Cogn Affect Behav Neurosci (2017) 17:406–421
DOI 10.3758/s13415-016-0487-3

Electrophysiological correlates reflect the integration of model-based and model-free decision information

Ben Eppinger[1,2] · **Maik Walter**[3] · **Shu-Chen Li**[1,4]

Published online: 3 January 2017

Abstract In this study, we investigated the interplay of habitual (model-free) and goal-directed (model-based) decision processes by using a two-stage Markov decision task in combination with event-related potentials (ERPs) and computational modeling. To manipulate the demands on model-based decision making, we applied two experimental conditions with different probabilities of transitioning from the first to the second stage of the task. As we expected, when the stage transitions were more predictable, participants showed greater model-based (planning) behavior. Consistent with this result, we found that stimulus-evoked parietal (P300) activity at the second stage of the task increased with the predictability of the state transitions. However, the parietal activity also reflected model-free information about the expected values of the stimuli, indicating that at this stage of the task both types of information are integrated to guide decision making. Outcome-related ERP components only reflected reward-related processes: Specifically, a medial prefrontal ERP component (the feedback-related negativity) was sensitive to negative outcomes, whereas a component that is elicited by reward (the feedback-related positivity) increased as a function of positive prediction errors. Taken together, our data indicate that stimulus-locked parietal activity reflects the integration of model-based and model-free information during decision making, whereas feedback-related medial prefrontal signals primarily reflect reward-related decision processes.

Keywords Decision making · Model-based · Model-free · EEG · Reinforcement learning

Ben Eppinger and Maik Walter shared first authorship

Electronic supplementary material The online version of this article (doi:10.3758/s13415-016-0487-3) contains supplementary material, which is available to authorized users.

✉ Ben Eppinger
ben.eppinger@concordia.ca

[1] Chair of Lifespan Developmental Neuroscience, Department of Psychology, TU Dresden, Dresden, Germany

[2] Department of Psychology, Concordia University, Loyola Campus, 7141 Sherbrooke Street W., Montreal, Quebec, Canada

[3] Institute for Customer Insight, University of St. Gallen, St. Gallen, Switzerland

[4] Center of Lifespan Psychology, Max Planck Institute for Human Development, Berlin, Germany

Investment decisions (such as buying stocks) can be driven by different types of information: In many cases, past experience regarding the performance (i.e., the value) of a company may serve as a good indicator of its future success. However, considering other information about the economic environment in which the company is operating can also be helpful in guiding decision making (Hodgkinson, Brown, Maule, Glaister, & Pearman, 1999; Pezzulo, Rigoli, & Friston, 2015). In analogy to this economic example, most current psychological theories of value-based decision making propose that two qualitatively distinct reinforcement-learning (RL) systems are involved in regulating choice behavior (Balleine & O'Doherty, 2010; Daw, Niv, & Dayan, 2005). The habitual or model-free RL system learns to choose actions on the basis of their rewarding or punishing consequences. Although computationally swift, model-free learning is slow in adapting to changes in contingencies, which is disadvantageous in dynamically changing environments (Dayan & Niv, 2008; Doll, Simon, & Daw, 2012). The goal-directed or model-based RL system uses experience to learn an internal model or cognitive map of the environment (Tolman, 1948). As a consequence, model-based decision making allows greater behavioral flexibility but is also more

costly in terms of cognitive resources (Dayan & Niv, 2008; Doll et al., 2012).

Recent studies that have investigated the neural mechanisms underlying model-free and model-based decision making have revealed inconclusive results. The findings from fMRI studies show no specific neural correlate of model-based processes, but instead point toward overlapping neural activity for both strategies in the ventral striatum (Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Deserno et al. 2015). In contrast, the results from a study using transcranial magnetic stimulation showed that the disruption of prefrontal cortex (PFC) function results in a specific deficit in model-based behavior, pointing to the important role of the PFC in strategic decision making (Smittenaar, FitzGerald, Romei, Wright, & Dolan, 2013). One explanation for the ambiguity in the current results might be that the temporal resolution of fMRI is not sufficient to uncover separable neural mechanisms underlying model-based and model-free decision making.

In this study, we sought to shed light on this issue by taking advantage of the high temporal resolution of electroencephalography (EEG) in combination with a two-stage Markov decision task (Daw et al., 2011) and computational (RL) modeling. On the basis of the previous literature, we focused on three components of the event-related potential (ERP) that have been associated with decision processes. To study model-based decision processes, we looked into the stimulus-locked P300 component. This component is a parietally oriented positivity that occurs around 300–600 ms after stimulus onset. The P300 is typically observed in experimental situations in which an internal world model has to be updated on the basis of task-relevant information (Donchin & Coles, 1988; Nieuwenhuis, Aston-Jones, & Cohen, 2005). Thus, the P300 might reflect the updating of model-based representations (Cavanagh, 2015; Gläscher, Daw, Dayan, & O'Doherty, 2010). To investigate outcome-related decision processes, we focused on the feedback-related negativity (FRN) and the feedback-related positivity (FRP). Both of these components have been linked to reward processing. The FRN is elicited by negative outcomes and occurs between 200 and 300 ms after feedback onset (Miltner, Braun, & Coles, 1997). The results of previous studies suggest that the FRN is sensitive to negative prediction errors during RL (Holroyd & Coles, 2002; Nieuwenhuis et al., 2002; Walsh & Anderson, 2012). However, more recent work has indicated that the FRN may signal surprise (unsigned prediction errors; Cavanagh, Figueroa, Cohen, & Frank, 2012; Talmi, Atkinson, & El-Deredy, 2013). The FRP is a positive-going deflection in the same time window (200–300 ms) that seems to reflect learning and is sensitive to positive reward prediction errors (Arbel, Goforth, & Donchin, 2013; M. X.Cohen, Elger, & Ranganath, 2007; Eppinger, Kray, Mock, & Mecklinger, 2008; Eppinger, Mock, & Kray, 2009).

To examine the neural dynamics of the interplay of model-free and model-based decision mechanisms, we used a two-stage Markov decision task in combination with ERPs and computational modeling. To manipulate the demands on model-based decision making, we applied two experimental conditions with different probabilities of transitioning from the first to the second stage of the task. In the 60–40 condition, the transition probabilities between task stages were more difficult to tell apart (60% common, 40% rare transitions), and participants should be less able to predict the upcoming state. In the 80–20 condition, the transition probabilities were more differentiated, which should make it easier for individuals to anticipate the state that they were transitioning to. Based on the idea that the P300 component reflects the updating of model-based state predictions (Donchin, 1981; Donchin & Coles, 1988; Nieuwenhuis et al., 2005), we predicted that the component should reflect the transition probability structure. That is, the P300 at the second stage should be more differentiated for the 80–20 than for the 60–40 condition. In contrast, the prediction error information during outcome processing should be reflected in the feedback-related ERP components. In line with previous findings, we expected the FRN amplitude to increase with the magnitude of negative reward prediction errors, whereas we predicted that the FRP would reflect positive prediction errors (M. X.Cohen et al., 2007; Eppinger et al., 2008; Holroyd & Coles, 2002; Holroyd, Nieuwenhuis, Yeung, & Cohen, 2003; Nieuwenhuis et al., 2002).

Method

Participants

Twenty-one healthy young adults (mean age = 24.1, *SD* = 3.58; 11 male, 10 female) participated in the study. All participants gave informed written consent prior to participation. The ethics committee of the Max Planck Institute for Human Development approved the study. Participants received a minimum payment of €21. An additional amount (bonus) of up to €7 was paid, depending on the amount of reward that participants earned in the task.

Stimuli

We generated 24 colored figures ("GoGos") using free software (available online on the www.gogos-crazybones.com website) as the stimuli. For presentation purposes, the stimuli were further processed in Photoshop. Stimulus ratings, performed in pilot studies, had revealed no significant differences in dominance, valence, complexity, and recognizability. Each learning block involved a new set

of six stimuli, to avoid carry over effects. All stimuli were randomly assigned to task conditions for each participant.

Task

Participants performed a modified version of the two-stage Markov decision task developed by Daw and colleagues (2011; see also Eppinger, Walter, Heekeren, & Li, 2013, for a detailed description). The task involves two decision stages (see Fig. 1). At the second stage of the task, participants have to learn which of four stimuli is associated with the highest reward probability. The reward probabilities fluctuate over time (see the lower part of Fig. 1a). Thus, participants have to constantly update the expected reward values of these stimuli to perform optimally. To get to the second stage of the task, participants have to choose one of two options at the first stage (one of the figures with a yellow background color in the online version of Fig. 1a). In the illustrative example in Fig. 1a, the left option at the first stage is associated with a higher probability of transitioning (60%/80%) to the lower left options at the second stage (blue background color), and a lower probability (40%/20%) of transitioning to the lower right options at the second stage (brown background color). The reverse is true for the right stimulus at the first stage.

Each experimental session consisted of four blocks, which were separated by breaks. One block consisted of 116 trials of the two-stage Markov decision task. Each transition probability condition was completed twice. The conditions alternated within participants and were counterbalanced across participants. Before each block, participants were fully informed about the transition probability conditions and received a cue regarding the actual transition probabilities.

To support understanding of the task, we applied a cover story. The cover story concerned a businessman who had to decide between two airline carriers (represented by the figures with yellow background colors in online Fig. 1a), each of which would bring him to one of two islands. The airlines were somewhat unreliable with respect to their destinations (80/20 and/60/40 transition structures). At each of the islands, the businessman could trade with one of two populations of inhabitants (represented online by the figures on blue and brown background colors). The productivities (reward probabilities) of the populations changed across time. The task of the businessman was to make as much money as possible by tracking information about the reward probabilities of the options at the second stage and the transition structure at the first stage. Each of the four conditions involved a new set of stimuli.

