Task Learnability Modulates Surprise but Not Valence Processing for Reinforcement Learning in Probabilistic Choice Tasks

Franz Wurm[1,2,3*], Wioleta Walentowska[4,5*], Benjamin Ernst[1], Mario Carlo Severo[5], Gilles Pourtois[5], and Marco Steinhauser[1]

Abstract

■ The goal of temporal difference (TD) reinforcement learning is to maximize outcomes and improve future decision-making. It does so by utilizing a prediction error (PE), which quantifies the difference between the expected and the obtained outcome. In gambling tasks, however, decision-making cannot be improved because of the lack of learnability. On the basis of the idea that TD utilizes two independent bits of information from the PE (valence and surprise), we asked which of these aspects is affected when a task is not learnable. We contrasted behavioral data and ERPs in a learning variant and a gambling variant of a simple two-armed bandit task, in which outcome sequences were matched across tasks. Participants were explicitly informed that feedback could be used to improve performance in the learning task but not in the gambling task, and we predicted a corresponding modulation of the aspects of the PE. We used a model-based analysis of ERP data to extract the neural footprints of the valence and surprise information in the two tasks. Our results revealed that task learnability modulates reinforcement learning via the suppression of surprise processing but leaves the processing of valence unaffected. On the basis of our model and the data, we propose that task learnability can selectively suppress TD learning as well as alter behavioral adaptation based on a flexible cost–benefit arbitration. ■

INTRODUCTION

How do agents optimally learn to select actions that lead to the achievement of their long-term goals? Reinforcement learning provides an elegant answer to this question by offering a normative computational framework of decision-making for goal-directed agents interacting with an uncertain environment (Sutton & Barto, 2018). Although reinforcement learning spans a very broad range of different algorithms, the most prominent candidate is temporal difference (TD) learning, as it offers a simple yet powerful formalization of trial-and-error learning. More importantly, the existing evidence suggests distinct neural footprints of TD in both animals and humans (Daw & O'Doherty, 2013; Lee, Seo, & Jung, 2012; Botvinick, Niv, & Barto, 2009; Niv, 2009; Holroyd & Coles, 2002). At its core, TD learning revolves around the idea of a prediction error (PE), which is calculated as the difference between the expected and the actual outcome of an action or state in the environment. This PE can then be used as an internal teaching signal as it carries two independent bits of information: (1) information about valence, which describes if an outcome was better (+) or worse (−) than expected, and (2) information about surprise, which quantifies how far off the expectation was. For subsequent computations, this information can be used to update expectations and translate them into future behavior.

Although the idea of TD learning emerged as an important principle for explaining human decision-making (Daw & O'Doherty, 2013; Lee et al., 2012; Botvinick et al., 2009; Niv, 2009; Holroyd & Coles, 2002), little is known on how its application is influenced by explicit knowledge about the environmental structure. Crucial but still open questions are whether TD learning can be suppressed if a task is not learnable (i.e., outcome feedback is uninformative for optimizing behavior) and, if so, for which bit of information this suppression occurs. In this study, the effect of learnability is investigated by contrasting brain activity in a learning task and a gambling task. Considering electrophysiological correlates of feedback processing in a model-based analysis, we asked which aspect of the PE is affected by learnability—information about valence or information about surprise.

The significance of TD learning for behavior is supported by evidence coming from both animal and human studies (Dayan & Niv, 2008; Balleine, 2005; Dickinson & Balleine, 2002). Although early work showed that the firing pattern of midbrain dopamine neurons in monkeys closely resembles a PE as predicted by TD algorithms (Schultz, Dayan, & Montague, 1997; Montague, Dayan,

[1]Catholic University of Eichstätt-Ingolstadt, Germany, [2]Leiden University, [3]Leiden Institute for Brain and Cognition, [4]Jagiellonian University, Krakow, Poland, [5]Ghent University
*These authors contributed equally to this work.

& Sejnowski, 1996), similar findings were also later observed in humans (Schonberg et al., 2010; D'Ardenne, McClure, Nystrom, & Cohen, 2008; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006). Since then, correlates of the PE have been replicated extensively in subpopulations of several neural structures such as the striatum, the amygdala, and multiple areas of the (pre)frontal cortex (Schultz, 2016).

In recent years, an increasing number of studies have used ERPs to investigate the neural dynamics underlying reinforcement learning. It has consistently been shown that ERPs are sensitive to feedback valence already at around 200 msec after feedback onset in a component called the feedback-related negativity (FRN; Miltner, Braun, & Coles, 1997). The FRN manifests as a negative deflection at frontocentral electrodes about 200–350 msec after feedback onset. The FRN is typically larger for negative than positive feedback (for reviews, see San Martín, 2012; Walsh & Anderson, 2012), a phenomenon we refer to as the ΔFRN.[1] Crucially, numerous studies could show that the ΔFRN scales with outcome magnitude and likelihood, supporting theoretical assumptions that the FRN amplitudes also reflect information about surprise as suggested by the reinforcement learning theory (Holroyd & Coles, 2002; for reviews, see Sambrook & Goslin, 2015; San Martín, 2012; Walsh & Anderson, 2012). In addition to the ΔFRN, feedback elicits a P3, which has been associated with a broad range of cognitive phenomena (for reviews, see Polich, 2020; San Martín, 2012). Manifesting as a parietal positivity around 300–600 msec after feedback onset, this feedback-P3 has been mainly reported to reflect surprise (Kopp et al., 2016; Seer, Lange, Boos, Dengler, & Kopp, 2016; Kolossa, Kopp, & Fingscheidt, 2015; Mars et al., 2008) and learning (Nassar, Bruckner, & Frank, 2019; Jepma et al., 2016, 2018; Fischer & Ullsperger, 2013). Importantly, recent evidence suggests that the relationship between P3 and learning is strongly modulated by the task context, thereby supposedly reflecting a mediated response to surprise (Nassar et al., 2019). Despite the obvious link between the P3 and surprise, its role within a reinforcement learning perspective of human decision-making is still unclear.

