation cross. The fixation cross was subtended $0.6°\times0.6°$ of visual angle. Horizontal saccades subtended 20° and vertical saccades subtended 15°. This exercise enabled us to determine linear components associated with eye blinks and saccades (using principal component analysis) that were subsequently projected out of the EEG data recorded for the main experiment (Parra et al., 2003). Trials with strong eye movement or other movement artifacts were manually removed by inspection.

Modeling of behavioral data

We used a reinforcement-learning (Sutton and Barto, 1998) model to compute trial-by-trial PEs using each subject's behavioural responses. The model assigned each option i an expected value $q_i(t)$,

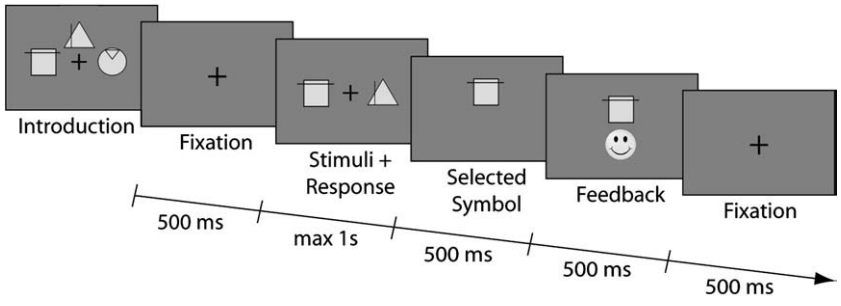

Fig. 1. Schematic representation of the experimental design. Each trial began with a fixation screen for 500 ms, followed by the presentation of two abstract symbols (selected from a combination of three symbols) for a maximum of 1 s. During this time subjects had to select, by pressing one of two buttons, the symbol that was most likely to lead to a reward. Upon response execution, the selected symbol appeared on the screen for 500 ms and it was followed by positive or negative feedback (smiley or frowney) for another 500 ms before a new trial began. See text for more details.

Please cite this article as: Philiastides, M.G., et al., Temporal dynamics of prediction error processing during reward-based decision making, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.05.052

which was used to derive choice probabilities $p_i(t)$ of choosing option i in trial t according to the softmax choice rule defined in Eq. (1).

$$p_i(t) = \exp[\gamma \cdot q_i(t)] / \sum_{j=1}^{n} \exp[\gamma \cdot q_i(t)] \quad (1)$$

where n is the number of options and γ is the sensitivity parameter that determined how strongly expected rewards influenced choices. After a decision was made in favor of one option, the received reward $r_i(t)$ was compared to the expected value $q_i(t)$, with the deviation formalized as PE: $\delta_i(t) = r_i(t) - q_i(t)$. Reinforcement learning models assume that learning is driven by these deviations, such that PEs are used to optimize reward predictions about the different choice options by updating the expected values:

$$q_i(t+1) = q_i(t) + \alpha \cdot \delta_i(t) \quad (2)$$

where α is the learning rate that determines the influence of the PE on the expectation updating process.

To derive PEs for the model-based EEG analysis, we first estimated two parameters individually for each participant: the sensitivity parameter of the choice rule γ and the learning rate α. We initially determined reasonably good parameters by a grid search while applying the following parameter constraints: $\gamma > 0$ and $0.1 < \alpha < 0.9$. The best parameters from the grid search were then used as starting points for a simplex optimization procedure, which determined the final parameter estimates. As a goodness-of-fit measure, we used the log likelihood of the observed choices over all trials T given the model and its parameters: $LL = \sum_{t=1}^{T} \ln[f_t(y|\theta)]$, where $f_t(y/\theta)$ denotes the probability of choice y in trial t given the model's parameter set θ. Predicted choice probabilities were calculated based on 1000 simulations per parameter set (combinations of the two free parameters), whereby in each simulation the model determined the choices used to update reward expectations (as opposed to observed choices). To obtain single-trial estimates of PE values, the two estimated participant specific parameters were then re-entered into the reinforcement learning algorithm, this time based on participants' observed choices.

All trials were sorted in four different groups according to their PE value. Initially, the trials were sorted according to the PE valence in positive and negative PE trials. Within each group the trials were further divided into two equally sized bins by performing a median split on the PE magnitudes. This resulted in a total of four groups of trials, which we used for all subsequent analyses: (i) high negative PE trials, (ii) low negative PE trials, (iii) high positive PE trials, and (iv) low positive PE trials.

Single trial EEG analysis

We used single-trial analysis of the EEG (Parra et al., 2002, 2005; Philiastides et al., 2006; Philiastides and Sajda, 2006, 2007; Ratcliff et al., 2009) to perform binary discriminations between conditions of interest. Specifically, we performed six pair-wise comparisons: (1) high negative vs. high positive PE trials, (2) low negative vs. low positive PE trials, (3) all negative vs. all positive PE trials, (4) low negative vs. high negative PE trials, (5) low positive vs. high positive PE trials, and (6) all low vs. all high PE trials. The first three comparisons were used to identify PE valence effects whereas the latter three were used to identify PE magnitude effects. The analysis was repeated for each subject separately.

For each comparison the method tries to identify, within short predefined time windows of interest, a projection in the multidimensional EEG space that maximally discriminates between each of the relevant conditions. Specifically, we defined time windows of interest with duration δ and onset time τ, and used logistic regression to estimate weighting vectors $\mathbf{w}_{\delta,\tau}$ (spatial filters) to generate one-dimensional projections $y_\tau(t)$ from D channels in the EEG data, denoted with $\mathbf{x}(t)$:

$$y_\tau(t) = \mathbf{w}_{\delta,\tau}^T \mathbf{x}(t) = \sum_{c=1}^{D} w_{\delta,\tau}^c x^c(t) \quad (3)$$

We used the reweighted least squares algorithm to learn the optimal discriminating spatial weighting vector $\mathbf{w}_{\delta,\tau}$ (Jordan and Jacobs, 1994). Samples within each time window τ are treated as independent and identically distributed. For all comparisons we used a window length $\delta = 50$ ms and feedback-locked onset times τ varying from 0 to 1000 ms, in increments of 5 ms. Compared to individual channel data the resulting "discriminating component" $y_\tau(t)$ is a better estimator of the underlying neural activity and is often thought to have better signal-to-noise ratio (SNR) and reduced interference from sources that do not contribute to the discrimination (Parra et al., 2005). We use the term "component" instead of "source" to make it clear that this is a projection of all the activity correlating with the underlying source.

Given the linearity of the model we also computed scalp topographies of the discriminating components resulting from Eq. (3) by estimating a forward model for each component:

$$\mathbf{a}_\tau = \frac{\mathbf{X}\mathbf{y}_\tau}{\mathbf{y}_\tau^T \mathbf{y}_\tau} \quad (4)$$

where the EEG data and discriminating components are now in a matrix and vector notation, respectively, for convenience (i.e., time is now a dimension of $\mathbf{X}$ and $\mathbf{y}_\tau$). Eq. (4) describes the electrical coupling of the discriminating component $\mathbf{y}_\tau$ that explains most of the activity in $\mathbf{X}$. Strong coupling indicates low attenuation of the component and can be visualized as the intensity of vector $\mathbf{a}_\tau$.

To visualize the profile of the discriminating components across all trials, we constructed discriminant component maps (as seen in Fig. 3). Specifically, after aligning trials to the onset of feedback the optimal vector $\mathbf{w}_{\delta,\tau}$ estimated for a given window τ is applied across an extended time window (here from 0 to 1000 ms after feedback). Each row of one such discriminant component map represents a single trial across time (i.e., $y^i(t)$, where i indexes trials). Discriminant component maps for each condition are shown with the mean (across trials) of the other condition subtracted (i.e., $y_1^i(t) - \bar{y}_2(t)$ and vice versa). Importantly, for each comparison the single-trial discriminator amplitudes for one condition are mapped to negative values and those of the other condition to positive values (shown with blue and red on the discriminant component maps, respectively). This mapping is arbitrary. For our purposes, we mapped the negative PE trials to negative values and the positive PE trials to positive values when discriminating along the PE valence dimension (i.e., comparisons 1–3 above). When discriminating along the PE magnitude dimension (i.e., comparisons 4–6 above) we mapped low PE trials to negative values and high PE trials to positive values.