Trial procedure

The trial procedure (see Fig. 1b) started with a fixation period (500 ms). After fixation, the first-stage choice options were displayed randomly on either the left or the right side of the screen. Participants had to indicate their choice within 2 s of stimulus presentation by using the "f" (left) or the "j" (right) key on a standard computer keyboard. If no response occurred within 2 s, the trial was aborted. The first-stage stimuli were followed by a fixation period (500 ms). At the second stage, we presented two colored squares for 1 s. The colors of these

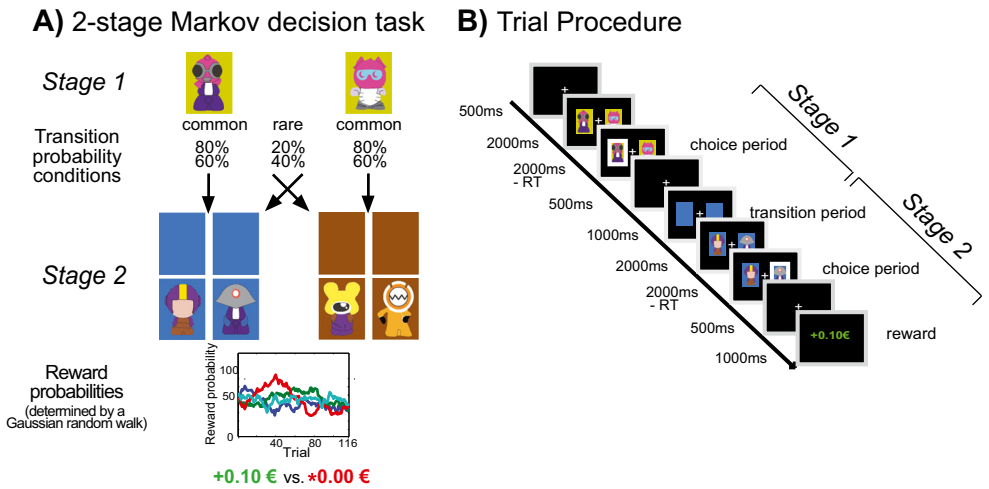

Fig. 1 **a** Schematic figure of the two-stage Markov decision task. In this task, participants have to constantly update reward predictions at the second stage (model-free decision making) and use these reward predictions to make goal-directed (model-based) decisions at the first stage. To manipulate the degree of model-based decision making, we applied two different versions of the task: In the 60–40 condition, the transition probabilities were difficult to differentiate (60% common, 40% rare transitions). Therefore, participants should be uncertain regarding their predictions of the upcoming state. In the 80–20 condition, the transition probabilities were easier to tell apart (80% common, 20% rare transitions). Thus, predictions regarding these state transitions should be more differentiated. Reward probabilities for the four decision options at the second stage of the task changed slowly and independently, according to Gaussian random walks. **b** Trial procedure of the two-stage task

squares indicated the state that participants had transitioned to (we will refer to this time period as the *state transition phase*). After 1 s, the corresponding stimuli (GoGo figures) were presented on top of the colored squares for 2 s (see Fig. 1b). Participants had to make their decision between the two stimuli (GoGo figures) within 2 s by using the same keys as in the first decision ("f" and "j"). If no response occurred within 2 s, three white question marks appeared on the screen for 1 s, and the trial was aborted (<1% of the trials across all participants). Below, we will refer to this stage as the *choice period*. This period was followed by a fixation phase (500 ms). The choices were either rewarded (10 € cents) or not rewarded (0 € cents). The probability of getting a reward was determined by a Gaussian random walk with a standard deviation of .025 and reflecting boundaries of .25 and .75 (Doll et al., 2012). The feedback stimuli were displayed for 1 s and were followed by another fixation period (500 ms). Overall, each trial lasted 7.5 s (see Fig. 1b).

Procedure

During preparation for the electroencephalogram (EEG), participants completed a demographic questionnaire and the BIS/BAS personality questionnaire (Carver & White, 1994). Prior to the experimental task participants completed a computerized training session, which was supervised by a research assistant. In the first part of the training, participants were introduced to the reward probability structure of the second stage of the task. To familiarize participants with the probabilistic reward structure, they had to perform ten choices between options with a fixed reward probability of 60%. To support their understanding of probabilistic information, we always referred to the reward probabilities in terms of absolute numbers (i.e., receiving a reward in approximately six of ten cases). Thereafter, participants were given ten additional trials, in which they had to find the option with the highest reward probability (out of two choice options). Upon their successful completion of this part, the research assistant explained that the reward probabilities would change slowly across the experiment. For illustration purposes, two examples of the random walks were displayed (see Fig. 1). In the next training phase, participants were introduced to the transition probabilities connecting the first and the second stage of the task. That is, we informed them about the fact that there were common and rare transitions and showed them a graphical illustration of the transition structure (similar to Fig. 1). Then participants performed ten trials in which they practiced the transitioning from the first-stage to the second-stage options. At the end of the practice session, participants conducted 30 trials of the experimental task (involving all stages as well as the probabilistic rewards) using a different stimulus set of "GoGo" figures. Thereafter, participants were placed ~1 m in front of a computer (CRT) screen in an electrically shielded room to perform the four blocks of the two-stage Markov decision task.

Data analysis

Behavioral data were analyzed using SPSS (SPSS Inc., Chicago, IL) and R (R Development Core Team, 2010). The RL model was implemented and fitted in MATLAB (The Mathworks Inc., Natick, MA). EEG data were processed using BrainVision Analyzer 2 (Brain Products GmbH).

Behavioral data

We defined stay–switch behavior as the probability to repeat a choice at the first stage as a function of the transition (common or rare) and the outcome (reward, no reward) on the previous trial. The mean stay probabilities were analyzed using mixed-effects logistic regression, as implemented in the lme4 package (Bates, Maechler, Bolker, & Walker, 2013) in R (R Development Core Team, 2010). The analysis involved the within-participants factors Transition Probability Condition (80–20, 60–40), Previous Transition Type (common, rare), and Previous Outcome (rewarded, no reward).

On the basis of our model simulations and the results of previous studies (Daw et al., 2011; Deserno et al., 2015; Eppinger et al., 2013), we expected that a pure model-free decision strategy would be reflected in a main effect of reward (Fig. 2a). That is, participants should stick to options that had been rewarded previously and should switch away from options that had been punished on the previous trial. Model-based decision making should be reflected in an interaction between the factors Transition on the Previous Trial and Reward on the Previous Trial (Fig. 2a).

For our follow-up analyses, we calculated model-based (mb) and model-free (mf) difference values based on stay probabilities (see Fig. 2a).

- Model-based (mb) difference values were calculated as:

$$\begin{aligned}\Pr(\mathrm{mb}) = &[\text{common rewarded} + \text{rare unrewarded}] \\ &-[\text{common unrewarded} + \text{rare rewarded}]\end{aligned}$$

- Model-free (mf) difference values were calculated as:

$$\begin{aligned}\Pr(\mathrm{mf}) = &[\text{common rewarded} + \text{rare rewarded}] \\ &-[\text{common unrewarded} + \text{rare unrewarded}]\end{aligned}$$

We then compared the model-based and model-free difference values, using repeated measures analyses of variances (ANOVAs) with the factor Transition Probability (80–20 vs. 60–40) (Fig. 2c).

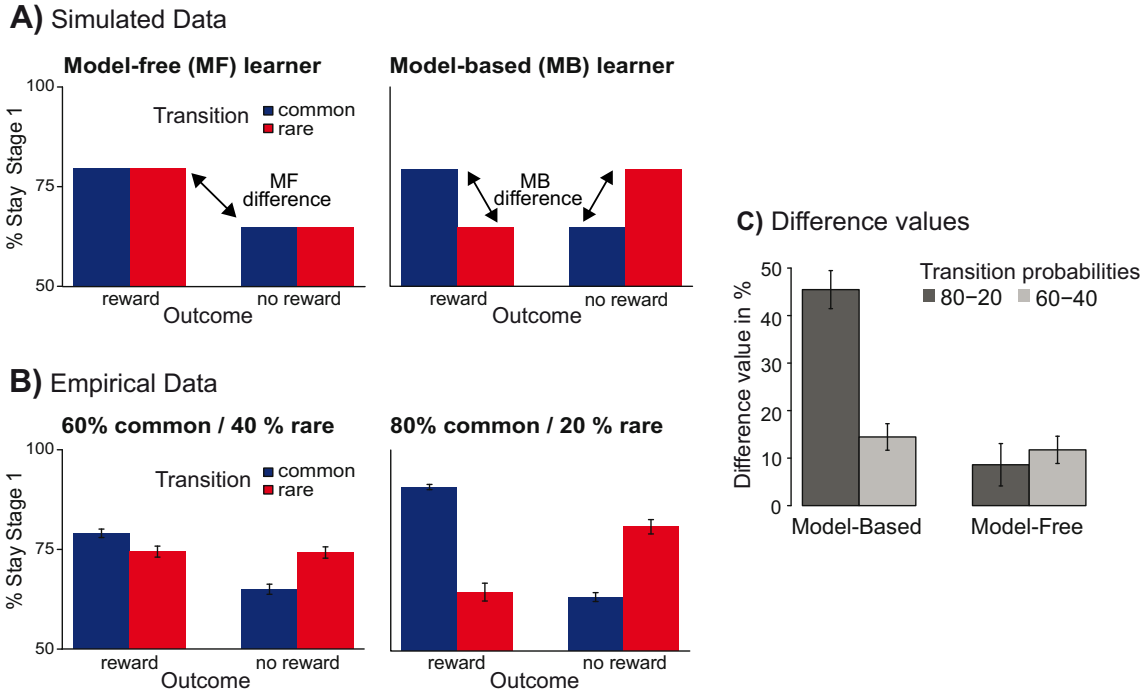

Fig. 2 Figure 2a): Model predictions. Left panel: Simulations show that model-free decision-making isreflected in a main effect of reward. That is, the probability of repeating the same first stage choice (staybehavior) depends on whether the choice on the previous trial was rewarded or not. Right panel: Modelbasedbehavior is reflected in an interaction between transition on the previous trial and reward on theprevious trial. That is, model-based behavior takes reward information as well as knowledge of the transitionstructure into account.Figure 2b): Empirical data. Probability of repeating the same first stage choice (stay behavior) as a functionof the transition on the previous trials (common, rare transition) and the outcome received on the previoustrial (reward, no reward). Stay probabilities are displayed separately for the two transition probabilityconditions (60-40 vs. 80-20), error bars reflect the standard error of the mean (SEM).Figure 2c): Difference values (stay probability) for model-based behavior [common rewarded + rareunrewarded] – [rare rewarded + common unrewarded] and model-free behavior [common rewarded + rarerewarded] – [common unrewarded + rare unrewarded]. Difference values are displayed separately for thetwo transition probability conditions (60-40 vs. 80-20), error bars reflect the standard error of the mean(SEM).