Crucial yet unanswered questions are whether and how top–down processes can modulate TD learning. These questions have recently been addressed in studies manipulating the goal relevance of feedback, that is, the informativeness of feedback for optimizing behavior. These studies revealed rather similar effects for the ΔFRN and the P3. When participants were instructed on the reliability of feedback in probabilistic learning (i.e., the probability that the same action will lead to the same outcome on future trials), the ΔFRN as well as the P3 was increased for reliable feedback as compared to unreliable feedback (Di Gregorio, Ernst, & Steinhauser, 2019; Ernst & Steinhauser, 2017, 2018; Schiffer, Siletti, Waszak, & Yeung, 2017). The modulation of the ΔFRN could partially be attributed to a top–down modulation because the effect was observed even when the objective reliability of feedback was held constant (Di Gregorio et al., 2019; Schiffer et al., 2017). Similar results were obtained in a speeded go/no-go task in which feedback could indicate whether responses were correct and faster than an unknown response time cutoff (Walentowska, Moors, Paul, & Pourtois, 2016) and in a time estimation task (Severo, Paul, Walentowska, Moors, & Pourtois, 2020). Again, the ΔFRN and the P3 were larger when feedback provided valid and relevant information for optimizing task performance.

The results described above could be taken as evidence that all aspects of TD learning are suppressed when feedback is goal irrelevant. However, such a conclusion might be premature for several reasons. First, as we have seen, the two main ERP components of feedback processing (ΔFRN and P3) cannot be unequivocally mapped onto the aspects of the PE. Valence and surprise information might contribute to both components, which raises the possibility that only one of these processes caused the ΔFRN and P3 results. A viable method to deal with this problem is to use computational modeling to provide direct estimates of the PE that can inform model-based analyses of the neural data (Mars, Shea, Kolling, & Rushworth, 2012; Gläscher & O'Doherty, 2010). Second, although the abovementioned studies manipulated the information content of feedback, behavioral adaptation could still lead to performance improvements in all conditions, at least in some of the studies (e.g., Di Gregorio et al., 2019). A stronger manipulation would be to vary the learnability of tasks, that is, to contrast a probabilistic learning task with a mere gambling task. Whereas performance can be improved by utilizing feedback in a learning task, gambling tasks do not allow for improving performance at all. From a normative perspective, the calculation of a PE has no utility in the latter condition. Therefore, the comparison between learning and gambling is a strong test for revealing top–down processes in TD learning. First evidence for such a modulation stems from a study in which the ΔFRN has been shown to vary with PEs in a learning task but not in a gambling task (Holroyd, Krigolson, Baker, Lee, & Gibson, 2009).

The aim of this study is to investigate how the calculation of a PE is influenced by the learnability of the environment, which was manipulated by varying the causal structure of a task. On the basis of the idea that the PE carries information about both valence and surprise of an outcome, we asked which of these aspects is affected when participants are confronted with a learnable task and a nonlearnable task, respectively. In both tasks, we used a simple variant of a two-armed bandit problem (Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006). Whereas participants could adapt to the varying reward probabilities by utilizing feedback in the learning task, feedback was fully unrelated to participants' choice behavior in the gambling task. To prevent the results from reflecting systematic differences in the frequency and order of feedback, the two tasks were matched with respect to feedback sequence (see Methods for details). Crucially, we used a computational modeling

approach to separate the effects of learnability on the two aspects of the PE. By modeling behavior in the two tasks using multiple instantiations of the same computational algorithm, we could robustly estimate PEs for each single trial and use its constituent parts to predict neural data in a linear model. On the basis of the idea that valence information is strongly related to the FRN amplitudes whereas surprise information is reflected by the P3 (Nassar et al., 2019; Mars et al., 2008), we expected to find correlates of valence at earlier frontocentral locations and correlates of surprise at later posterior locations. Importantly, if TD learning is influenced by learnability, we hypothesized to find either a reduced ΔFRN or a reduced P3, or both, for the gambling task relative to the learning task. Moreover, any influence of task learnability on the calculation of PEs should be reflected by a modulation of the strength by which its constituent parts predict the respective neural activity.