To quantify the discriminator performance we used a split-half cross validation approach. Specifically, for each comparison we randomly split the trials into two halves (this resulted in at least 100 trials per split, per condition). The first half of the trials (training set) was used to estimate the vector $\mathbf{w}_{\delta,\tau}$, which was subsequently applied to the data from the second half (testing set). This in turn resulted in a test discriminator output, y_τ^{test}. Test performance was estimated by measuring the area under a Receiver Operating Characteristic (ROC) curve, also referred to as A_z value. An ROC curve plots the false positive rate vs. the true positive rate for a binary classification, which in our case means a comparison of the predicted trial classification based on y_τ^{test} with the true trial classification. These steps were repeated after flipping the training and testing sets. We reiterated through this procedure for a total of 10 split-half

Please cite this article as: Philiastides, M.G., et al., Temporal dynamics of prediction error processing during reward-based decision making, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.05.052

randomizations, which resulted in a total of 20 A_z values (per condition and window τ of interest). The final test performance was estimated by averaging across all A_z values.

Finally, to assess the significance of the resultant discriminating components, we used a bootstrapping technique to compute an A_z value leading to a significance level of $p = 0.01$. Specifically, for each of the 20 split-half runs we randomized the truth labels of the testing set 100 times, each time computing a new A_z value. This procedure yielded an A_z randomization distribution (2000 values) which we used to compute the A_z value leading to a significance level of $p = 0.01$.

Trial-by-trial association of PE magnitude and EEG component activity

To establish a relationship between model-based PEs and our single-trial EEG-derived components we performed a separate regression analysis. Specifically, we used the single-trial discriminator amplitude test values y_τ^{test}, derived from the different magnitude comparisons (4–6), to predict the single-trial PE magnitudes estimated using the model:

$$PE = \beta_0 + \beta_1 \times \mathbf{y}_\tau^{test} \quad (5)$$

Importantly, we ran this single-trial regression analysis for each condition (low and high PE magnitude groups) separately. For each subject, we repeated the analysis for all cross-validation runs and averaged the regression coefficients across runs. To establish a significant positive trial-by-trial association between PE magnitude and discriminator output we tested (for each of the low and high PE groups separately) whether the regression coefficients resulting from all subjects (β_1's), come from a distribution with mean greater than zero (using a one-tailed t-test).

Conventional ERP analysis

Average event-related potential (ERP) analysis was also performed to visualize the effects on pre-determined sensors of interests (i.e., sensors that according to the forward model $\mathbf{a}_\tau$ contribute most to the discrimination between trial types) in a more conventional manner. For each of the six comparisons performed using the single-trial analysis we identified the sensors with the strongest coupling to the discriminating components (i.e., highest $\mathbf{a}_\tau$ values) during time windows with significant A_z values (see Results). Note that, on average, these time windows overlap with the time windows for which reward and prediction error effects have been reported before. We subsequently computed feedback-locked ERPs and ERP difference waves, for each of the six comparisons and windows of interest, by averaging across subjects, trials, and sensors of interests.

Results

PE valence and magnitude components

To detect EEG components relating to the valence and magnitude of the PE signal we used a machine learning approach which identifies linear spatial weightings of the EEG sensors, for specific temporal windows, which optimally discriminated between negative and positive PEs (comparisons 1–3) and low and high magnitude PEs (comparisons 4–6), respectively. To quantify the discriminator's performance at each time window we used the area under the ROC curves (A_z value) resulting from each of these comparisons. Fig. 2A summarizes these results. For the PE valence comparisons we observed significant discrimination in a time range between 180 and 380 ms after the onset of feedback, with a pronounced peak around 220 ms. The PE magnitude comparisons resulted in significant discrimination in a narrower time range, which peaked around 320 ms after feedback onset.

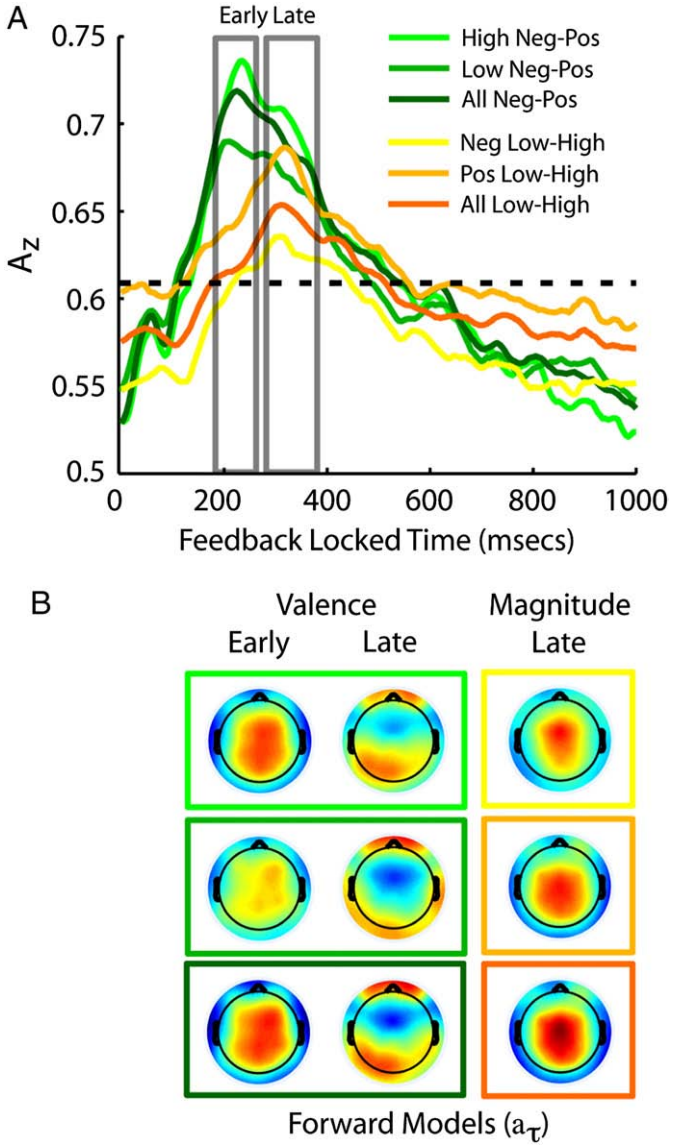

Fig. 2. Single-trial analysis reveals components related to PE valence and magnitude. (A) Single-trial discriminator performance (A_z) as a function of feedback-locked time. Shades of green represent the three valence comparisons (1–3) whereas shades of yellow represent the three magnitude comparisons (4–6). Results are averaged over all twenty cross-validation runs and over all subjects. The dotted black line represents the A_z leading to a significance level of $p = 0.01$. The gray boxes represent the early and late time windows. Note that discrimination for the valence comparisons remains significant during both of these time windows, while discrimination for the magnitude comparisons peak primarily in the late window. (B) Forward models (scalp maps, $\mathbf{a}_\tau$) for the three valence comparisons in the early and late windows (left and middle columns) and for the three magnitude comparisons in the late window (right column). For each cross-validation run we extracted the forward model from the time point with the highest A_z value (in each of the relevant windows—early or late) and then averaged across all cross-validation runs. Each subject's average forward model was normalized prior to averaging across subjects. Red represents positive correlation between the sensor readings and the extracted discriminating components whereas blue represents negative correlation. The coloured boxes follow the colour notation used in panel A. Note the distinct scalp distributions of each of the late valence and magnitude components.

Next, we looked at the forward models (scalp maps) that resulted from these comparisons at several time windows spanning the range of significant discrimination. The results are summarized in Fig. 2B. For the valence comparisons, in the range 180 to 380 ms after feedback we observed two distinct scalp distributions; an early one, peaking around 220 ms, with a broad spatial distribution over central electrodes and a later one, peaking around 300 ms, with major

Please cite this article as: Philiastides, M.G., et al., Temporal dynamics of prediction error processing during reward-based decision making, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.05.052

contributions from centrofrontal and occipitoparietal electrode sites. Due to the poor spatial resolution of the EEG and the possible effects of volume conduction it is unclear whether individual scalp maps reflect a single or multiple neural sources. However, significant variability across maps can often be attributed to different neural generators. Therefore, the distinct distribution of the early and late scalp maps (compare first and second columns in Fig. 2B) suggests the presence of two separate EEG valence components. For the magnitude comparisons, we found only a later component, peaking around 320 ms but with a strong central scalp distribution that was different from the late valence component. Note that the latter magnitude representation seems to extend to both positive and negative PEs (note similarity of forward models along the third column of Fig. 2B). The sign of the scalp distributions (forward model, $\mathbf{a}_T$) depends on the negative/positive mapping of the relevant conditions onto the discriminator output (see Materials and methods).

Based on these results we defined two time windows (henceforth, early and late windows) ranging from 180 to 260 ms, and from 280 to 380 ms after feedback, respectively. The distinct scalp distributions and the profile of the average ERP components (see Fig. 5) within each of these windows suggest that this is a reasonable differentiation. We refer to these two time windows for all further analyses.