Computational modeling

As had been described in prior studies (Daw et al., 2011; Eppinger et al., 2013; Wunderlich, Smittenaar, & Dolan, 2012), we fitted each participant's choice behavior using a hybrid RL algorithm. The model acquired state–action values via separate model-free and model-based decision-making algorithms. Both Q values were weighted by the parameter omega (Ω), to compute the overall state–action value of the first-stage options. The "model-basedness" parameter Ω was held constant across trials and constrained to range between 0 and 1. If omega approached 0, the behavior was mostly model-free (primarily driven by reward). In contrast, an omega of close to 1 would indicate mostly model-based choice behavior—that is, choices reflecting an interaction of transition structure and reward. We assumed that participants would select actions according to a softmax function. The choice probabilities were determined by the state–action values.

For the model fitting, we estimated the free parameters of the hybrid model for each block and participant individually via maximum likelihood. We first iterated all parameters individually by using grid search to get a rough estimate. Subsequently, we extracted the 12 best-fitting parameter combinations in both transition probability conditions and entered them as starting values for precise parameter estimation, using the MATLAB routine fMincon. The parameters were held constant across trials but were allowed to vary across participants and between blocks within participants. We used the individually estimated parameters of the model to compute reward prediction errors (RPEs) and Q values at the second stage for each trial. For the ERP analyses, the RPEs were split between valences (positive, negative). Furthermore, prediction errors, Q_{net} values at the first stage, and Q values for the chosen option at the second stage were split between magnitudes (low, medium, high). We defined the RPe magnitude factor as the 33rd, 66th, and 100th percentiles of the respective range (within each individual). This procedure guaranteed comparable values for each category of RPe magnitude and similar amounts of trials per category (see Fig. 4c in the Results). We acknowledge that this split between small, medium, and large prediction errors (Q values) was somewhat arbitrary. Future studies should try single-trial regression approaches

(Fischer & Ullsperger, 2013) to uncover the neural dynamics of learning-related changes in prediction error signaling.

Description of the computational model

The task consisted of two stages and three states (first stage, S_1 –State A; second stage, S_2 –States B & C; see Fig. 1). Each state was associated with two actions (a_A, a_B). At both stages (i), a state–action value function $Q_{Si}(a)$ was learned that mapped each state–action pair to its expected value. We refer to the model-based value function at the first stage as $Q_{MB\ |\ S1}$, and to the model-free value function as $Q_{MF\ |\ Si}$.

Model-free state–action values Model-free state–action values at the second stage were updated using SARSA(λ) temporal difference learning (Rummery & Niranjan, 1994). The state–action pairs were updated in each trial t according to the equation

$$Q_{MF|S2}(a,t+1) = Q_{MF|S2}(a,t) + \alpha_2\left[r(t) - Q_{MF|S2}(a,t)\right],$$

where α_i is the learning rate at a given stage (here, 2), and $r(t)$ is the reward received in that trial.

The state–action value and the reward at the second stage were then used to update the model-free values at the *first stage*. This updating mechanism followed the same temporal-difference learning rule, with an additional parameter λ, allowing for eligibility traces:

$$\begin{aligned} Q_{MF|S1}(a,t+1) = {} & Q_{MF|S1}(a,t) \\ & + \alpha_1\left[Q_{MF|S2}(a_{\text{chosen}},t) - Q_{MF|S1}(a,t)\right] \\ & + \alpha_1\lambda\left[r(t) - Q_{MF|S2}(a,t)\right]. \end{aligned}$$

Eligibility traces are not assumed to carry over from trial to trial because the task structure involved constantly changing reward probabilities (determined by the random walks) for each option.

Model-based state–action values Model-based state–action values are computed for each trial using Bellman's equation (Sutton & Barto, 1998) by taking the model-free state–action values from the second stage and the transition probabilities into account.

$$\begin{aligned} Q_{MB|S1}(a_1) &= \text{HighTran}*\max\left[Q_{MF|S2}(a)\right] + \text{LowTran}*\max\left[Q_{MF|S2}(a)\right], \\ Q_{MB|S1}(a_2) &= \text{LowTran}*\max\left[Q_{MF|S2}(a)\right] + \text{HighTran}*\max\left[Q_{MF|S2}(a)\right]. \end{aligned}$$

In this equation, HighTran is defined as the higher transition probability of the current condition (either .8 or .6) and LowTran is defined as the lower transition probability of that condition (either .2 or .4). Participants were explicitly instructed about the nature of the transition probabilities prior to each block.

Finally, in the full hybrid model the Q_{Net} state–action value was calculated as the weighted sum of the model-based and model-free values:

$$Q_{\text{Net}|S1} = \Omega * Q_{MB|S1}(a) + (1-\Omega) * Q_{MF|S1}(a),$$

where Ω is the weighting parameter. At the second stage, the Q_{Net} state–action value is equal to the model-free state–action value ($Q_{\text{Net}\ |\ S2} = Q_{MF\ |\ S2}$).

Softmax rule Choice probabilities at each stage were calculated according to a softmax rule:

$$P_{S1}(a_1,t) = \frac{\exp\left(\beta_1 * \left[Q_{\text{Net}|S1}(a_1,t) + \pi * rep(a_1)\right]\right)}{\left(\exp\left(\beta_1 * \left[Q_{\text{Net}|S1}(a_1,t) + \pi * rep(a_1)\right]\right)\right) + \left(\exp\left(\beta_1 * \left[Q_{\text{Net}|S1}(a_2,t) + \pi * rep(a_2)\right]\right)\right)},$$

where β_i is the inverse softmax temperature parameter, controlling the distinctiveness of the choices. We allowed both learning parameters (α_1, α_2) and the softmax temperature parameters (β_1, β_2) to differ between both stages. The indicator function rep(a) is defined as 1 if a is a top-stage action and is the same as was chosen on the previous trial and zero otherwise. Taken together, the function rep(a) and the parameter π captures the degree of perseveration ($\pi > 0$) or switching ($\pi < 0$) at the first-stage options (Lau & Glimcher, 2005).

Choice probabilities were calculated similarly as:

$$P_{S2}(a_1,t) = \frac{\exp\left[\beta_2 * Q_{\text{Net}|S2}(a_1,t)\right]}{\exp\left(\beta_2 * Q_{\text{Net}|S2}(a_1,t)\right) + \exp\left(\beta_2 * Q_{\text{Net}|S2}(a_2,t)\right)}$$

Taken together, the model contained seven parameters (α_1, α_2, β_1, β_2, π, λ, Ω), with $\Omega = 0$ indicating pure model-free learning and $\Omega = 1$ indicating pure model-based decision making.

EEG recordings and ERP analysis

EEG and electrooculography (EOG) were recorded continuously from 64 passive Ag/AgCl electrodes embedded in an elastic plastic cap, using BrainVision Recorder (Brain Products GmbH, Gilching, Germany). The recording locations were based on the international 10–10 system; recording electrodes were referenced online to the right mastoid, and re-referenced offline to the average of the left and right mastoids. The EEG signals were filtered with a band-pass filter in the range of 0.01 and 100 Hz and were digitized with a sampling rate of 1000 Hz. The ground electrode was placed above the forehead. Vertical and horizontal EOGs were recorded next to each eye and below the left eye. Electrode impedances were kept below 5 kΩ.

For the statistical analyses, the EEG data were low-pass filtered at 30 Hz using an Infinite Impulse Response filter. EEG data were epoched (–200 to 600 ms) and averaged dependent on stage, transition, and obtained reward. The epochs were baseline-corrected by subtracting the average of the 200-ms prestimulus activity. All epochs were time-locked to the onset of the stimulus. Vertical and horizontal eye movements were corrected using a regression approach (Gratton, Coles, & Donchin, 1983). Trials containing remaining ocular, or other artifacts, were rejected using a threshold criterion (200-mV difference, 30-μV gradient). To avoid differences in trial numbers of the ERP averages, the trial numbers in the frequent conditions (80%, 60%, 40%) were adjusted to the trial numbers in the lowest-frequency condition (20%). In this procedure, 21 trials were randomly chosen from each transition probability condition and used to calculate the individual participant ERP averages at feedback onset. The averages at the onset of the second-stage stimuli contained a minimum number of 44 trials per condition.

Minimum trial numbers for the FRN averages were determined on the basis of previous work, suggesting that the number of trials (21 in the present study) was sufficient to obtain reliable FRN amplitudes in younger adults (Marco-Pallares, Cucurell, Münte, Strien, & Rodriguez-Fornells, 2011). For stimulus-evoked P300 components, previous studies had recommended trial numbers around 30–60 trials (Cohen & Polich, 1997; Luck, 2005). Thus, with 44 trials per condition, we were well within this range. Nevertheless, future work should consider the use of bootstrapping methods to improve sensitivity and reliability of the EEG data.