METHODS

Participants

Thirty participants (24 women) between 18 and 28 years old (M = 21.9 years, SD = 2.6 years) with normal or corrected-to-normal vision and free from neurological and psychiatric history (based on self-reports) participated in the study. Participants were recruited at Ghent University. They received €20 for participation and a performance-dependent bonus (M = €2.16, SD = €1.14). All participants provided informed consent, and the study was approved by the local ethics committee (Faculty of Psychology and Educational Sciences of Ghent University).

Stimuli

The experimental task was programmed and executed using Presentation software (Neurobehavioral Systems). The stimulus set consisted of 112 black-and-white drawings obtained from the online database of the International Picture Naming Project (Szekely et al., 2005). Pictures were converted into 300 × 300 pixel images corresponding to a visual angle of 6.4° in height and 6.4° in width at a viewing distance of 80 cm. For each block, eight pictures were randomly sampled from the stimulus array (without replacement) and then further divided into four stimulus pairs, which were shown to the left and right of a centrally presented fixation cross at a distance of 0.7°. The left or right position for each stimulus in a pair was randomly determined for each trial. The feedback stimuli were a white diamond (win) and a white rock (loss) that were centrally presented at a height of 2.8° and a width of 2.5°. All stimuli were presented on a black background.

Task and Procedure

We employed two variants (learning task vs. gambling task) of a simple two-armed bandit task in which participants had to repeatedly choose between two actions corresponding to the two stimuli in a stimulus pair (Figure 1A). Both tasks differed regarding their objective learnability, that is, the utility of feedback for improving performance. For the learning task, the mapping between actions and feedback was highly coherent, making behavioral adaptation based on feedback relevant for optimizing outcome. The two possible actions of each stimulus pair were associated with reward probabilities p and $(1 - p)$, and p varied throughout a block according to a Gaussian random walk with reflecting boundaries at 0 and 1 (e.g., Sambrook, Hardwick, Wills, & Goslin, 2018; Daw, 2011). On each repetition of this stimulus pair, p was constantly updated by adding a value drawn from a Gaussian distribution with M = 0 and SD = 0.1 (Figure 1B). If this update led to values below 0 or above 1, p was set to 0 or 1, respectively. For the gambling task, the mapping between action and feedback was pseudorandom. More specifically, we used the same feedback sequence as in the preceding learning-task block while presenting a randomly shuffled stimulus pair sequence. Because of this combination of a yoked feedback sequence and a

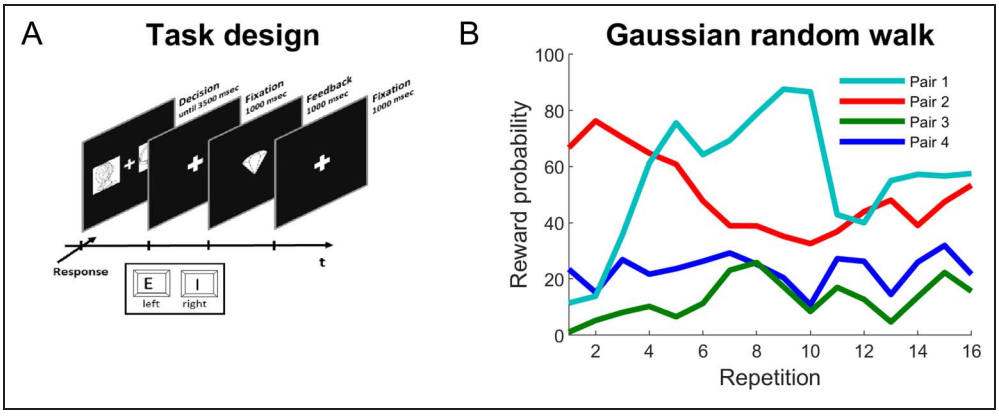

Figure 1. (A) Graphical illustration of a trial. Participants had to choose between two stimuli. After an intermediate presentation of a fixation cross, feedback was shown followed by another fixation cross. (B) Example of a Gaussian random walk. Over the course of a block, reward probabilities p associated with one of the stimuli in a stimulus pair gradually changed according to a Gaussian random walk with a mean of 0 and a standard deviation of 0.10. The probability of reward for the other stimulus in the stimulus pair was $1 - p$.

randomized stimulus sequence, participants' actions could not influence subsequent feedback making behavioral adaptation irrelevant for optimizing outcome.

The procedure of a trial was identical in the two different task variants and is illustrated in Figure 1A. At the beginning of each trial, a stimulus pair was presented for up to 3500 msec. An action had to be made by pressing one of two keys on an English standard keyboard (E with the left index finger for choosing the left stimulus, I with the right index finger for choosing the right stimulus). After the keypress, the stimulus pair disappeared, and a fixation cross was presented for a fixed time of 1000 msec. Afterward, feedback was centrally presented for 1000 msec. Diamonds indicated wins, whereas rocks indicated losses. After the disappearance of the feedback stimulus, a fixation cross displayed for 1000 msec marked the end of the trial. If no action was carried out during presentation of the stimulus pair, the trial was aborted and counted as a miss. The feedback stimulus for misses was a white question mark. The participants were informed that misses were associated with a loss of 5 cents, whereas positive and negative feedback was associated with a win/loss of ±2 cents, respectively. Participants worked through 10 blocks with 64 experimental trials each. Learning and gambling tasks alternated across blocks, and the task of the first block was counterbalanced across participants. Before each block, participants were informed about which task was used in the upcoming block. In each block, four new stimulus pairs were introduced. Blocks consisted of 16 subblocks in which all four stimulus pairs were presented in random order. To avoid unequal attention allocation to the feedback stimuli between tasks, we included 10 catch trials in each block, which were randomly interspersed among the experimental trials. Participants' task in catch trials was to recall the previously shown feedback. Participants were presented a prompt ("What feedback was shown last?") and indicated the feedback by pressing the E key in case of a win, the I key in case of a loss, or the space key if they did not remember (or if the last trial was a miss).