Fig. 3 illustrates the effects of valence and magnitude for a single subject and for the group (Fig. 3A and B, respectively) in the form of discriminant component maps. For valence, the component maps were constructed from discriminating between all negative versus all positive PE trials during the early (left) and late (middle) windows. For magnitude, the component maps were constructed from discriminating between all low versus all high magnitude PE trials during the late (right) windows. All maps were constructed by averaging across the discriminant component maps resulting from each cross-validation run (and across subjects for the group results in Fig. 3B).

These discriminant component maps highlight a number of important points relating to the interpretability of our single-trial analysis results. First, note that for each comparison, trials from one condition are mapped to negative discriminator amplitudes (top row; effect shown in blue), while those from the second condition to positive amplitudes (bottom row; effect shown in red). Next, note the trial-to-trial variability represented in the discriminator output for each of the valence and magnitude components. Finally, note how this variability relates to the trial-to-trial variability in the model-based PE estimates for the magnitude component, but not for the valence components. To visualize this we sorted the trials in each of these maps based on the magnitude of the model-based PEs so that trials with the smallest PE magnitude values are on the top of each map.

On the one hand, low PE trials in the magnitude comparison (right top map in Fig. 3A and B) were mapped to negative discriminator amplitudes. From those, PE trials near zero were mapped to the most negative discriminator amplitudes. These maps show that as PE values increased, discriminator output gradually increased as well (amplitudes became less negative–less blue). On the other hand, high magnitude PE trials (right bottom map in Fig. 3A and B) were mapped to positive discriminator amplitudes. From those, PE trials near one were mapped to the most positive discriminator amplitudes. These maps, too, show that as the PE values increased, so did the discriminator output (amplitudes became more positive–more red). We quantify the degree of correlation between the single-trial discriminator amplitudes and the model-based PE values in the next section.

Trial-by-trial PE magnitude indexed by late EEG component

To provide quantitative support that the magnitude of the PE signal is encoded by the late magnitude component we performed an additional regression analysis (Eq. (5)). Specifically, for each of the three magnitude comparisons we used the single-trial discriminator amplitudes (at the

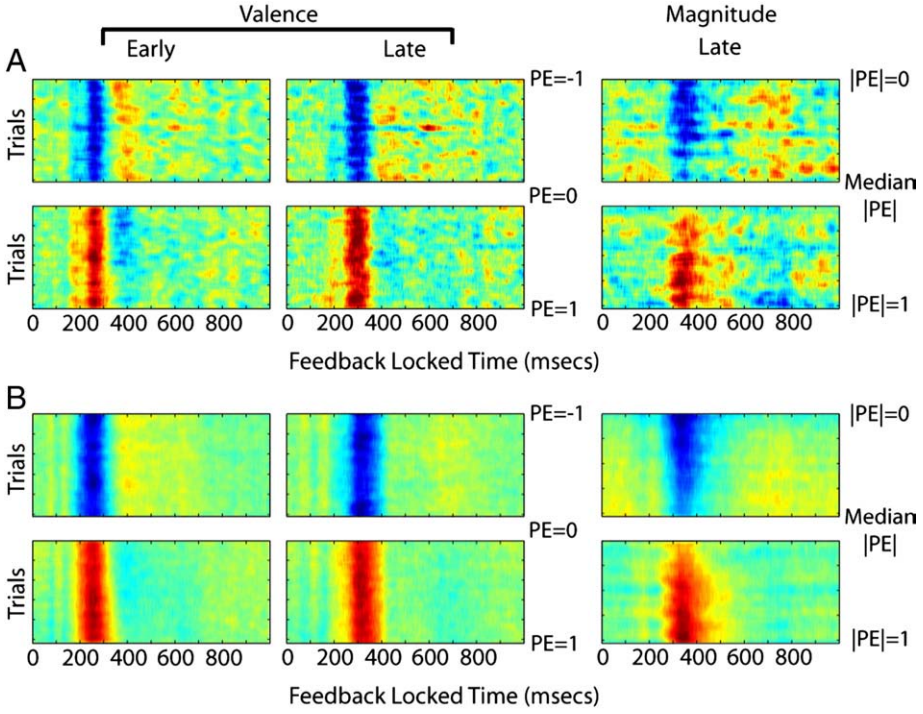

Fig. 3. Single-trial PE discriminant component maps for valence and magnitude. (A) Discriminating component maps from a representative subject showing single-trial data for the early and late valence components (left and middle panels) and the late magnitude component (right panels). (B) Average discriminant component maps across all subjects. Red represents positive and blue negative discriminator amplitudes. All trials are aligned to the onset of feedback and sorted by PE magnitude (in each map PE magnitude increases from top to bottom). For valence, the maps resulted from discriminating between all negative (top panels) versus all positive (bottom panels) PE trials. For magnitude, the maps resulted from discriminating between all low (top panel) versus all high (bottom panel) magnitude PE trials. To construct these maps we extracted the spatial weighting vectors $\mathbf{w}_{\delta,T}$ from the time point with the highest A_z value (in each of the relevant windows—early or late), applied it across all trials and all time (from 0 to 1000 ms after feedback) and then averaged the resulting maps across all cross-validation runs (and across all subjects in (B)).

Please cite this article as: Philiastides, M.G., et al., Temporal dynamics of prediction error processing during reward-based decision making, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.05.052

time of maximum discrimination) to predict the single-trial PE magnitudes estimated by the reinforcement learning model. Importantly, we ran the single-trial regression analysis for each condition (low and high PE magnitude groups) separately (i.e., we looked for correlation within–not across–the low and high PE magnitude groups). This ensured that the discriminator's ability to separate low and high PE magnitude trials would not dictate the trial-by-trial correlation between discriminator amplitudes and PE magnitudes.

We established a significant positive trial-by-trial association between PE magnitude and discriminator output by ensuring that the estimated regression coefficients (β_1's in Eq. (5)) were significantly greater than zero across participants. Note that for each condition (low and high PE magnitude groups) the regression coefficients were indeed significantly greater than zero (one-tail, *t*-test, all $p<0.01$) over all three magnitude comparisons. Results are shown in Fig. 4A.

Comparison to ERP component analysis

Conventional ERP component analysis relies on trial averaging to reveal mean differences across experimental conditions. Though trial averaging compromises single-trial variability, ERP data are sometimes easier to interpret and as such can help reinforce the main findings of our study. To visualize the valence and magnitude effects of PE using this approach we computed average ERP waves for the conditions in each of the six comparisons of interest along with a difference ERP waveform between each pair of conditions.

To allow for direct comparisons with our single-trial analysis methods we selected sensors of interest based on the forward models (scalp maps) shown in Fig. 2B. Specifically, to visualize the PE valence effects in the early and late windows we constructed ERP data from a cluster of central electrodes and a cluster of occipitoparietal electrodes, respectively. To visualize the PE magnitude effects in the late window we constructed ERP data from a cluster of centrofrontal electrodes. These results appear inline with our single-trial discriminator findings and they are illustrated in Fig. 5 where prototypical FRN and FRP ERP components can be seen. A major advantage of our single-trial analysis technique is, however, that it provides a rigorous, quantitative account of the effects of PE valence and magnitude across time and it helps dissociate their individual contributions (both in time and space as seen in Fig. 2A and B respectively).

Another important advantage of our single-trial analysis technique is that it spatially integrates information across the entire sensor space such that the resulting discriminating components have higher SNR compared to ERP data from individual or small subsets of sensors. To highlight this important advantage we compared the predictive power of single-trial information from individual sensors of interest to that of our discriminator output by introducing a second regression model:

$$PE = \beta_0 + \beta_1 \times \mathbf{y}_\tau^{test} + \beta_2 \times ERP_\tau^{test} \quad (6)$$

where y_τ^{test} is defined as in Eq. (5) and ERP_τ^{test} indicates single-trial data from a pre-defined time window and sensor of interest. To select the time and sensor of interest, we computed ERP difference scalp maps (across multiple feedback-locked time windows) for each of the three magnitude comparisons and selected the time and sensor with the highest ERP difference amplitude. We then defined a 50-ms window around this peak time and computed single-trial ERP amplitudes by averaging the samples within the 50 ms window. This was done to acquire more robust single-trial ERP estimates and to parallel the way we trained the single-trial discriminator (with 50 ms training windows).

This analysis addressed the question of how much variance in the PEs could be explained from single-trial discriminator and ERP amplitudes, respectively, when both regressors were present. If the performance of the single-trial analysis relied mostly on signals from individual sensors, regression coefficients for discriminator amplitudes should decrease drastically in the presence of the channel with the highest ERP difference.