The P300 component at the second stage was measured as the mean amplitude in the 330- to 430-ms time window after stimulus onset. The FRP was measured as the mean amplitude in the 280- to 380-ms time window after feedback onset. The time windows were placed at the peaks of the components, which were determined by visual inspection of the grand averages. The EEG signals at the different electrode sites were averaged into six topographical regions of interest: left anterior (F7, F5, F3), middle anterior (F1, Fz, F2), right anterior (F4, F6, F8), left central (T7, C5, C3), middle central (C1, Cz, C2), right central (C4, C6, T8), left posterior (P7, P5, P3), middle posterior (P1, Pz, P2), and right posterior (P4, P6, P8). The 9 ROIs were then pooled in two experimental factors Anterior-Posterior and Hemisphere, each involving three levels (anterior, central and posterior as well as left, medial and right, respectively). To examine the mean P300 amplitudes, we applied a repeated measures ANOVA involving the within-participant factors Transition Probability Condition (80–20, 60–40), Transition Type (common, rare), Anterior–Posterior (anterior, central, posterior), and Hemisphere (left, medial, right). For the analysis of the FRP component, the ANOVA design involved the factor Outcome (reward, no reward) as an additional predictor. For the model-based ERP analyses, we used ANOVA designs involving the factors Transition Probability Condition (80–20, 60–40), Q Value/RPE Magnitude (low, medium, high), Anterior–Posterior (anterior, central, posterior), and Hemisphere (left, medial, right).

The FRN was defined using peak-to-peak measures at electrode FCz (Frank, Woroch, & Curran, 2005; Yeung & Sanfey, 2004). To calculate the peak-to-peak measures, we first applied a low-pass filter of 15 Hz. Subsequently, a semiautomatic algorithm was used to identify the maximum positive peak in a time window of 130–240 ms after feedback onset. From that latency, the most negative peak until 325 ms after stimulus onset was identified. The amplitude of the component was defined as the difference between the two peak measures. For our statistical analyses, we applied a repeated measures ANOVA involving the within-participants factors Transition Probability Condition (80–20, 60–40) and Reward (rewarded, unrewarded). For the model-based analyses of the FRN, a repeated measures ANOVA with the factors Prediction Error Valence (positive, negative), Magnitude (low, medium, high), and Transition Probability Condition (80–20, 60–40) was conducted.

Bonferroni corrections were applied when necessary, and corrected p values are reported throughout (p level < .05). Whenever necessary, the Greenhouse–Geisser correction (Geisser & Greenhouse, 1958) was applied. The original F value, the adjusted p values, and the epsilon values (ε) are given. Effect sizes (eta-squared, η^2) are provided where applicable (Cohen, 1973).

Correlation analysis

We investigated the relationships between the behavioral, electrophysiological, and computational model parameters across individuals by calculating Pearson correlation coefficients across both transition probability conditions. As behavioral measures, we entered the model-based and model-free difference scores (see the Method section and Fig. 2a) in the correlation analysis. To test for relationships with parameters from the computational model, we used the fitted parameters for each individual. Furthermore, we calculated amplitude difference values for the

second-stage P300 transition effects (P300-Tran: Common – Rare), as well as the second-stage P300 expected value effects (P300 Q_{Val}: High Q_{Value} – Low Q_{Value}), and entered them into the correlation analysis. Finally, we used each individual's fitted model parameters to calculate the optimal action on each trial, and calculated the probability with which each participant chose this option. This procedure resulted in a value that indicated the likelihood that a participant would choose the best option on a given trial.

Results

Choice behavior

As in previous studies (Daw et al., 2011; Eppinger et al., 2013), choice behavior at the first stage (proportion stay trials) was analyzed depending on whether the choice on the previous trial had been rewarded and whether the transition on the previous trial had been common (60%/80%) or rare (40%/20%). Given that the first-stage choice proportions were binomial (stay = 1, switch = 0), we used a mixed-effects logistic regression (see Table 1 and the Method section; Daw et al., 2011; Eppinger et al., 2013). This analysis revealed a significant main effect of reward ($p < .001$), showing that participants stayed with options that had previously been rewarded and switched away from options had not been rewarded on the previous trial (reward-based, model-free decision making; see Fig. 2a and b). A significant interaction between reward and transition type ($p < .001$) indicated that in addition to rewards, the participants also took the transition structure into account (model-based decision making; see Fig. 2a and b). Most interestingly, the analysis also revealed a significant three-way interaction between transition probability condition, transition type, and reward ($p < .001$). To follow up on the three-way interaction, we performed separate analyses for each of the transition probability conditions. For the 80–20 condition, these analyses revealed significant main effects of transition type and reward ($ps < .001$), as well as a significant interaction between the two factors ($p < .001$). For the 60–40 condition, the main effect of transition type was not significant ($p = .26$). However, we did find a significant main effect of reward and a significant interaction between transition type and reward ($ps < .001$). Together, these findings suggest that even though participants were less able to differentiate between common and rare transitions in the 60–40 condition, there was still evidence for a significant model-based contribution to their decision making in this condition. To directly test for differences in model-based behavior between the transition probability conditions, we calculated model-based and model-free difference values (for more details, see Fig. 2a and the Method section). As is shown in Fig. 2b and c, these analyses revealed enhanced model-based behavior in the 80–20 relative to the 60–40 condition ($p < .001$, $\eta^2 = .66$). No effect of transition probability was observed for model-free behavior ($p = .51$).

Computational modeling results

To analyze differences in the model parameters between the two transition probability conditions, we used paired *t* tests. The analysis revealed no difference in the weighting (model-basedness) parameter **Ω** ($t = 0.09$, $p = .92$). However, it did reveal a significant main effect of transition probability condition on the inverse temperature parameter (β_1) at the first stage ($t = 3.37$, $p < .01$). As is shown in Table 2, β_1 was lower for the high (80–20) than for the low (60–40) transition probability condition. This finding suggests that in the high transition probability condition it was easier for participants to distinguish the first-stage choice options. In addition, the learning rate at the second stage (α_2) was significantly higher for the 80–20 than for the 60–40 condition, indicating that the participants gave more weight to recent reward outcomes in the high transition probability condition.

Stimulus-locked ERPs at the first stage The analysis of the stimulus-locked P300 at the first stage revealed no significant main effects or interactions involving the factor Transition Probability Condition, Transition Type, or Reward ($ps > .05$). Furthermore, an additional analysis of the previous

Table 1 Estimates of the logistic regression analysis

Predictor	Estimate	*p* Value
(Intercept)	1.18	<.001
Condition	0.09	<.001
Transition	0.09	<.01
Outcome	0.19	<.001
Condition × Transition	0.11	<.001
Condition × Outcome	0.03	>.05
Transition × Outcome	0.43	<.001
Condition × Transition × Outcome	0.24	<.001

Table 2 Optimal model parameters in each condition

	α_1	α_2	β_1	β_2	λ	π	**Ω**	-LL
60–40 Transition Probability Condition								
25th percentile	0.25	0.24	3.81	3.04	0.01	0.04	0.42	106.19
Median	0.41	0.42	8.61	4.10	0.10	0.09	0.66	130.53
75th percentile	0.73	0.65	14.68	5.42	0.28	0.19	0.84	145.81
80–20 Transition Probability Condition								
25th percentile	0.11	0.46	4.09	3.39	0.01	0.07	0.39	107.67
Median	0.46	0.60	5.94	4.43	0.36	0.13	0.58	114.28
75th percentile	0.76	0.76	7.79	6.49	0.79	0.24	0.76	120.24

transitions and rewards on the first-stage ERPs (in analogy to the analysis of stay–switch behavior) did not reveal significant main effects of previous reward or transition or the expected interaction between previous transition and reward (*p*s > .05). In a further analysis, we used the individually estimated parameters of the hybrid RL model to calculate state–action values at the first stage (Q_{Net} values). The calculated state–action values were then split into low ($Q_{Net} \leq$ 33rd percentile), medium (33rd < Q_{Net} < 66rd percentile), and high ($Q_{Net} \geq$ 66rd percentile) magnitudes, and used to inform the stimulus-locked ERP analysis at the first stage. Again, as is shown in Supplementary Fig. 3, we did not find significant main effects of condition or magnitude, or a significant interaction of condition and magnitude (*p*s > .05).

Stimulus-locked ERPs at the second stage: Transition phase An analysis of the ERPs in the state transition period at the second stage (when the background colors were presented; see Fig. 1b) revealed main effects of probability condition in the N200 and the P300 components, *F*s(1, 20) > 5.31, *p*s < .03, η_g^2s > .03. As is shown in Supplementary Fig. 4, these modulations seem to reflect a general amplitude shift in the ERPs in the high-demand (60–40) as compared to the low-demand (80–20) transition probability condition. Interestingly, we also found evidence for a main effect of transition type, $F(1, 20) = 15.60$, $p < .001$, $\eta_g^2 = .04$, as well as a significant interaction between probability condition and transition type in the late time window of the ERP, $F(1, 20) = 4.73$, $p = .04$, $\eta_g^2 = .01$. Consistent with previous reports, we refer to this component as the *late positive*