At the beginning of the experiment, participants received written instructions on the task. The instructions included information on decision-making and feedback in the task in general, as well as the basic properties of the learning task and the gambling task. For the learning task, participants were instructed that they could learn which action is most likely to be followed by a win. Furthermore, they were informed about the implementation of the random walk and the resulting changes in reward probabilities. For the gambling task, participants were told that they are unable to learn which action is most likely to be followed by reward. Afterward, they worked through two short practice blocks, one from each task, in which two stimulus pairs were presented 10 times. These blocks were then followed by two long practice ones, again one from each task, in which four stimulus pairs were presented 16 times. The order of tasks in the practice blocks was the same as in the subsequent experimental blocks. For participants starting the experiment with the gambling task, the long training block for the learning task was used for initial yoking. In practice blocks with the gambling task, the feedback sequence was random and not yoked to a learning block.

After the experiment, participants completed a number of questionnaire items to collect data on subjective information and use of feedback in the two tasks using a computer-mouse slider on the visual analog scale. For a detailed description of the items, refer to Table 1. Each aspect was rated separately for the learning and gambling tasks on a continuous scale ranging between 0 and 100. For Items 1–5, the scale ranged between "not at all" and "a lot." For Items 6 and 7, the scale ranged between "very unhappy" and "very happy." Both the catch trials and the subjective ratings were implemented to quantify that the feedback between learning and gambling tasks was equally attended.

Behavioral Data Analysis

Choice behavior was analyzed in two steps. First, we considered the proportion of correct actions in the learning task. Correct actions were defined as those actions that were associated with a higher reward probability. The total proportion of correct actions was tested against chance level using a two-tailed paired-sample t test. Furthermore, the proportion of correct actions calculated separately for each of the 16 subblocks was submitted to a one-way repeated-measure ANOVA with the variable subblocks. For both significance tests, the proportion of correct action was arcsine square root transformed (Winer, Brown, & Michels, 1991). Second, we analyzed the proportion of stay behavior, which was defined as the probability of repeating the same action as on the previous encounter with the same stimulus pair. The goal of this analysis was to investigate whether we find the commonly observed win–stay lose–shift (WSLS) behavior as an indicator of behavioral adaptation and whether this pattern was more pronounced in the learning task than in the gambling task. We therefore applied a logistic regression with choice type as criterion and previous outcome, task, and their interaction as predictors. Choice type could either be a stay or a switch (coded as 1 and 0), previous outcome could either be a win or a loss (coded as 1 and −1), and task could either be learning or gambling (coded as 1 and −1). All within-participant variables (intercept, previous outcome, task, interaction) were implemented as random effects and therefore were allowed to vary across participants (e.g., Gillan, Otto, Phelps, & Daw, 2015; Daw, Gershman, Seymour, Dayan, & Dolan, 2011). Performance on catch trials was analyzed by calculating the proportion of catch trials on which the feedback of the previous trial was correctly indicated. This accuracy was compared between the learning-task blocks and the gambling-task blocks using a paired-sample t test. Finally, mean ratings for each item of the final questionnaire

Table 1. Subjective Ratings

		Task		
		Learning	*Gambling*	
Item	*Question*	*M (SD)*	*M (SD)*	*t Value*
Interest in task	How interesting did you find the rounds?	62 (16)	41 (27)	3.65**
Attention to feedback	How much attention did you pay to the outcome in the rounds?	79 (14)	34 (26)	7.22***
Contribution to win	How much did you credit yourself for a win in the rounds?	70 (18)	58 (28)	2.49*
Blame for loss	How much did you blame yourself for a loss in the rounds?	59 (24)	20 (22)	6.18***
Usefulness	How useful was the outcome to guide your next action?	68 (18)	26 (29)	5.90***
Feeling after win	How did you feel when you won?	79 (14)	71 (23)	1.80
Feeling after loss	How did you feel when you lost?	26 (13)	39 (12)	−5.69***

M and *SD* represent the mean and standard deviation, respectively. Asterisks indicate the significance level.

* $p < .5$.

** $p < .01$.

*** $p < .001$.

were compared between learning and gambling tasks using paired-sample *t* tests.