We found that our discriminator output was indeed a better predictor than the ERP data from a single sensor of interest. Specifically, we showed that for each condition (low and high PE magnitude groups) the regression coefficients associated with y_τ^{test} (β_1's) were significantly greater than zero (one-tail, *t*-test, all $p<0.01$) over all three magnitude comparisons, while the majority of those associated with ERP_τ^{test} (β_2's) were not. These results are summarized in Fig. 4B. Orthogonalizing y_τ^{test} with respect to ERP_τ^{test} so that the common variance is absorbed by ERP_τ^{test} (average correlation between these two predictors was 0.29) did not alter the significance profile of the results as depicted in Fig. 4B.

This analysis clearly demonstrates the superiority of the single-trial method and emphasizes that using traditional ERP analysis techniques (e.g., simple correlation of sorted ERP derived amplitudes) could not unequivocally associate the late component with the encoding of the magnitude of the PE signal.

PEs and EEG component activity predict future choices

Reinforcement learning models postulate that PE signals are used to update future reward expectations thereby influencing future choice behavior. Here, we investigated how the PE valence and magnitude on the current trial could be used to predict choice on the subsequent trial. With respect to PE valence, participants should be more likely to repeat a choice after positive feedback, and less likely to repeat a choice after negative feedback. Fig. 6A shows that participants were indeed more likely to choose the same option on the next trial after positive compared to negative feedback (two-tailed, *t*-test, $p<0.01$).

With respect to the relation of PE magnitude and stay/switch decisions we found that the probability of repeating the same choice in the next trial was higher for low magnitude positive PEs (compared to high magnitude positive PEs) and for high magnitude negative PEs (compared to low magnitude negative PEs) as shown in Fig. 6A. While these findings appear at odds with previous reports (e.g., Cohen and Ranganath (2007)) the intricacies of our behavioral paradigm justify

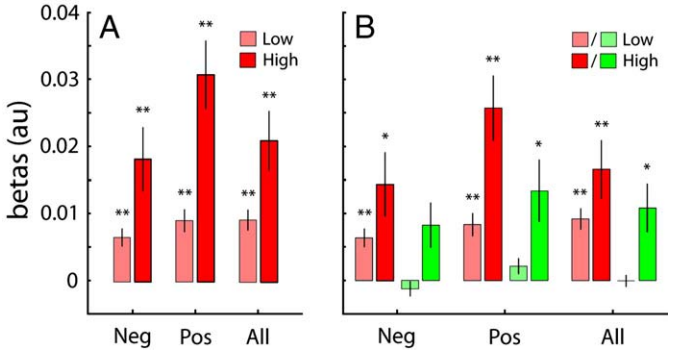

Fig. 4. PE magnitude encoded by late EEG component. (A) Regression coefficients (Eq. (5)) reflecting a significant trial-by-trial association between PE magnitude and discriminator output for each of the low (faint red) and high (solid red) PE magnitude groups over all three magnitude comparisons (negative PEs, positive PEs and both combined). (B) Regression coefficients from a regression that predicts single trial PEs with both, discriminator output (red) and ERP amplitudes of a single sensor of interest (green) (Eq. (6)). For the negative PE conditions ERP data were extracted from sensor FCz, while for the remaining conditions ERP data were extracted from sensor Cz. Note that unlike the discriminator output only two out of the possible six conditions showed a significant trial-by-trial association between PE magnitude and the single-electrode ERP data. Significance was established using a two-tailed *t*-test (null hypothesis: regression coefficients = 0) at $p<0.01$ (*) and $p<0.001$ (**). Error bars represent standard errors across subjects.

Please cite this article as: Philiastides, M.G., et al., Temporal dynamics of prediction error processing during reward-based decision making, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.05.052

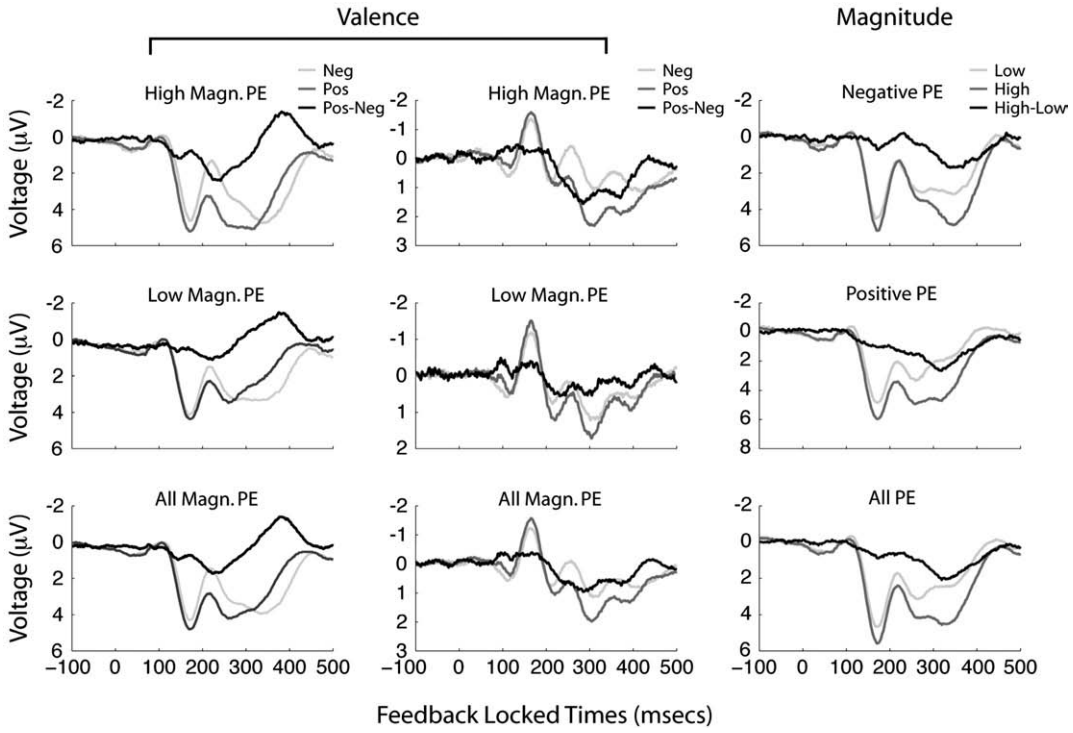

Fig. 5. PE valence and magnitude average ERP components. We identified sensors of interest from the forward models shown in Fig. 2. Specifically, for the early and late valence components we computed ERP averages over electrodes [FC1, FCz, FC2, C1, Cz, C2, CP1, CPz and CP2] and [P5, P3, P1, PO5, PO3 and POz] respectively. For the late magnitude component we computed ERP averages over electrodes [FC1, FCz, FC2, C1, Cz and C2]. For the valence comparisons light gray traces represent negative PE trials while dark gray traces represent positive PE trials. For the magnitude comparisons light gray traces represent low magnitude PE trials while dark gray traces represent high magnitude PE trials. For both the valence and magnitude comparisons the black traces represent the difference between the dark and light gray waveforms. Note that ERPs are plotted with the negative *y*-axis pointing up.

these effects (see also Fig. 7 in Discussion). Specifically, on any given trial subjects were required to choose one of two symbols that were selected among three alternatives, only one of which was associated with high reward probability. This experimental manipulation meant that the low magnitude positive PEs generally arose from choices of the good symbol (i.e., the one with high reward probability), whereas the high magnitude positive PEs came primarily from choices of one of the bad symbols. This in turn suggests that subjects were more likely to keep selecting the good symbol in future trials despite the occasional unexpected reward after choosing one of the bad symbols.

In contrast, low magnitude negative PEs generally arose from choosing one of the bad symbols since participants, who have already learned the reward associations, were aware that negative feedback would be the most likely outcome (regardless of their choice), whereas the high magnitude negative PEs came primarily from unexpected negative feedback on trials containing the good symbol. As with the positive PE trials, subjects were more likely to keep selecting the good symbol in future trials despite the occasional unexpected loss after choosing a good symbol. (Note that despite these trial-by-trial dynamics participants would still adapt to reversals, but this adaptation relied on changes in reward expectations that accumulated over repeated deviations from expectations).

Another way to realize these findings is in the form of PE magnitudes before switch and stay trials as shown in Fig. 6B. Specifically, the mean PE magnitude for negative PE trials was higher before stay than switch trials (two-tailed *t*-test, $p<0.01$) while the mean PE magnitude for positive PE trials was higher before switch than stay trials (two-tailed *t*-test, $p<0.01$).