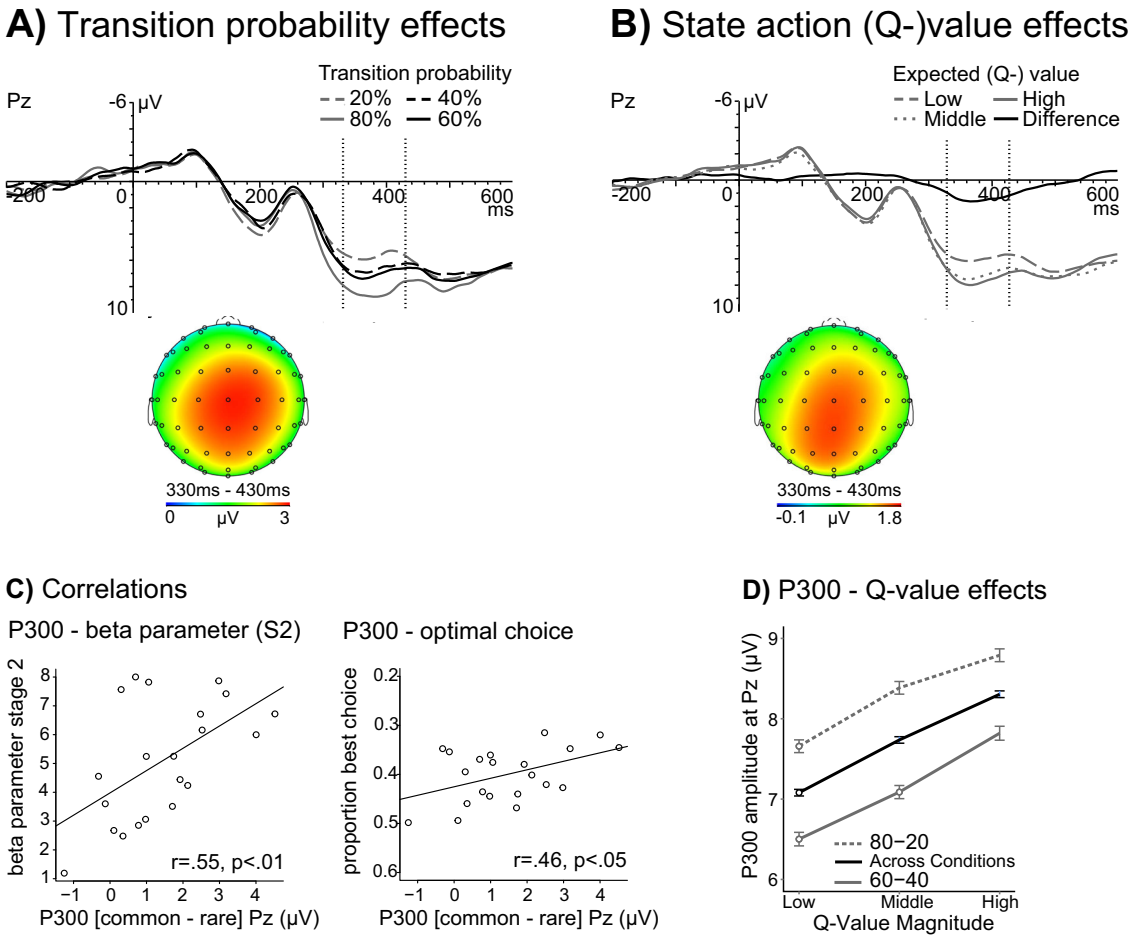

Fig. 3 **a** *Top*: ERPs elicited by second-stage stimuli at electrode Pz, displayed separately for the 80–20 condition (gray) and the 60–40 condition (black), as well as the common transitions (solid lines) and rare transitions (dashed lines). Dashed vertical lines indicate the time window that was used for the analysis as well as for generation of the topographic maps. *Bottom*: The topographic map displays the difference between common and rare transitions in the 80–20 condition in a time window of 330–430 ms. **b** *Top*: ERPs elicited by second-stage stimuli at electrode Pz, displayed separately for low (gray dashed), middle (gray dotted), and high (gray solid) state action (Q) values. Black solid lines reflect the difference between ERPs to the high and low Q values. Dashed vertical lines indicate the time window that was used for the statistical analysis as well as for generation of the topographic maps. *Bottom*: The topographic map displays the difference between high and low Q values in a time window of 330–430 ms. **c** Scatterplot of correlations between the inverse temperature parameter at the second stage (β_2) and the P300 amplitude at Pz (averaged across conditions). **d** Mean P300 amplitudes as a function of Q-value magnitude, displayed separately for the 80–20 condition (dashed gray line) and the 60–40 condition (solid gray line), as well as averaged across conditions (black solid line). Error bars reflect the standard errors of the means (*SEM*s)

complex (LPC) (S. Sutton & Ruchkin, 1984). Previous work suggested that the LPC is sensitive to physical expectancy deviations during language processing (Kutas & Hillyard, 1980). However, some more recent evidence has suggested that the LPC might reflect the volatility of decision rules during learning (Bland & Schaefer, 2011). As is shown in Supplementary Fig. 4, the interaction effect reflects a greater LPC for rare than for common transitions in the 80–20 condition ($p < .001$, $\eta_g^2 = .09$), but no such effect in the 60–40 condition ($p = .12$, $\eta_g^2 = .01$). These results suggest that, on the basis of the background colors, the participants were able to differentiate between common and rare transitions in the low-demand (80–20) condition, but not in the high-demand (60–40) condition. To make sure that the baseline period for the subsequent ERPs in the choice phase was not confounded by condition differences in the transition phase, we analyzed the late time window in the state transition ERPs (see the shaded area in Supplementary Fig. 4). This analysis revealed no significant differences between conditions (see Supplementary Fig. 5).

Stimulus-locked ERPs at the second stage: Choice period The analysis of the stimulus-locked P300 at the second stage revealed a significant main effect of transition type, $F(1, 20) = 22.15$, $p < .001$, $\varepsilon = 1$, $\eta^2 = .53$, which reflected a greater P300 amplitude for common (80% and 60%) than for rare (20% and 40%) transitions. Moreover, we found a significant interaction between probability condition and transition type, $F(1, 20) = 6.48$, $p < .05$, $\varepsilon = 1$, $\eta^2 = .25$. Separate analyses for each of the probability conditions showed a greater P300 amplitude after common than after rare transitions in the 80–20 condition ($p < .001$, $\eta^2 = .56$). This was not the case for the 60–40 condition ($p = .23$). As is shown in Fig. 3a, the more model-based behavior in the 80–20 condition was associated with a greater P300 component for common than for rare transitions. No such effect was observed for the 60–40 condition.

To examine whether the P300 component reflected the expected value of choice options at the second stage, we used the individually estimated parameters of the hybrid RL model (the state–action Q values at the second stage) to inform the ERP analysis (see the Method section for details). This analysis showed a significant main effect of expected value magnitude, $F(2, 40) = 15.41$, $p < .001$, $\varepsilon = .98$, $\eta^2 = .44$. As is shown in Fig. 3b and d, the P300 amplitude increased with the magnitude of the state–action values. Furthermore, we obtained a main effect of transition probability condition, $F(1, 20) = 4.54$, $p < .05$, $\varepsilon = 1$, $\eta^2 = .19$. As is shown in Fig. 3d, the P300 amplitude was higher overall for the 80–20 than for the 60–40 condition. However, the analysis did not reveal a significant interaction between probability condition and expected value magnitude ($p = .52$), indicating that the expected-value effects in the P300 were independent of model-based (transition probability) effects.

Outcome-locked ERPs

To examine the signatures of model-free and model-based processes at the reward stage, we focused on the ERP components that have previously been shown to covary with prediction error information: the FRN elicited by negative outcomes, and the FRP, which occurs in response to rewarding outcomes.

Feedback-related negativity The analysis of FRN amplitudes revealed a significant main effect of reward, $F(1, 20) = 10.27$, $p < .01$, $\varepsilon = 1$, $\eta^2 = .34$, demonstrating higher amplitudes for no reward than for reward feedback, which was in line with previous studies of the FRN. As is shown in Fig. 4a the analysis did not show significant main effects or interactions involving the factor Probability Condition or Transition Type (ps > .13, η^2s < .10).

To investigate whether the FRN reflected RPEs at the second stage, we used the individually estimated RPE estimates of the hybrid RL model (averaged into three levels of magnitude) to inform the ERP analysis. This analysis showed a significant main effect of RPE valence, $F(1, 20) = 8.09$, $p = .01$, $\eta^2 = .29$, which reflected the greater FRN for negative than for positive RPEs. However, the analysis did not reveal a significant effect of RPE magnitude on the FRN ($p > .70$). Thus, the present results suggest that the FRN (as defined using peak-to-peak measures) is affected by the valence, but not by the magnitude, of prediction errors (see Fig. 4b and d).

Feedback-related positivity The FRP is a positive component in the time window of the FRN that is elicited by reward and that has been shown to reflect learning (Arbel et al., 2013; Cohen et al., 2007; Eppinger et al., 2008; Eppinger et al., 2009). As with the FRN, we examined whether the FRP was modulated by the transition probability condition as well as by the valence and magnitude of RPEs. The results of this analysis showed a significant main effect of RPE magnitude, $F(2, 40) = 5.00$, $p < 05$, $\varepsilon = .96$, $\eta^2 = .20$. Moreover, the analysis revealed a significant interaction between RPE valence and magnitude, $F(2, 40) = 9.68$, $p < .001$, $\varepsilon = .99$, $\eta^2 = .33$. Separate analyses for the factor RPE Valence showed a significant main effect of magnitude for positive RPEs, $F(2, 40) = 16.66$, $p < .001$, $\varepsilon = .87$, $\eta^2 = .45$ (see Fig. 4d). No magnitude effect was observed for negative RPEs ($p = .50$). Taken together, our results show that the FRP is sensitive to the magnitude of positive, but not of negative, RPEs.