Electrophysiological Recordings and Analyses

Participants were seated comfortably in a dimly lit, sound-attenuated, and electrically shielded cabin. The EEG was recorded using a BioSemi Active-Two system (BioSemi) with 64 Ag–AgCl electrodes from channels Fp1, AF7, AF3, F1, F3, F5, F7, FT7, FC5, FC3, FC1, C1, C3, C5, T7, TP7, CP5, CP3, CP1, P1, P3, P5, P7, P9, PO7, PO3, O1, Iz, Oz, POz, Pz, CPz, Fpz, Fp2, AF8, AF4, AFz, Fz, F2, F4, F6, F8, FT8, FC6, FC4, FC2, FCz, Cz, C2, C4, C6, T8, TP8, CP6, CP4, CP2, P2, P4, P6, P8, P10, PO8, PO4, and O2, as well as the left and right mastoid. The horizontal and vertical EOG was monitored by means of four electrodes, placed above and below the right eye and the outer canthi of both eyes. Sampling rate was 512 Hz.

As with the behavioral data, EEG data were analyzed using custom-made routines in MATLAB (The MathWorks). For the preprocessing, we used EEGLAB 13.5.4b (Delorme & Makeig, 2004), an open-source toolbox for EEG data analysis (EEGLAB toolbox for single-trial EEG data analysis, Swartz Center for Computational Neurosciences; www.sccn.ucsd.edu/eeglab). EEG data were offline rereferenced to averaged mastoids, band-pass filtered to exclude frequencies below 0.1 Hz and above 35 Hz and divided into epochs from 1000 msec before to 1500 msec after feedback onset. Baseline activity from 200 msec before feedback onset was removed. Bad channels were interpolated using spherical spline interpolation if they met the joint probability criterion (threshold = 5) as well as the kurtosis criterion (threshold = 5) in EEGLAB's channel rejection routine. Epochs were excluded whenever neural activity in a channel deviated more than ±300 μV from the epoch mean. This criterion was not applied to those channels that are typically contaminated by blinks (Fp1, Fpz, Fp2, AF7, and AF8) as this activity was corrected later. In a next step, an infomax-based independent component analysis (Bell & Sejnowski, 1995) was conducted. After visual inspection of the derived independent components, those representing eye blinks and muscular artifacts were identified and removed from the data. The remaining epochs were averaged separately for each participant and task. On average, this resulted in the following numbers of artifact-free trials in the respective feedback/task conditions: 170.7 (*SD* = 17.0) for win/learning, 127.1 (*SD* = 13.3) for loss/learning, 170.0 (*SD* = 17.3) for win/gambling, and 126.3 (*SD* = 14.7) for loss/gambling.

FRN amplitudes were quantified using the mean amplitude in a time window of 200–400 msec after feedback presentation at electrode FCz (Sambrook & Goslin, 2014). The P3 amplitude was measured using a peak amplitude approach (Ernst & Steinhauser, 2018; Sailer, Fischmeister, & Bauer, 2010; Pontifex, Hillman, & Polich, 2009) because P3 peaks are often shifted across conditions, making a quantification in a fixed time interval difficult. We first identified the maximum amplitude at electrode Pz in a time window of 200–700 msec and analyzed the peak amplitude as well as the latency of the peak. For statistical analysis, we applied repeated-measure ANOVAs involving the

variables Outcome (win, loss) and Task (learning, gambling) for amplitudes in both the FRN and P3 time window.

Computational Modeling

We used computational modeling for two reasons. First, we sought to investigate the mechanisms at play during feedback processing in general. By constructing different instantiations of computational reinforcement learning models, we aimed to identify the mechanisms that are most likely implemented on a group level. Second, we sought to leverage this mechanistic explanation of the behavioral data for a subsequent model-based analysis of the EEG data. By deriving single-trial estimates of PEs and using their constituent parts (valence and surprise) as predictors for the neural data, we aimed to reveal the specific patterns of TD learning, especially regarding its modulation via the learnability of the environment.

The fundamental rationale, which is shared between all computational models detailed below, is that learning occurs in two successive steps: the calculation of a (reward) PE and the updating of action values based on this specific PE signal. We therefore implemented a learning policy that can be characterized as a TD learning architecture (Sutton & Barto, 2018; Gläscher & O'Doherty, 2010). The calculation of the PE follows the general approach with the equation

$$PE(t) = [r(t) - Q(a,s,t)] \tag{1}$$

where $r(t)$ denotes the outcome received in that trial and $Q(a,s,t)$ denotes the expected value of the chosen action in the specific stimulus pair. Please note that this calculation of the PE is most accurately called Rescorla–Wagner learning rule, as it does not include a term for the expected value of future states, which determines how agents navigate in more complex environments. On the basis of the PE, the expected value of the chosen action is incrementally updated according to the equation

$$Q(a,s,t+1) = Q(a,s,t) + \alpha \times PE(t) \tag{2}$$

where α is the learning rate, which controls for the speed of updating. The probability of selecting action a in a specific stimulus pair s is then simply determined by inserting the updated action values in a soft-max decision rule

$$P(a_t = a|s) = \frac{\exp(\beta \times Q(a,s,t) + p \times rep(a,s))}{\sum_{a'} \exp(\beta \times Q(a',s,t) + p \times rep(a',s))} \tag{3}$$

where the inverse temperature β guides the stochasticity of the choices and the perseveration parameter p captures choice perseveration ($p > 0$) or switching ($p < 0$; Lau & Glimcher, 2005). The indicator function $rep(a)$ takes a value of 1 if action a was chosen on the last trial of the same stimulus pair and 0 otherwise.