To test whether trial-by-trial EEG component activity associated with PE valence and magnitude also carries information necessary to guide future choices we performed another regression analysis. Specifically, we tested whether the discriminator amplitudes associated with our EEG components on the current trial could be used to predict stay/switch patterns on the next trial:

$$P_{\text{stay}} = \beta_0 + \beta_1 \times \mathbf{y}_{\tau}^{\text{test}} \tag{7}$$

P_{stay} indicates the probability of repeating ones current choice in the next trial and it is realized as a vector of 1's (for stay) and 0's (for switches). y_{τ}^{test} indicates the single-trial discriminator amplitudes for the current trial. τ Indexes our early and late PE valence and our late PE magnitude components. For the PE valence components the analysis was run on all negative and positive PE trials whereas for the PE magnitude component the analysis was run on all low and high magnitude trials but for each of the negative and positive PE trials separately.

The results of this analysis were in line with the behavioural observations depicted in Fig. 6. Specifically, we found that single-trial discriminator information obtained from both the early and late PE valence components predicted that subjects were more likely to choose the same option on the next trial after positive compared to negative feedback (β_1's>0 in Eq. (7), two-tailed, *t*-test, $p<0.01$). With respect to the PE magnitude component, we found a positive association between P_{stay} and discriminator output for negative PE trials (β_1's>0 in Eq. (7), two-tailed, *t*-test, $p<0.01$) and a negative association between P_{stay} and discriminator output for positive PE trials (β_1's<0 in Eq. (7), two-tailed, *t*-test, $p<0.01$). In other words, as EEG activity from the late PE magnitude component increased subjects were more likely to repeat their choice after negative feedback and switch their choices after positive feedback. These results provide further support that the trial-by-trial fluctuations in our PE valence and magnitude EEG components represent information required to update reward expectations in order to adjust future behavior.

Please cite this article as: Philiastides, M.G., et al., Temporal dynamics of prediction error processing during reward-based decision making, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.05.052

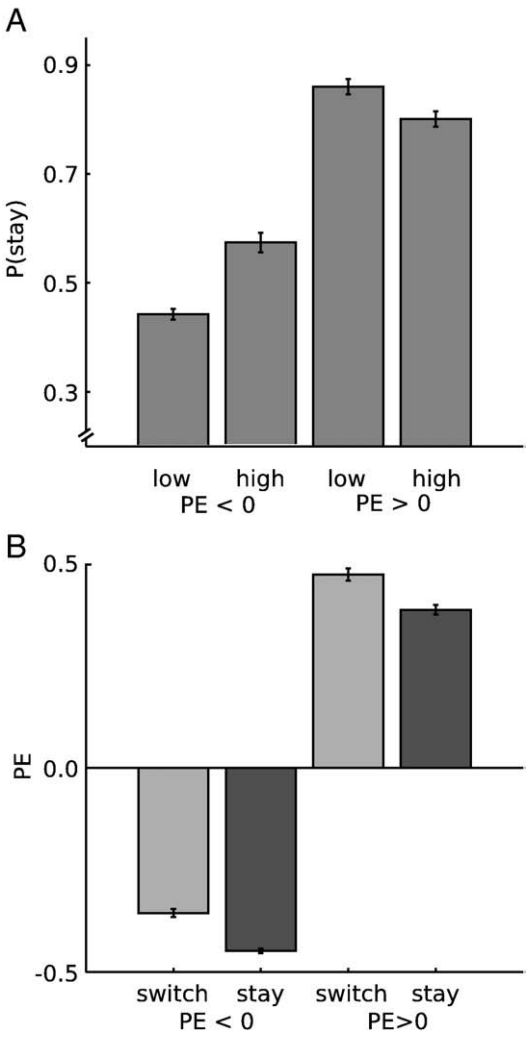

Fig. 6. Relationship between PEs on the current trial and stay or switch decisions on the next trial. (A) Probability to stay (P(stay)) for PE quartiles obtained by performing separate median splits for negative and positive PEs. As expected, stay decisions were generally more likely after positive than after negative PE trials. (B) PEs before stay or switch decisions, separately for negative and positive PE trials. Note that, as already suggested by the average P(stay) values in (A), there were more stay than switch decisions after higher magnitude negative PEs. In contrast, after higher magnitude positive PEs there were more switch than stay decisions. See text for more details.

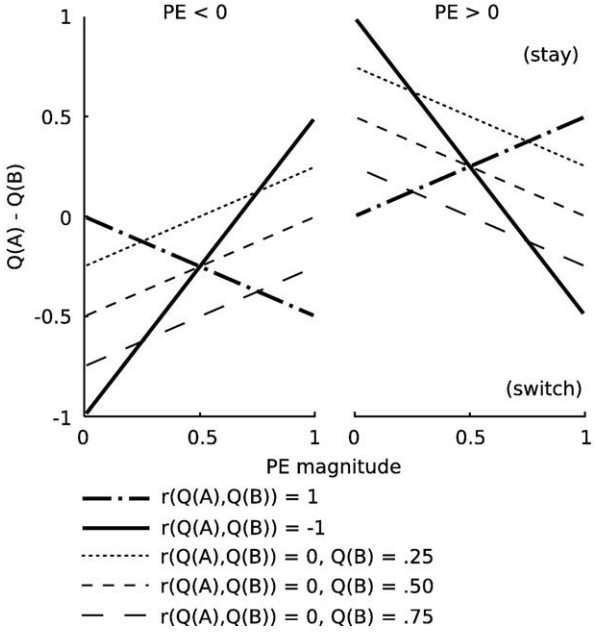

Fig. 7. Influence of positive and negative PEs on subsequent choices for different reward structures. Note that choices depend on reward expectations, so that the option with a higher reward expectation, Q, is more likely to be chosen. The *x*-axis shows magnitude of PEs. The *y*-axis shows the difference between the reward expectation for the current trial for two options, Q(A) and Q(B), after option A was chosen in the previous trial and it was either rewarded (positive PEs, right panel) or not rewarded (negative PEs, left panel). PE was used to update Q(A) with the delta learning rule (learning rate = 0.5). Larger positive differences between Q(A) and Q(B) imply higher probabilities to choose the same option again (i.e., higher stay probabilities). Note that a positive correlation between stay probability and magnitude of positive PEs (right panel) and a negative correlation between stay probability and magnitude of negative PEs can only be observed when Q(A) and Q(B) are positively correlated. When Q(A) and Q(B) are either negatively correlated or uncorrelated, positive PEs are negatively correlated with stay probability and negative PEs are positively correlated with stay probability. Importantly, even though this correlation is negative within positive PEs, the same PE magnitude results in higher stay probability after positive than after negative PEs. Indeed, the data from our experiment, where payoffs were by design negatively correlated, show this pattern: the stay probability is higher after positive feedback than after negative feedback and within positive (negative) PEs larger magnitude PEs are associated with a smaller (greater) stay probability.

Discussion

The goal of this work was to provide a temporal characterization of cortical PE processing during instrumental learning by using a model-based single-trial EEG approach. To date, only few EEG studies have used reinforcement learning models to estimate PEs. In particular, Cohen and Ranganath (2007) calculated ERPs based on PEs from a reinforcement learning model and showed that in a strategic economic game choices of the next trial can be predicted by PE valence and magnitude on the current trial. Cavanagh et al. (2010) incorporated PEs from a reinforcement learning model into their analysis to relate EEG power in the theta band to PE valence and magnitude. Those authors showed that PEs and theta band activity in the current trial are associated with changes in reaction time in the next presentation of the same stimuli. These recent advances over more traditional, model-free, ERP analysis are important, because only a quantitative trial-to-trial association between PEs and the underlying neuronal activity can provide unequivocal support that EEG signals carry information instrumental for trial-by-trial learning and not about global learning effects like uncertainty reduction through learning (see Fig. S1). By using a single-trial-analysis technique in combination with PEs from a reinforcement-learning model we advanced the approaches taken by Cohen and Ranganath (2007) and Cavanagh et al. (2010) in three important ways. First, using a rigorous cross validation scheme we showed that neural activity could be used to *predict* PE valence and magnitude on a trial-by-trial level. Second, we established a quantitative trial-by-trial association between neural activity and PE valence and magnitude. Third, we showed that these trial-by-trial changes in the EEG could be used to *predict choices* from one trial to the next.

To examine how PE signals are encoded in the EEG, we designed a probabilistic reversal learning task to elicit a broad range of PEs and used a machine learning approach to identify linear spatial weightings of the EEG sensors for specific temporal windows, which optimally discriminated between trials conditioned along different PE valence and magnitude dimensions. We then used the trial-by-trial information extracted from these discriminating components to predict trial-by-trial changes in PEs estimated with a reinforcement learning model. Moreover, we showed that trial-by-trial choice dynamics could be predicted from model-based PEs as well as from neuronal activity associated with PE valence and magnitude.