ERP–behavioral correlations

To investigate relationships between the behavioral, electrophysiological, and computational model parameters across individuals, we calculated Pearson correlation coefficients. As is shown in Supplemental Table 1, we found

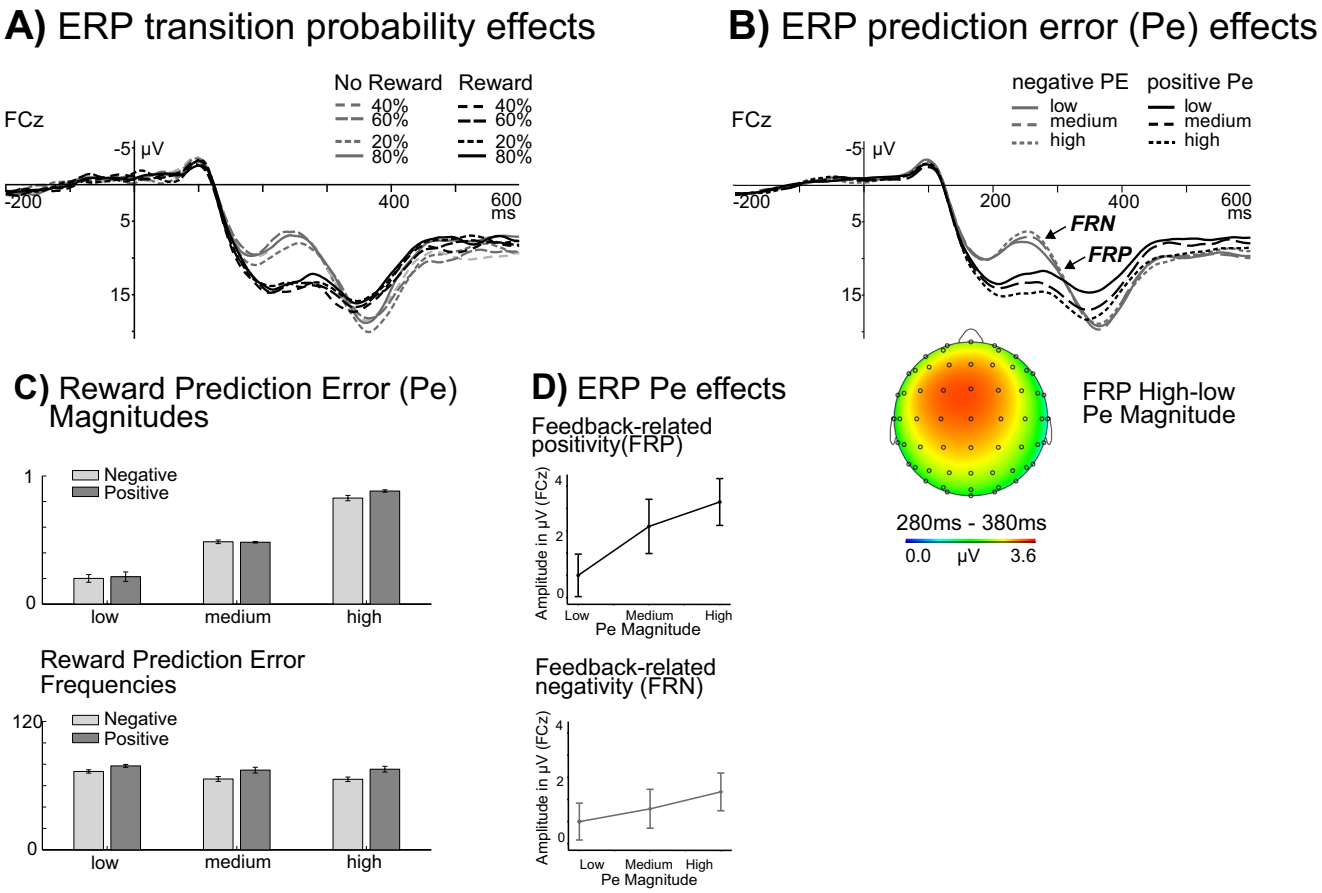

Fig. 4 **a** Feedback-locked ERPs at electrode FCz for rewards (black) and no rewards (gray), displayed separately for the two transition probability conditions (80–20 vs. 60–40). **b** *Top*: Feedback-locked ERPs at electrode FCz for negative (gray) and positive (black) reward prediction errors, displayed separately for low, medium, and high magnitudes of prediction error. *Bottom*: The topographic map displays the difference in the FRPs between high- and low-magnitude prediction errors in a time window of 280–380 ms. **c** Mean magnitudes and frequencies of positive and negative reward prediction errors, averaged across conditions in a time window of (280–380 ms). Error bars reflect the standard errors of the means (*SEM*s). **d** Amplitudes of the FRN and FRP at electrode FCz as function of prediction error magnitude (low, medium, high). Error bars reflect the standard errors of the means (*SEM*s)

a significant positive correlation between the P300 amplitude difference (common minus rare transitions) and the inverse temperature at the second stage ($r = .55$, $p < .01$). Moreover, we found that the P300 difference correlated positively with the probability of choosing the better option at that stage ($r = .46$, $p < .05$). These findings suggest that a greater ability to predict the upcoming state on the basis of knowledge of the transition structure (as reflected in the P300) is associated with more optimal choice behavior. Interestingly, we did not find significant correlations between the Q-value P300 effect and any of the behavioral measures (see Table 3), indicating that the present associations cannot be explained in terms of differences in the expected values of the choice options. Taken together, our results show that the P300 amplitude at the second stage reflects individual differences in the ability to integrate model-based knowledge of the transition structure with the decision values of the choice options, and that it predicts optimal performance.

Discussion

Current decision-making theories suggest that value-based decisions arise from the interaction of two distinct decision systems: (1) a model-free system that is involved in making habitual decisions based on past experience, and (2) a model-based system that subserves goal-directed decisions that are based on a cognitive model of the environment (Balleine & O'Doherty, 2010; Daw et al., 2005). So far, neuroimaging studies have been inconclusive as to whether the neural mechanisms underlying model-free and model-based decision processes can be dissociated (Daw et al., 2011; Deserno et al., 2015). In contrast to previous approaches, which had used functional imaging (fMRI), we took advantage of the high temporal resolution of ERPs in combination with computational (reinforcement-learning) modeling to examine the neural dynamics of model-free and model-based decision mechanisms.

Consistent with previous studies, we applied a two-stage Markov decision task that allowed for a dissociation of model-

free and model-based contributions to choice behavior (adapted from Daw et al., 2011; Eppinger et al., 2013; see Fig. 1). The idea of the task was to integrate model-free information about the reward as well as model-based information about the (Markovian) transition structure. A pure model-free learner learns an expected value for a decision option from his past experiences with that same option, but ignores the state transition structure of the task (see Fig. 2a). A model-based learner integrates model-free information about the values of the decision options with information about the transition structure of the task to inform choice behavior (see Fig. 2).[1] To investigate the dynamics in the interaction between these two decision systems, we manipulated the demands on model-based decision making by applying two different transition probability structures (see Fig. 1). In the high-demand structure (60–40 condition), the common and rare transitions were difficult to differentiate, and participants should have been uncertain about the upcoming stimulus (state). In the low-demand structure (80–20 condition), the transition probabilities were easier to tease apart, which should support model-based behavior.

Consistent with our prediction, the behavioral results showed that model-based decision making was sensitive to changes in the transition probabilities. As is shown in Fig. 2c, reduced demands on the representation of the transition structure in the 80–20 condition led to greater model-based behavior, as in the high-demand structure (the 60–40 condition). Contrary to previous findings, an analysis of the computational model parameters revealed no significant difference in the "model-basedness" parameter Ω between transition probability conditions (Daw et al., 2011; Wunderlich et al., 2012). Instead, the analysis showed a significantly lower inverse temperature parameter (β_1) at the first stage of the task for the high as compared to the low transition probability condition (see Table 2).

Taken together, the behavioral results suggest that the manipulation of the transition probability structure led to a more differentiated model-based choice pattern at the first stage of the task. At first sight, the absence of condition differences in the "model-basedness" parameter (Ω) seems surprising. However, when considering Fig. 2b, it becomes clear that in both conditions participants showed a Transition × Reward interaction (which has been interpreted as the hallmark of model-based behavior). In fact, our results in the high-demand (60–40) condition look very similar to the results in younger adults published in previous studies using a 70–30 transition matrix (Daw et al., 2011; Deserno et al., 2015; Wunderlich et al., 2012). What differs between the two transition probability conditions is the *degree* of model-based behavior. In the results of the model fitting, this greater differentiation in the model-based choice pattern was reflected in changes in the inverse temperature parameter of the softmax function at the first stage of the task. To verify whether we could replicate the empirical results using the computational model, we simulated the data by manipulating Ω while holding the temperature parameter at the first stage constant, and vice versa. The results of the simulations mimicked the empirical results, showing that increasing Ω leads to a shift from model-free to model-based behavior, whereas changing the temperature parameter makes the first-stage choice pattern more differentiated (see Supplementary Fig. 1). Furthermore, rerunning the model fitting with a fixed inverse temperature parameter at the first stage resulted in the expected differences between transition probability conditions in the Ω parameter (see Supplementary Fig. 2), This suggests that the β at the first stage was capturing variance that would otherwise be captured by the model-basedness parameter Ω.

Thus, consistent with the findings of the regression analysis, the results of the model fitting suggest that even in the high-demand (60–40) condition, younger participants showed evidence for model-based behavior. Making the state transitions more predictable in the low-demand (80–20) condition made the model-based choice pattern more distinct; that is, participants had a clearer representation of the upcoming options. Consistent with this interpretation, we found that the inverse temperature parameter at the first stage was negatively correlated with the probability of choosing the best option at the second stage (see Supplemental Table 1). This correlation suggests that the more differentiated the choice patterns at the first stage of the task, the higher the probability of choosing the option with the highest reward probability at the second stage of the task. Thus, the better participants were at predicting their transition to the second stage, the better they were at choosing the option with the highest expected value.

In summary, the present behavioral results suggest that the manipulation of the transition probability structure led to more distinct model-based choice behavior. The more participants were able to differentiate common and rare transitions at the first stage of the task, the better they were at predicting the upcoming stimulus, which boosted their performance at the second stage of the task.

An analysis of the ERPs at the state transition period showed a greater LPC for rare relative to common transitions in the low-demand (80–20) condition, but no such effect in the high-demand (60–40) condition. This finding suggests that with more differentiated transition probabilities, participants are able to predict the upcoming states. The reduced LPC component for common (80%) transitions reflects a higher expectedness of that transition and a reduced need for the updating of task-relevant information (Bland & Schaefer, 2011; Donchin, 1981; Donchin & Coles, 1988).

[1] It should be noted that the transition probabilities were instructed and trained in order to avoid learning effects. Moreover, the transition probabilities were explicitly cued at the beginning of each block to make participants aware of the condition they were in.