We hypothesized that the instruction on the nature of the task (learning vs. gambling) causes behavioral differences between these tasks by affecting parameter settings. In this study, we identified three possible mechanisms that could lead to differences between learning and gambling behavior. Each putative mechanism is linked to one specific model parameter. For models with a variable learning rate, the learning rate α is estimated separately for each task condition. This allows the updating of action values (Equation 2) to be modulated between the two tasks. As this parameter controls for the integration of new experience over trials, we expect to find higher learning rates for the learning task compared to the gambling task. For models with variable inverse temperature, the inverse temperature β is estimated separately for each task condition. This allows action selection (Equation 3) to be modulated between the tasks. As this parameter controls for the stochasticity of choice behavior (with higher values increasing experience-based action selection), we expect to find higher inverse temperatures for the learning task compared to the gambling task. For models with variable policy mixture, we took into consideration previous findings showing that behavioral and neural data can be best explained by a mixture of different choice policies (e.g., Daw et al., 2011; Keramati, Dezfouli, & Piray, 2011). Core mechanism of this mixture model is the eponymous mixture of action values for two separate and independent choice policies. Here, these policies are TD learning, as described earlier, and random choice or guessing behavior. Formally, the mixture of policies was realized by calculating a net action value Q_{NET} as the weighted combination of the independent action values according to the equation

$$\begin{aligned} Q_{\text{NET}}(a,s,t) = {} & \omega Q_{\text{LEARN}}(a,s,t) \\ & + (1-\omega) Q_{\text{GUESS}}(a,s,t) \end{aligned} \tag{4}$$

where $Q_{\text{LEARN}}(a,s,t)$ denotes action values derived from a TD mechanism (Equations 1 and 2) and $Q_{\text{GUESS}}(a,s,t)$ denotes action values derived from a guessing mechanism. To allow random action selection, $Q_{\text{GUESS}}(a,s,t)$ is initialized as 0 and kept constant during the experiment. For action selection, Q_{NET} values are submitted to Equation 3. The RL-basedness parameter ω describes the contribution of the learning policy, thus guiding the trade-off between the contribution of guessing and learning policy to overt behavior. If the RL-basedness parameter approaches 1, the learning policy is predominant and drives action selection. If the RL-basedness parameter approaches 0, the guessing policy is predominant and drives action selection. For models with this mechanism, this RL-basedness parameter ω was estimated separately for each task condition. In the gambling task, feedback cannot be predicted, which might make a mere stochastic action selection viable (Wurm, Ernst, & Steinhauser, 2020) and sometimes even advantageous to learning policies (Tervo et al., 2014). We therefore expect to find higher RL-basedness for the learning task compared to the gambling task.

Given the task design and instruction, feedback on a trial cannot only be used to update the value of the chosen action. In addition, the forgone action on each trial can be updated using the reversed PE. In contrast to the partial

updating mechanism, as detailed in Equation 2, this full updating mechanism should speed up learning. In line with the literature (Burnside, Fischer, & Ullsperger, 2019; Palminteri, Lefebvre, Kilford, & Blakemore, 2017), it is plausible to assume that the participants exploited this feature of the task. Therefore, we incorporated each of the three architectures for both the partial and full updating mechanisms, resulting in six candidate models.

All models were implemented in MATLAB v8.6 and the *mfit* toolbox (Gershman, 2016) was used for parameter fitting. Parameters were fitted to the simulated or observed data using nonuniform priors and specific bounds (Gershman, 2016). The learning rate was bound in [0; 1], and the respective prior was $\alpha \sim \text{beta}(1.2, 1.2)$. The inverse temperature was bound in [0; 20], and the prior was $\beta \sim \text{gamma}(2, 1)$. The perseveration parameter was bound in [−5; 5], and the prior was $p \sim \text{gauss}(0, 1)$. Finally, RL-basedness was bound in [0; 1], and the prior was $w \sim \text{beta}(1.2, 1.2)$.

To show that the above-described models lead to distinct and identifiable behavioral patterns, we conducted a model recovery analysis. We simulated each of the six candidate models for 200 participants. Consistent with the empirical task, each simulated participant worked through 10 blocks with 64 trials each; task identity alternated across blocks, and feedback for the gambling blocks was yoked. The parameters for each model and participant were drawn from the same priors that are used for model fitting (see above). To ensure behavioral modulation in line with our predictions for the empirical data, the variable parameters for each model (learning rate, inverse temperature, or RL-basedness) were sampled independently for each task condition. For the variable learning rate models, learning rates for the learning task were uniformly drawn from the interval [0.5; 1], whereas learning rates for the gambling task were drawn from the interval [0; 0.5]. For the variable inverse temperature models, inverse temperatures for the learning task were drawn from the interval [1; 5], whereas inverse temperatures for the gambling task were drawn from the interval [0; 1]. For the variable policy mixture models, RL-basedness for the learning task was drawn from the interval [0.5; 1], whereas RL-basedness for the gambling task was drawn from the interval [0; 0.5]. This simulation procedure ensures that the models generate meaningful behavioral patterns that are generated via quantitative differences in model parameters. The simulated data from each model were then subjected to model comparison in the same way as for the empirical data (see below) to verify that the model that best fit to the data is also the model that generated the data.