Specifically, using our single-trial linear discriminator, we identified two PE processing stages that evolve gradually over time while providing a clear dissociation between the encoding of valence and magnitude of the PE signal. The timing and scalp distribution of the

Please cite this article as: Philiastides, M.G., et al., Temporal dynamics of prediction error processing during reward-based decision making, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.05.052

early, valence component are consistent with previous reports on the FRN (Gehring and Willoughby, 2002; Cohen and Ranganath, 2007; Frank et al., 2007; Bellebaum and Daum, 2008; Eppinger et al., 2008; Holroyd et al., 2009) and suggest an early quick categorical evaluation whether the outcome is negative or positive. The temporally overlapping representations of the later valence and magnitude components suggest a second processing stage where the complete information needed to update reward expectation is simultaneously represented in what appear to be spatially distinct neural populations. Hence, our results suggest that, on the cortical level, the brain represents the necessary information for updating reward expectations within around 300 ms after feedback. Importantly, by combining the modeling of behavioral data with single-trial analysis of EEG data we were able to establish a quantitative relationship between PEs and feedback-induced potentials, strongly suggesting that the PE signals encoded in late potentials are instrumental in learning the expected reward of different choice options. The finding that the late discriminator amplitudes were predictive of next-trial choices further supports their relevance to learning.

PE valence

The similarity between the timing (~200 ms post feedback) and the spatial distribution of our early valence component and the previously reported FRN (i.e., strong differential activity between gains and losses at centrofrontal electrode sites) suggest that this potential is indeed part of the complex representing PE valence. The FRN was observed in a number of tasks, including gambling tasks that do not require learning (Gehring and Willoughby, 2002; Yasuda et al., 2004; Yeung and Sanfey, 2004; Hajcak et al., 2007) and instrumental learning tasks (Holroyd and Coles, 2002; Cohen et al., 2007; Bellebaum and Daum, 2008; Eppinger et al., 2008). While our data are consistent with previous studies in that the FRN encodes feedback valence, previous results are inconsistent regarding the role of the FRN in the representation of PE magnitude.

One influential theory proposes that the FRN provides a quantitative representation of negative PEs in learning tasks (Holroyd and Coles, 2002), but empirical results do not generally support this theory (Bellebaum and Daum, 2008; Mathewson et al., 2008; Holroyd et al., 2009). In particular, none of the studies reporting a representation of magnitude for negative PEs did so on a trial-by-trial basis, leaving open the question of whether these signals reflect local PE signals or signal changes due to more global effects like overall changes in reward probabilities or uncertainty reduction due to learning (see Fig. S1 for a Bayesian perspective on the relationship between uncertainty on the one hand, and reward probability and experience on the other hand). However, Cavanagh et al. (2010) showed with a trial-by-trial analysis that theta band activity over the medial prefrontal cortex is related to PE valence and magnitude, which in turn they associated with the FRN. Still, similar to earlier ERP studies, these authors do not find a magnitude effect for positive PEs in the FRN time window over medial sensors. This is relevant because even if we consider the FRN as a quantitative signal for negative PEs, there is no evidence for a quantitative signal for positive PEs, which are equally important for learning (e.g., the new good option in a reversal learning task can only be learned from positive PEs).

Our results, however, offer a different account. Specifically, we showed that 200 ms after feedback the brain encodes outcome evaluation primarily categorically, i.e., it is signaling whether the outcome was positive or negative without distinct PE magnitude representations. Note that this finding is ecologically sound because to survive in an uncertain world one first needs to quickly assess whether an outcome is good or bad in order to take immediate action (if necessary) before engaging in estimating all relevant information required to plan future actions.

In addition to the early PE valence encoding the brain appears to maintain a valence representation for a later processing stage as evident by our late valence component. Unlike previous ERP studies that provided inconsistent evidence for the role of late potentials (e.g., FRP) in the encoding of PE valence (e.g., some showed larger FRP amplitudes for losses (Cohen et al., 2007; Mathewson et al., 2008), others for gains (Johnston and Holcomb, 1980; Bellebaum and Daum, 2008), and some reported no difference at all (Yeung and Sanfey, 2004; Sato et al., 2005), we found clear and distributed differences between positive and negative PE trials. Inspection of the scalp distribution of our late valence component suggests that the neural generators of this component are different from those of the early valence component indicating that early valence representations are subsequently relayed to a separate set of brain regions for further processing. This finding confirms our original hypothesis, which postulates that PE processing evolves gradually over time in a distributed network.

PE magnitude

In addition to an effect of PE valence in the later window (~300 ms after feedback), our single-trial analysis approach allowed us to identify a temporally overlapping but separate representation of PE magnitude. This finding suggests that the complete set of information needed to update reward expectations is simultaneously available around 300 ms after feedback. In addition, comparing the scalp distributions of the late valence and magnitude components clearly demonstrates that these components, too, have distinct neural generators. The spatial distribution of the late PE magnitude component suggests that central electrodes contribute most to the discriminator's performance by exploiting the higher signal positivity after unexpected as compared to expected events.

A stronger positivity for unexpected events is generally reported in the context of odd-ball tasks (Polich, 2007), but was recently also reported in the context of instrumental learning (Cohen and Ranganath, 2007; Bellebaum and Daum, 2008). One influential theory in particular, proposes that the signal positivity observed 300 ms after the onset of an oddball stimulus (i.e., P300 component) reflects a process of context updating, where new, unexpected, and task relevant information is incorporated into the representation of the environment (Donchin and Coles, 1998; Polich, 2007). While the notion of context updating relates to the idea of updating reward expectations, it is not clear if the same cognitive and neural processes are at work in both situations, as the oddball task differ substantially from the probabilistic reinforcement learning task we used here. Nonetheless, the similarity of the neural signatures observed in both situations suggests that updating of context and reward expectation might share some common processes.

Importantly, the trial-by-trial neuronal variability obtained from the late PE magnitude component correlated significantly with the single-trial PE information estimated by a reinforcement learning model applied to the behavioral data of individual subjects. This finding provides a quantitative link between the two measures that would not have been detected using conventional ERP analysis and strongly suggests that late EEG-potentials are instrumental in learning expected rewards. In exploiting the trial-to-trial variability in the data, our approach is a substantial improvement over traditional average ERP approaches (Holroyd and Coles, 2002, 2009; Cohen et al., 2007) and as such could also be used to identify other error-related process (e.g. error processing before and after learning or in situations of variable uncertainty) in the future.

An alternative interpretation of the role of our PE magnitude component is that it merely reflects the degree of surprise (or expectancy) of an outcome and not a real PE signal. From a computational point of view this is difficult to address because the magnitude of the PE and the degree of surprise are equivalent. However, the fact that neuronal activity predictive of PE magnitude

Please cite this article as: Philiastides, M.G., et al., Temporal dynamics of prediction error processing during reward-based decision making, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.05.052

(resulting from our single-trial discriminator) can be used to predict future stay/switch decisions suggests that this signal is directly involved in learning and decision making, most likely through updating of future reward expectations.

Prediction of future choices

Our results showed that participants were more likely to choose the same action after they were rewarded. Consistent with this observation, stay/switch decisions were also associated with single-trial neuronal information obtained from a classifier trained to discriminate between negative and positive outcomes. While these results are intuitively appealing, we also reported results that may at first glance seem somewhat counterintuitive. Specifically, we found that as PE magnitude increased stay decisions were more likely after negative feedback and switch decisions were more likely after positive feedback. Similarly, as neuronal activity from our late PE magnitude component increased subjects were more likely to repeat their choice after negative feedback and switch their choices after positive feedback.

Despite the fact that these results contradict the pattern reported by Cohen and Ranganath (2007) they are still perfectly justifiable when taking into account the payoff structure of our task, as already highlighted in the Results section. Fig. 7 provides a graphical representation of the mathematical analysis of the relation of payoff structures and stay/switch decisions as predicted by a reinforcement learning model. Specifically, the relationship between PEs and stay/switch decisions we observed holds for tasks in which the correlation between the payoffs of different options is zero or smaller. Intuitively, when payoffs are negatively correlated as in our task, high positive PEs result from unexpected rewards coming from an option that mostly produces no rewards. By comparison, low positive PEs result from expected rewards coming from an option that mostly produces rewards. Most learners will, however, persist in choosing the option that was more frequently associated with rewards in the past, which in turn leads to the pattern of behavioral choices reported here. This pattern is only broken when payoffs of the available options are positively correlated, as might happen in interdependent economic games like the one studied by Cohen and Ranganath (2007), where periods in which participants outsmarted their opponent and earned rewards on most trials (irrespective of which option they chose) could be interleaved with periods in which the reverse was true (i.e., participants were outsmarted by their opponent, and received no rewards in most trials).