In line with our predictions, the stimulus-evoked P300 at the second stage reflected model-based decision processes—that is, processes that were sensitive to the expected value of the options as well as to the transition structure of the task. As is displayed in Fig. 3a the analysis showed a greater P300 amplitude after common than after rare transitions in the 80–20 condition (see Fig. 3a). No such effect was observed for the 60–40 condition. Thus, the P300 was enhanced in the condition in which participants had more differentiated (and reliable) predictions regarding the state transitions. This is in line with the suggestion that the choice-related P300 component may reflect state prediction errors (Gläscher et al., 2010). Importantly, as is shown in Fig. 3b, the same P300 component that was sensitive to the probability of state transitions was also sensitive to the expected values of the choice options. Moreover, the expected-value effect in the P300 was independent of the transition probability condition (see Fig. 3b and d). Taken together, these findings indicate that the P300 component may reflect the integration of model-based information about the transition structure of the task with model-free information about the expected values of the choice options during decision making. In support of this idea, we found positive correlations between the P300 amplitude difference between common and rare trials and both the inverse temperature parameter (β_2) and the probability of choosing the currently best option at the second stage. These findings indicate an association of the P300 effect with (a) the degree to which participants were able to differentiate the choice options at the second stage (the inverse temperature parameter β_2) and (b) the degree to which they were able to implement optimal choice behavior.

What remains unclear is the relationship between the LPC effects during the state transition period and the P3 effects at the second-stage choice period. In our interpretation of the data, the state transition effects reflect the predictability of the upcoming states, which is higher in the 80–20 than in the 60–40 condition. In contrast, the effects in the second-stage choice period seem to reflect the fact that participants are predicting the values of options and that these predictions are more differentiated in the high- than in the low-demand condition. Thus, it seems that the phase reversal of the two ERP components reflects the functional demands that are induced by the task.

Consistent with our results, findings from a recent fMRI study provided initial evidence for a key role of the ventromedial prefrontal cortex (vmPFC) in the integration of model-based and model-free value signals (Lee, Shimojo, & O'Doherty, 2014). Consistent with this idea, findings from a simultaneous EEG–fMRI study suggested that the P300 is generated by an attentional/arousal network involving the brain stem (presumably locus coeruleus, LC) as well as the anterior cingulate (ACC) and the ventromedial/orbitofrontal cortex (Walz et al., 2013). Together, these findings are consistent with the LC–P300 hypothesis, proposing that the P300 reflects the response of the locus coeruleus–norepinephrine (LC-NE) system to the outcome of decision making processes in the ACC and orbitofrontal cortex (Aston-Jones & Cohen, 2005; Nieuwenhuis et al., 2005). Taking these findings and theoretical ideas into account, it could be argued that the ventromedial/orbitofrontal cortex is involved in the integration of model-based and model-free value information (Wilson, Takahashi, Schoenbaum, & Niv, 2014). This integration process may trigger selective attention in parietal areas (as reflected in the P300 response) to facilitate optimal choice behavior. In support of this interpretation, recent evidence from single-unit cell recordings in monkeys showed a clear relationship between LC activity and the pupil response (Joshi, Li, Kalwani, & Gold, 2016), which has repeatedly been shown to be associated with P300 activity (Hong, Walz, & Sajda, 2014; Murphy, Robertson, Balsters, & O'Connell, 2011; Nieuwenhuis et al., 2005).

However, given of a lack in direct evidence for a relationship between LC activity and the P300 response during decision making, the LC–P300 account remains speculative. An alternative interpretation of the present results may be provided by the context-updating theory of the P300 (Donchin & Coles, 1988). According to this theory, the P300 reflects the updating of task-relevant information (such as model-based information about state transitions), which should be enhanced in the condition with more differentiated transition probabilities (80–20 condition). In contrast, if participants cannot make reliable predictions regarding the upcoming stimuli, such as in the 60–40 condition, it may be more difficult to update the task-relevant information at the second stage. That is, this theory can account for the observed transition probability effects in the P300, but it seems difficult to reconcile with the increase of P300 ERP activity as a function of expected value.

In a more exploratory analysis, we examined the relationship between the outcome-related ERPs and prediction error information. To investigate whether the FRN is related to the magnitude of RPEs (Holroyd & Coles, 2002), we used RPE estimates from the RL model to inform the ERP analysis. Consistent with the condition-based analyses, we observed an effect of prediction error valence. As is shown in Fig. 4b, the FRN amplitude was larger for negative than for positive RPEs. However, the FRN showed no clear relationship to the magnitude of negative RPEs (see Fig. 4b and d). These findings are in line with several previous studies, suggesting that the FRN may not be as tightly coupled to negative prediction errors as had previously been thought (Arbel et al., 2013; Eppinger et al., 2008; Eppinger et al., 2009; Hämmerer, Li, Mueller, & Lindenberger, 2011). In fact, several recent findings indicate that the FRN may be sensitive to surprise (unsigned prediction errors), rather than to signed prediction errors (Cavanagh & Frank, 2014). As such, the FRN may reflect a more general teaching signal generated in the ACC, which is not valence-specific (Botvinick, 2007; Johansen & Fields, 2014) but may be involved in the hierarchical organization of effortful behavior (Holroyd & McClure, 2015). This interpretation would be consistent with findings suggesting that FRN-like signals reflect surprise information in

the medial frontal cortex that triggers control adjustments (Cavanagh & Frank, 2014). It should also be noted that in the present study, negative RPEs reflected the absence of a reward, not a truly aversive outcome or punishment (such as an electric shock). Thus, it could still be that with truly aversive outcomes, the FRN might reflect the magnitude of negative prediction errors (Talmi et al., 2013; Talmi et al., 2012).

Most interestingly, in contrast to the FRN, which did not reflect RPE magnitude, we found evidence for an effect of RPE magnitude in the reward-related ERPs signals. As is shown in Fig. 4b, an FRP around 300 ms after feedback onset increased as a function of the magnitude of positive RPEs. These results are consistent with previous findings suggesting that the FRP might be sensitive to RPEs (Arbel et al., 2013; Cohen et al., 2007; Eppinger et al., 2008; Eppinger et al., 2009; Hämmerer et al., 2011; Herbert, Eppinger, & Kray, 2011). A recent study using single-trial frequency analyses revealed that the FRP reflected delta band activity that scaled with positive prediction errors (Cavanagh, 2015). However, this prediction error information in the feedback-related delta band response did not predict behavior. As in the present study, correlations with behavioral adaption effects were only found for stimulus-locked delta band (P300) activity. Consistent with previous neuroimaging work (Gläscher et al., 2010), these findings suggest that activity in the delta band may reflect hierarchically distinct types of prediction error information. The results of the present study support this idea by showing that stimulus-evoked P300 activity reflects violations in state predictions and is tightly coupled to behavior. In contrast, feedback-locked activity does seem to reflect RPEs but is not reflective of behavioral adjustments.

Counter to our initial expectation, we did not find evidence for model-based or model-free effects (or an interaction of such effects) on the ERPs at the first stage. At first sight, this result seems surprising, given that the behavioral dissociation is based on choice behavior at the first stage. Thus, one would expect to see correlates of these decision processes in the ERPs. Please note, however, that previous fMRI studies using the same two-state Markov task also did not show evidence for dissociable neural correlates of model-based or model-free decision processes at the first stage of the task (Daw et al., 2011; Deserno et al., 2015). We currently see two possible interpretations for the absence of these effects. On the one hand, the neural computations to update model-based information may not be confined to the first stage of the task, but rather may happen continuously throughout the task. Therefore, the behavioral measure may only reflect the application of a decision strategy that relies on the integration of information across the different stages of the paradigm. On the other hand, due to averaging, ERPs may not be sensitive enough to detect subtle fluctuations in choice behavior. That is, the neural mechanisms that lead to the behavioral dissociation at the first stage might be less tightly coupled to the eliciting stimulus than is the case at the second stage and reward delivery. Subsequent research should address this questions by using parametric analyses of single-trial EEG measures (Fischer & Ullsperger, 2013), or by using analysis techniques that are sensitive to induced rather than to evoked EEG signals (Herrmann, Munk, & Engel, 2004).

Finally, it should be noted that our electrophysiological data are also consistent with more integrated RL architectures, such as DYNA, which do not assume two competing controllers (Gershman, Markman, & Otto, 2014; Sutton, 1990). In the DYNA architecture, behavior is completely controlled by the model-free system; the model-based system only has an indirect influence, by training the model-free system offline using simulations of the state space of the task. Although the DYNA architecture might be more in line with our electrophysiological data, several aspects of this model are less straightforward in terms of its predictions regarding neural correlates. For example, DYNA assumes that the model-based system replays experienced state–action pairs and uses this information to simulate the state space. The neural correlates of such a replay process are unclear, but presumably would involve the hippocampus. Future work should try to disambiguate the two approaches and test the neural predictions arising from the two competing models.

Conclusions

The present findings support the idea of integrated neural processing of model-based and model-free information during decision making, but they also point to dissociable mechanisms. As we showed in stimulus-locked analyses at the second stage, the parietal activity (the P300) seems to reflect the integration of model-based information about the transition structure of the task with model-free information about the expected values of choice options during decision making. Moreover, the P300 component was associated with the ability to differentiate between choice options at the second stage, and is predictive of optimal choice behavior. In contrast to the P300, outcome-locked medial prefrontal activity only reflected reward-related processes. Similar to previous studies, we found that the FRN was modulated by outcome valence. However, the FRN did not reflect negative RPEs. This finding may indicate that the FRN is sensitive to the relative valences of outcomes, but does not reflect signed prediction errors. In contrast, the FRP was sensitive to the magnitude of positive prediction errors during learning, indicating that it may reflect cortical processes involved in the updating of expected reward value in the medial PFC.

Author note B.E., M.W., and S.-C.L. designed the study. M.W. and B.E. acquired and analyzed the data, and B.E., M.W., and S.-C.L. wrote the manuscript. This research was funded by the German Federal Ministry of Education and Research through Bernstein Focus Neuronal Basis of Learning Grants (Nos. FKZ 01GQ0913, FKZ 01GQ1313) as well as the German Research Foundation (Deutsche Forschungsgesellschaft), SFB 940, subproject B7 (B.E.).