In a next step, we performed model selection based on the empirical data. We fitted each of the previously detailed models to the observed behavioral data from the participants. Three measures are reported for model selection: the Bayesian information criterion (BIC), the Akaike information criterion (AIC), and the protected exceedance probability (PXP). In contrast to the two former metrics, the latter is derived using a Bayesian approach with PXP quantifying the probability that any model considered for model comparison is more frequent than all other possible models, given equal prevalence in the population (Rigoux, Stephan, Friston, & Daunizeau, 2014). In contrast to BIC- or AIC-informed approximation of log evidence, and as implemented as default in the mfit toolbox, the PXP in our study is derived under the Laplace approximation (e.g., Friston, Mattout, Trujillo-Barreto, Ashburner, & Penny, 2007). Moreover, a Bayesian approach for model selection allows us to derive posterior probabilities for each participant, which quantify the posterior belief that a specific model generated the data for that participant. As the model recovery analysis revealed that the PXP metric leads to a better recoverability than AIC and BIC, we relied our model selection procedure on this metric but also report BIC and AIC for comparison.[2]

Subsequently, we compared parameter estimates from the empirical data between the learning task and the gambling task using two-tailed paired-sample tests. Because learning rate and RL-basedness are interpreted as a ratio, their values were arcsine square root transformed (Winer et al., 1991).

For the model-based single-trial analysis, we constructed a general linear model to predict single-trial EEG activity at each electrode and time point, separately for each participant. Regressors included the task condition as well as the information about valence and surprise. Task could either be learning or gambling (coded as 1 and −1). Valence was defined as the sign of the PE (+1 vs. −1), and surprise was defined as the absolute value of the PE. PEs for each candidate model and participant were simulated by feeding the estimated model parameters back into the same model that was initially used to calculate the model parameters. On the basis of the idea of random effects (different models may explain the data from different participants), we utilized the Bayesian model selection approach and weighted PEs from all models according to their participant-specific posterior belief to obtain a weighted average PE. This weighted average PE was used to derive the valence and surprise regressors. Before regression, the feedback-locked EEG data were downsampled to 125 Hz. All regressors were *z*-scored. For the surprise regressor, individual values were *z*-scored separately for the learning task and the gambling task. Moreover, we included all possible interactions between regressors. The resulting linear equation took the following form for each task:

$$\begin{aligned} EEG \sim\ & \text{b0} + \text{b1} \times Task + \text{b2} \times Valence + \text{b3} \\ & \times Surprise + \text{b4} \times TaskValence + \text{b5} \\ & \times TaskSurprise + \text{b6} \times ValenceSurprise \\ & + \text{b7} \times TaskValenceSurprise + error, \end{aligned} \tag{5}$$

where composite terms (e.g., *TaskValence*) denote the interaction between the respective regressors. To ensure

comparability within task conditions and between participants and to penalize multicollinearity of predictors, the resulting beta values were standardized by their respective standard deviation (Fischer & Ullsperger, 2013). Only after this normalization procedure, the individual beta weights were tested against zero via two-tailed cluster-based permutation tests implemented in the Mass Univariate ERP Toolbox (Groppe, Urbach, & Kutas, 2011) conducted across sampling points and electrodes. Corrections for a family-wise alpha level of .05 were applied, and clusters were identified for all sampling points at which the uncorrected p value fell below .05. We used 10^5 permutations to obtain sufficient test distributions. All sampling points between 0 and 1000 msec after feedback presentation and all electrodes were considered for permutation testing. To investigate how reinforcement learning is modulated between the learning and gambling tasks, we applied the identical analysis to the difference between beta values from both tasks, separately for reward prediction errors and action values.

RESULTS

Behavioral Data

In a first analysis, we investigated whether participants show learning in the learning task by considering the mean percentage of correct actions. An action was defined as correct if the reward probability associated with this action exceeded 50%. Accordingly, a correct decision was not defined by actual positive feedback but by the underlying reward probability. Because of the nonlearnable structure of the gambling task, correct actions can be determined only for the learning task. Overall, participants' performance in this task (M = 62.1%, SEM = 1.3%) significantly exceeded chance level, $t(29) = 9.22$, $p < .001$, $d = 1.68$. Comparing performance across subblocks, we found a significant main effect of subblock, $F(15, 29) = 4.47$, $p < .001$, $\eta_p = .13$. Figure 2A suggests that no performance improvement occurs beyond the second trial. This reflects that reward probabilities are not only variable but even include reversals of the correct actions, which prevents a monotonically increasing performance. Nevertheless, these results indicate that participants were able to use feedback for improving performance in the learning task.