PE representation in one or multiple systems?

The proposal of a sequential and distributed representation of PEs at the cortical level diverges in some sense from the common view that PEs are signaled by firing of dopaminergic neurons in the midbrain that increases when positive PEs become larger and that pauses longer when negative PEs become larger. Even though the assumption of a sequential and distributed representation of PEs might seem less elegant than the assumption that the cortex simply reflects dopaminergic PE signals, it is more consistent with the available empirical results. First, note that dopaminergic firing from neurons in the ventral tegmental area and the substantia nigra has a latency of around 50–110 ms after reward-stimulus onset (Overton and Clark, 1997; Schultz et al., 1997; Redgrave et al., 1999) and peaks around 200 ms. The effect of dopaminergic neurons onto the cortex however, is comparatively slow (Lapish et al., 2007), making it unlikely that a rapid response like the FRN (e.g., our early valence component), which peak at around 220 ms, originates in the midbrain. Second, available results from EEG experiments are inconsistent with the proposal of a monotonic representation of PEs in the FRN. Note that a monotonic PE representation requires the following order for the strength of a neuronal signal: high positive PEs > small positive PEs > small negative PEs > high negative PEs, or the reverse order (Caplin and Dean, 2008). To our knowledge, no EEG study has so far reported such a pattern in the relevant ERP components or their associated frequency characteristics. For instance, Cohen and Ranganath (2007) did not find that FRN amplitudes differ significantly for positive PEs and Cavanagh et al. (2010) found that theta power is greater for high positive and negative PEs (over lateral but not medial prefrontal cortex). Given that the substantial number of EEG studies examining PEs failed to find conclusive evidence in favor of a monotonic relationship between PEs and EEG activity from a unitary system, it is reasonable to consider and examine alternative hypotheses.

The proposal of a sequential and distributed processing of PEs on the cortical level invites the question of how this PE representation relates to dopaminergic PE signals. Given that EEG does not record signals originating from deeper brain structures, we can only speculate about the interactions of the two systems. One possibility is that the central nucleus of the amygdala (CNA) provides input to dopaminergic midbrain regions, as well as to the ventromendial prefrontal cortex (vmPFC), ACC, and parietal cortex, which are potential generators for the observed feedback-related potentials. Indeed, the CNA represents values of conditioned stimuli (Ono et al., 1995; Baxter and Murray, 2002) and is involved in PE processing (Yacubian et al., 2006; Niv and Schoenbaum, 2008). Furthermore, the CNA has afferent projections to dopaminergic midbrain regions and also to vmPFC, ACC, and parietal cortex (Schultz and Romo, 1990; Phillips et al., 2003; Pierce et al., 2007). Future research, using simultaneously acquired EEG-fMRI data, combined with a single-trial methodology that simultaneously discriminates between both valence and magnitude (e.g., with multinomial logistic regression) will be required to provide a more complete representation of the spatiotemporal characteristics of PE processing in the human brain.

In conclusion, our results suggest that PE processing takes place in a distributed network and that it proceeds in two main stages that evolve gradually over time after feedback. Specifically, the first processing stage starts with a categorical assessment of the outcome as good or bad, which is reflected in an early EEG component arising around 220 ms after feedback. In the next 100 ms, this early evaluation of outcome valence is followed by a second processing stage, which reflects both PE valence and magnitude. These later PE valence and magnitude representations are reflected in two spatially distinct but temporally overlapping EEG components suggesting that updating reward expectations requires simultaneous access to both representations.

Acknowledgments

This work was supported by the Max Planck Society, the Deutsche Forschungsgemeinschaft (Emmy Noether Programme), and the Federal Ministry of Education and Research.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2010.05.052.

References

Apicella, P., Ljungberg, T., Scarnati, E., Schultz, W., 1991. Responses to reward in monkey dorsal and ventral striatum. Exp. Brain Res. 85, 491–500.
Balleine, B.W., Delgado, M.R., Hikosaka, O., 2007. The role of the dorsal striatum in reward and decision-making. J. Neurosci. 27, 8161–8165.
Baxter, M.G., Murray, E.A., 2002. The amygdala and reward. Nat. Rev. Neurosci. 3, 563–573.
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.
Bellebaum, C., Daum, I., 2008. Learning-related changes in reward expectancy are reflected in the feedback-related negativity. Eur. J. Neurosci. 27, 1823–1835.