References

Arbel, Y., Goforth, K., & Donchin, E. (2013). The good, the bad, or the useful? The examination of the relationship between the feedback-related negativity (FRN) and long-term learning outcomes. *Journal of Cognitive Neuroscience, 25,* 1249–1260.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus–norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience, 28,* 403–450. doi:10.1146/annurev.neuro.28.061604.135709

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: Cortico-striatal determinants of goal-directed and habitual action. *Neuropsychopharmacology, 35,* 48–69.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4. Retrieved from http://cran.r-project.org/web/packages/lme4

Bland, A. R., & Schaefer, A. (2011). Electrophysiological correlates of decision making under varying levels of uncertainty. *Brain Research, 1417,* 55–66.

Botvinick, M. M. (2007). Conflict monitoring and decision making: Reconciling two perspectives on anterior cingulate function. *Cognitive, Affective, & Behavioral Neuroscience, 7,* 356–366. doi:10.3758/CABN.7.4.356

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology, 67,* 319–333. doi:10.1037/0022-3514.67.2.319

Cavanagh, J. F. (2015). Cortical delta activity reflects reward prediction error and related behavioral adjustments, but at different times. *NeuroImage, 110,* 205–216.

Cavanagh, J. F., Figueroa, C. M., Cohen, M. X., & Frank, M. J. (2012). Frontal theta reflects uncertainty and unexpectedness during exploration and exploitation. *Cerebral Cortex, 22,* 2575–2586.

Cavanagh, J. F., & Frank, M. J. (2014). Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences, 18,* 414–421.

Cohen, J. (1973). Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and Psychological Measurement, 33,* 107–112. doi:10.1177/001316447303300111

Cohen, M. X., Elger, C. E., & Ranganath, C. (2007). Reward expectation modulates feedback-related negativity and EEG spectra. *NeuroImage, 35,* 968–978.

Cohen, J., & Polich, J. (1997). On the number of trials needed for P300. *International Journal of Psychophysiology, 25,* 249–255.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron, 69,* 1204–1215. doi:10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Human and rodent homologies in action control: Cortico-striatal determinants of goal-directed and habitual action. *Nature Neuroscience, 8,* 1704–1711.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology, 18,* 1–12.

Deserno, L., Huys, Q. J., Boehme, R., Buchert, R., Heinze, H.-J., Grace, A. A.,… Schlagenhauf, F. (2015). Ventral striatal dopamine reflects behavioral and neural signatures of model-based control during sequential decision making. *Proceedings of the National Academy of Sciences*, *112*, 1595–1600. doi:10.1073/pnas.1417219112

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology, 22,* 1075–1081.

Donchin, E. (1981). Surprise!…Surprise? *Psychophysiology, 18,* 493–513. doi:10.1111/j.1469-8986.1981.tb01815.x

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences, 11,* 357–374. doi:10.1017/S0140525X00058027. disc. 374–427.

Eppinger, B., Kray, J., Mock, B., & Mecklinger, A. (2008). Better or worse than expected? Aging, learning, and the ERN. *Neuropsychologia, 46,* 521–539.

Eppinger, B., Mock, B., & Kray, J. (2009). Developmental differences in learning and error processing: Evidence from ERPs. *Psychophysiology, 46,* 1043–1053.

Eppinger, B., Walter, M., Heekeren, H. R., & Li, S.-C. (2013). Of goals and habits: Age-related and individual differences in goal-directed decision-making. *Frontiers in Neuroscience, 7,* 253. doi:10.3389/fnins.2013.00253

Fischer, A. G., & Ullsperger, M. (2013). Real and fictive outcomes are processed differently but converge on a common adaptive mechanism. *Neuron, 79,* 1243–1255.

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

Geisser, S., & Greenhouse, S. W. (1958). An extension of Box's results on the use of the *F*-distribution in multivariate analysis. *Annals of Mathematical Statistics, 29,* 885–891.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General, 143,* 182–194. doi:10.1037/a0030844

Gläscher, J., Daw, N. D., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron, 66,* 585–595. doi:10.1016/j.neuron.2010.04.016

Gratton, G., Coles, M. G. H., & Donchin, E. (1983). A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology, 55,* 468–484. doi:10.1016/0013-4694(83)90135-9

Hämmerer, D., Li, S.-C., Mueller, V., & Lindenberger, U. (2011). Lifespan differences in electrophysiological correlates of monitoring gains and losses during probabilistic reinforcement learning. *Journal of Cognitive Neuroscience, 23,* 579–592.

Herbert, M., Eppinger, B., & Kray, J. (2011). Younger but not older adults benefit from salient feedback during learning. *Frontiers in Psychology, 2*(171), 1–9. doi:10.3389/fpsyg.2011.00171

Herrmann, C. S., Munk, M. H. J., & Engel, A. K. (2004). Cognitive functions of gamma-band activity: Memory match and utilization. *Trends in Cognitive Sciences, 8,* 347–355.

Hodgkinson, G. P., Brown, N. J., Maule, A. J., Glaister, K. W., & Pearman, A. D. (1999). Breaking the frame: An analysis of strategic cognition and decision-making under uncertainty. *Strategic Management Journal, 20,* 977–985.

Holroyd, C. B., & Coles, M. G. H. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review, 109,* 679–709. doi:10.1037/0033-295X.109.4.679

Holroyd, C. B., & McClure, S. M. (2015). Hierarchical control over effortful behavior by rodent medial frontal cortex: A computational model. *Psychological Review, 122,* 54–83. doi:10.1037/a0038339

Holroyd, C. B., Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2003). Errors in reward prediction are reflected in the event-related potential. *NeuroReport, 14,* 2481–2484.

Hong, L., Walz, J. M., & Sajda, P. (2014). Your eyes give you away: Prestimulus changes in pupil diameter correlate with poststimulus task-related EEG dynamics. *PLoS ONE, 9,* e91321. doi:10.1371/journal.pone.0091321

Johansen, J. P., & Fields, H. L. (2014). Glutamatergic activation of anterior cingulate cortex produces an aversive teaching signal. *Nature Neuroscience, 7,* 398–403.

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron, 89,* 221–234. doi:10.1016/j.neuron.2015.11.028

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207,* 203–205. doi:10.1126/science.7350657

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in Rhesus monkeys. *Journal of the Experimental Analysis of Behavior, 84,* 555–579.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron, 81,* 687–699.

Luck, S. J. (2005). Ten simple rules for designing and interpreting ERP experiments. In T. C. Handy (Ed.), *Event-related potentials: A methods handbook* (pp. 17–32). Cambridge: MIT Press.

Marco-Pallares, J., Cucurell, D., Münte, T. F., Strien, N., & Rodriguez-Fornells, A. (2011). On the number of trials needed for a stable feedback-related negativity. *Psychophysiology, 48,* 852–860. doi:10.1111/j.1469-8986.2010.01152.x

Miltner, W. H., Braun, C. H., & Coles, M. G. H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a "generic" neural system for error detection. *Journal of Cognitive Neuroscience, 9,* 788–798. doi:10.1162/jocn.1997.9.6.788

Murphy, P. R., Robertson, I. H., Balsters, J. H., & O'Connell, R. G. (2011). Pupillometry and P3 index the locus coeruleus–noradrenergic arousal function in humans. *Psychophysiology, 48,* 1532–1543. doi:10.1111/j.1469-8986.2011.01226.x

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus–norepinephrine system. *Psychological Bulletin, 131,* 510–532. doi:10.1037/0033-2909.131.4.510

Nieuwenhuis, S., Ridderinkhof, K. R., Talsma, D., Coles, M. G. H., Holroyd, C. B., Kok, A., & van der Molen, M. W. (2002). A computational account of altered error processing in older age: Dopamine and the error-related negativity. *Cognitive, Affective, & Behavioral Neuroscience, 2,* 19–36. doi:10.3758/CABN.2.1.19

Pezzulo, G., Rigoli, F., & Friston, K. (2015). Active Inference, homeostatic regulation and adaptive behavioural control. *Progress in Neurobiology, 134,* 17–35.

R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from www.R-project.org

Rummery, G. A., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems*. Unpublished manuscript. Retrieved from ftp://mi.eng.cam.ac.uk/pub/reports/auto-pdf/rummery_tr166.pdf

Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N., & Dolan, R. J. (2013). Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron, 80,* 1–6.

Sutton, R. S. (1990). *Integrated architectures for learning, planning, and reacting based on approximating dynamic programming.* Paper presented at the Seventh International Conference on Machine Learning, San Francisco, CA.

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

Sutton, S., & Ruchkin, D. S. (1984). The late positive complex: Advances and new problems. *Annals of the New York Academy of Sciences, 425,* 1–23. doi:10.1111/j.1749-6632.1984.tb23520.x

Talmi, D., Atkinson, R., & El-Deredy, W. (2013). The feedback-related negativity signals salience prediction errors, not reward prediction errors. *Journal of Neuroscience, 33,* 8264–8269.

Talmi, D., Fuentemilla, L., Litvak, V., Duzel, E., Duzel, E., & Dolan, R. J. (2012). An MEG signature corresponding to an axiomatic model of reward prediction error. *NeuroImage, 59,* 635–645.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review, 55,* 189–208. doi:10.1037/h0061626

Walsh, M. M., & Anderson, J. R. (2012). Learning from experience: Event-related potential correlates of reward processing, neural adaptation, and behavioral choice. *Neuroscience & Biobehavioral Reviews, 36,* 1870–1884. doi:10.1016/j.neubiorev.2012.05.008

Walz, J. M., Goldman, R. I., Carapezza, M., Muraskin, M., Brown, T. R., & Sajda, P. (2013). Simultaneous EEG-fMRI reveals temporal evolution of coupling between supramodal cortical attention networks and the brainstem. *Journal of Neuroscience, 4,* 19212–19222.

Wilson, R. C., Takahashi, Y. K., Schoenbaum, G., & Niv, Y. (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron, 81,* 267–279.

Wunderlich, K., Smittenaar, P., & Dolan, R. J. (2012). Dopamine enhances model-based over model-free behavior. *Neuron, 75,* 418–424.

Yeung, N., & Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *Journal of Neuroscience, 24,* 6258–6264. doi:10.1523/JNEUROSCI.4537-03.2004