Buch, E.R., Brasted, P.J., Wise, S.P., 2006. Comparison of population activity in the dorsal premotor cortex and putamen during the learning of arbitrary visuomotor mappings. Exp. Brain Res. 169, 69–84.
Caplin, A., Dean, M., 2008. Axiomatic methods, dopamine and reward prediction error. Curr. Opin. Neurobiol. 18, 197–202.
Cavanagh, J.F., Frank, M.J., Klein, T.J., Allen, J.J.B., 2010. Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. Neuroimage 49, 3198–3209.
Cohen, M.X., Ranganath, C., 2007. Reinforcement learning signals predict future decisions. J. Neurosci. 27, 371–378.
Cohen, M.X., Elger, C.E., Ranganath, C., 2007. Reward expectation modulates feedback-related negativity and EEG spectra. NeuroImage 35, 968–978.
Dayan, P., Niv, Y., 2008. Reinforcement learning: the good, the bad and the ugly. Curr. Opin. Neurobiol. 18, 185–196.
Donchin, E., Coles, M.G.H., 1998. Context updating and the P300. Behav. Brain Sci. 21, 149–168.
Eppinger, B., Kray, J., Mock, B., Mecklinger, A., 2008. Better or worse than expected? Aging, learning, and the ERN. Neuropsychologia 46, 521–539.
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., D'Lauro, C., Curran, T., 2007. Cross-task individual differences in error processing: neural, electrophysiological, and genetic components. Cogn. Affect. Behav. Neurosci. 7, 297–308.
Gehring, W., Willoughby, A., 2002. The medial frontal cortex and the rapid processing of monetary gains and losses. Science 295, 2279–2282.
Hajcak, G., Holroyd, C., Moser, J., Simons, R., 2005. Brain potentials associated with expected and unexpected good and bad outcomes. Psychophysiology 42, 161–170.
Hajcak, G., Moser, J., Holroyd, C., Simons, R., 2006. The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. Biol. Psychol. 71, 148–154.
Hajcak, G., Moser, J.S., Holroyd, C.B., Simons, R.F., 2007. It's worse than you thought: The feedback negativity and violations of reward prediction in gambling tasks. Psychophysiology 44, 905–912.
Holroyd, C., Coles, M.G.H., 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.H., 2008. Dorsal anterior cingulate cortex integrates reinforcement history to guide voluntary behavior. Cortex 44, 548–559.
Holroyd, C.B., Krigolson, O.E., Baker, R., Lee, S., Gibson, J., 2009. When is an error not a prediction error? An electrophysiological investigation. Cogn. Affect. Behav. Neurosci. 9, 59–70.
Horst, R.L., Johnson, R., Donchin, E., 1980. Event-related brain potentials and subjective probability in a learning task. Mem. Cogn. 8, 476–488.
Johnston, R., Donchin, E., 1978. On how P300 amplitude varies with the utility of the eliciting stimuli. Electroencephalogr. Clin. Neurophysiol. 44, 424–437.
Johnston, V.S., Holcomb, P.J., 1980. Probability learning and the P3 component of the visual evoked potential in man. Psychophysiology 17, 396–400.
Jordan, M.I., Jacobs, R.A., 1994. Hierarchical mictures of experts and the EM algorithm. Neural Comput. 6, 181–214.
Krugel, L.K., Biele, G., Mohr, P.N.C., Li, S.-C., Heekeren, H.R., 2009. Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. Proc. Nat. Acad. Sci. 106, 17951–17956.
Lapish, C.C., Kroener, S., Durstewitz, D., Lavin, A., Seamans, J.K., 2007. The ability of the mesocortical dopamine system to operate in distinct temporal modes. Psychopharmacology 191, 609–625.
Mathewson, K., Dywan, J., Snyder, P., Tays, W., Segalowitz, S., 2008. Aging and electrocortical response to error feedback during a spatial learning task. Psychophysiology 45, 936–948.
Matsumoto, M., Matsumoto, K., Abe, H., Tanaka, K., 2007. Medial prefrontal cell activity signaling prediction errors of action values. Nat. Neurosci. 10, 647–656.
Nieuwenhuis, S., Aston-Jones, G., Cohen, J.D., 2005. Decision making, the P3, and the locus coeruleus-norepinephrine system. Psychol. Bull. 131, 510–532.
Niv, Y., Schoenbaum, G., 2008. Dialogues on prediction errors. Trends Cogn. Sci. 12, 265–272.
O'Doherty, J., Critchley, H., Deichmann, R., Dolan, R.J., 2003. Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. J. Neurosci. 23, 7931–7939.
Ono, T., Nishijo, H., Uwano, T., 1995. Amygdala role in conditioned associative learning. Prog. Neurobiol. 46, 401–422.
Overton, P.G., Clark, D., 1997. Burst firing in midbrain dopaminergic neurons. Brain Res. Brain Res. Rev. 25, 312–334.
Parra, L.C., Alvino, C., Tang, A., Pearlmutter, B., Yeung, N., Osman, A., Sajda, P., 2002. Linear spatial integration for single-trial detection in encephalography. NeuroImage 17, 223–230.
Parra, L.C., Spence, C.D., Gerson, A.D., Sajda, P., 2003. Response error correction—a demonstration of improved human–machine performance using real-time EEG monitoring. IEEE Trans. Neural Syst. Rehabil. 11, 173–177.
Parra, L.C., Spence, C.D., Gerson, A.D., Sajda, P., 2005. Recipes for the linear analysis of EEG. NeuroImage 28, 326–341.
Philiastides, M.G., Sajda, P., 2006. Temporal characterization of the neural correlates of perceptual decision making in the human brain. Cereb. Cortex 16, 509–518.
Philiastides, M.G., Sajda, P., 2007. EEG-informed fMRI reveals spatiotemporal characteristics of perceptual decision making. J. Neurosci. 27, 13082–13091.
Philiastides, M.G., Ratcliff, R., Sajda, P., 2006. Neural representation of task difficulty and decision making during perceptual categorization: a timing diagram. J. Neurosci. 26, 8965–8975.
Phillips, A.G., Ahn, S., Howland, J.G., 2003. Amygdalar control of the mesocorticolimbic dopamine system, parallel pathways to motivated behavior. Neurosci. Biobehav. Rev. 27, 543–554.
Pierce, L., Krigolson, O., Tanaka, K., Holroyd, C., 2007. The ERN and reinforcement learning in a difficult perceptual expertise task. Can. J. Exp. Psychol. 61, 372-372.
Polich, J., 2007. Updating P300: an integrative theory of P3a and P3b. Clin. Neurophysiol. 118, 2128–2148.
Ratcliff, R., Philiastides, M.G., Sajda, P., 2009. Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. Proc. Nat. Acad. Sci. 106, 6539–6544.
Redgrave, P., Prescott, T.J., Gurney, K., 1999. Is the short-latency dopamine response too short to signal reward error? Trends Neurosci. 22, 146–151.
Rescorla, R., Wagner, A.D., 1972. A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: Black, A., Prokasy, W., Black, A., Prokasy, W. (Eds.), Classical conditioning II: current research and theory, pp. 64–99.
Rey, A., 1964. L'Examen Clinique en Psychologie. Presses Universitaire de France, Paris.
Sato, A., Yasuda, A., Ohira, H., Miyawaki, K., Nishikawa, M., Kumano, H., Kuboki, T., 2005. Effects of value and reward magnitude on feedback negativity and P300. Neuroreport 16, 407–411.
Schultz, W., 2006. Behavioral theories and the neurophysiology of reward. Annu. Rev. Physiol. 57, 87–115.
Schultz, W., Romo, R., 1990. Dopamine neurons of the monkey midbrain: contingencies of responses to stimuli eliciting immediate behavioral reactions. J. Neurophysiol. 63, 607–624.
Schultz, W., Dayan, P., Montague, P.R., 1997. A neural substrate of prediction and reward. Science 275, 1593–1599.
Sutton, R., Barto, A., 1998. Reinforcement learning, An introduction. .
Tremblay, L., Schultz, W., 2000. Reward-related neuronal activity during go–nogo task performance in primate orbitofrontal cortex. J. Neurophysiol. 83, 1864–1876.
Yacubian, J., Gläscher, J., Schroeder, K., Sommer, T., Braus, D.F., Büchel, C., 2006. Dissociable systems for gain- and loss-related value predictions and errors of prediction in the human brain. J. Neurosci. 26, 9530–9537.
Yasuda, A., Sato, A., Miyawaki, K., Kumano, H., Kuboki, T., 2004. Error-related negativity reflects detection of negative reward prediction error. Neuroreport 15, 2561–2565.
Yeung, N., Sanfey, A.G., 2004. Independent coding of reward magnitude and valence in the human brain. J. Neurosci. 24, 6258–6264.
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.

Please cite this article as: Philiastides, M.G., et al., Temporal dynamics of prediction error processing during reward-based decision making, NeuroImage (2010), doi:10.1016/j.neuroimage.2010.05.052

SUPPLEMENTARY MATERIAL

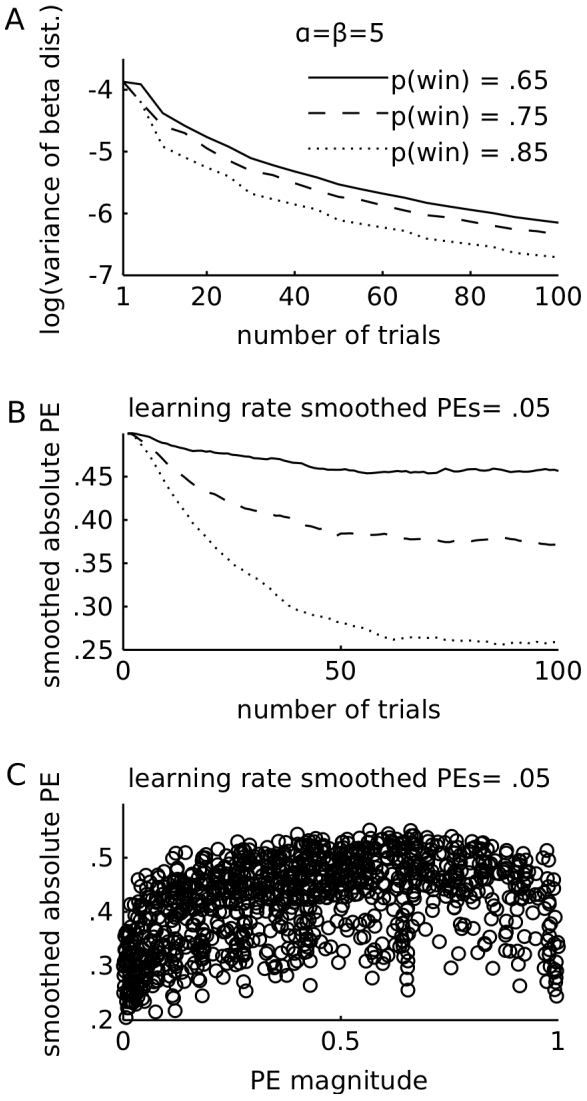

Figure S1: Uncertainty and prediction errors (PEs) over the course of learning. One way to measure uncertainty in a binary decision task is to model the reward expectation as a beta distribution (c.f. Behrens et al., 2007), whereby the first moment represents the expected winning probability and the second moment represents the uncertainty associated with the expectation. (A) Change in the variance of a posterior beta distribution that is

progressively updated with more samples, drawn from binary payoff distributions with different winning probabilities. Uncertainty measured as variance is reduced as sample size increases and uncertainty is generally higher for win probabilities near .5 (holding sample size constant). This is problematic for ERP studies using comparisons between trials during early and late learning or between trials with low (near .5) and high winning probabilities to examine representation of PE magnitude, because it shows that PE magnitude and uncertainty are confounded when using these simple approaches. (B) A simpler measure for uncertainty (which is likely easier to implement for the brain than bayesian updating) is the magnitude of PEs. PE magnitudes can be smoothed over time (e.g. with a delta learning rule with small learning rate) to obtain more stable estimates of uncertainty (c.f. Krugel et al., 2009). Note that using smoothed PEs as a measure of uncertainty implies the same qualitative relation between sample size and probability of a win [p(win)] and uncertainty as when using the variance of the beta distribution. Graphs show average smoothed PEs from 250 simulated learning paths for each p(win) based on model parameters from one exemplary participant. (C) Scatter plot showing that trial-by-trial PE magnitudes estimated with a RL model are independent of uncertainty (measured by smoothed PE magnitudes). Hence, the results of a trial-by-trial analysis of PE magnitudes cannot be explained by uncertainty.

SUPPLEMENTARY REFERENCES

Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS (2007) Learning the value of information in an uncertain world. Nat Neurosci 10,1214-1221.

Krugel LK, Biele G, Mohr PNC, Li S-C, Heekeren HR (2009) Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. Proc Nat Acad Sci 106,17951-17956.