Whereas the preceding analysis shows that participants can utilize coherent action–feedback mappings in the learning task to improve performance, we were now interested whether participants adapted behavior based on feedback by considering switch/stay behavior across consecutive encounters with the same stimulus pair. A common finding is that participants stay with their choices after wins but switch choices after losses (e.g., Daw et al., 2011; Cohen & Ranganath, 2007). This pattern of WSLS behavior is also possible in the gambling task, although it cannot lead to performance improvement because of the nonlearnable structure of this task. We hypothesized to find more evidence for WSLS behavior in the learning task provided that the learnability influences the strength of behavioral adaptation. As illustrated in Figure 2B, the probability of stay behavior was indeed higher after wins than after losses in both tasks, but this effect was larger for the learning task. Logistic regression analysis showed a significant main effect for previous reward, $\beta = 0.39$, $p < .001$, and task, $\beta = 0.32$, $p < .001$, as well as a significant interaction between both, $\beta = 0.27$, $p < .001$. Separate analyses for each task revealed that previous reward affected stay behavior for both the learning task, $\beta = 0.67$, $p < .001$, and the gambling task, $\beta = 0.11$, $p = .002$. These results demonstrate that the learnability of a task influences the strength of behavioral adaptation based on feedback.

In a last step, we analyzed catch trial performance and subjective ratings. Analyzing accuracy for catch trials, we found that although mean values were high for both the learning task (M = 94.0%, SEM = 1.0%) and the gambling task (M = 86.6%, SEM = 1.6%), performance was significantly higher in the learning task, $t(29) = 5.89$, $p < .001$, $d = 1.08$, indicating decreased attentional resource allocation in the gambling context. Such an interpretation receives further support when analyzing the subjective ratings obtained after the experiment (Table 2). Here, participants explicitly reported a significantly stronger allocation of attention toward feedback in the learning task compared to the gambling task. In addition, the learning task was associated with stronger interest in

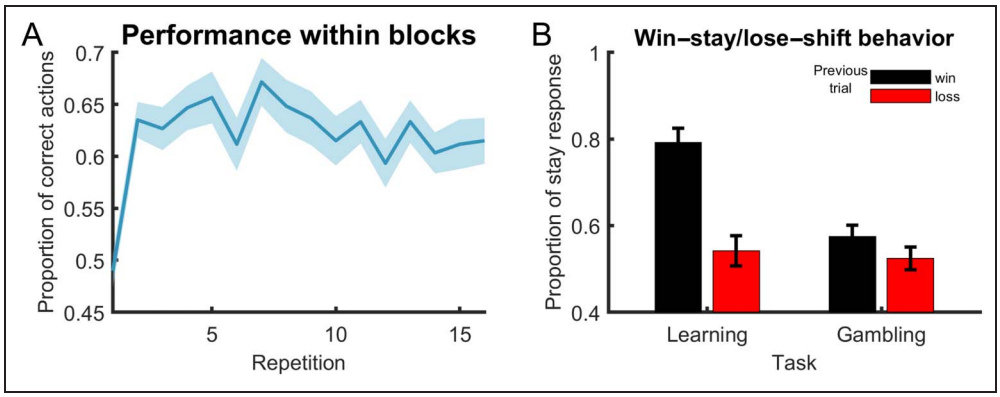

Figure 2. (A) Course of the proportion of correct actions within blocks of the learning task. Shaded areas depict the within-participant standard error of the mean. (B) Proportion of stay responses in the learning and gambling tasks. Error bars depict the within-participant standard error of the mean.

Table 2. Model Comparison

	Mechanism	*Updating*	*#*	*−LL*	*AIC*	*BIC*	*PXP*
1	α	Partial	4	−404.4 (49.6)	816.8 (99.1)	834.7 (99.1)	0.12
2	β	Partial	4	−405.6 (49.8)	819.2 (99.4)	837.1 (99.4)	0.12
3	ω	Partial	5	−404.68 (49.6)	819.4 (99.1)	841.7 (99.1)	0.12
4	α	Full	4	−401.6 (51.7)	811.2 (103.4)	829.0 (103.4)	0.15
5	β	Full	4	−403.5 (51.6)	815.0 (103.2)	832.9 (103.2)	0.13
6	ω	Full	5	401.8 (51.7)	813.6 (103.4)	835.9 (103.4)	0.36

The mechanism indicates which parameter was allowed to vary between tasks. α is the learning rate parameter, β is the inverse temperature, and ω the RL-basedness. Updating indicates which algorithm is implemented for value updating. Partial refers to the algorithm that only updates values for the chosen action on that trial. Full refers to the algorithm that also updates the forgone action. # is the number of free parameters within a model, −LL is the negative log likelihood, BIC is the Bayesian information criterion, AIC is the Akaike information criterion, and PXP is the protected exceedance probability. Values represent the mean, and those in brackets indicate the standard error of the mean.

the task, higher perceived contribution to win, and more blame after losses. Crucially, outcome in the learning task was perceived to be significantly more useful compared to the gambling task. Taken together, the results from both catch trial performance and subjective ratings provide support for the effectiveness of our manipulation of learnability.

ERPs

After finding differences in behavioral adaptation between tasks, we were interested to elucidate the underlying neural mechanisms by analyzing feedback-locked ERPs. If reinforcement learning is influenced by learnability, we hypothesized to find either a reduced ΔFRN or a reduced P3, or both, for the gambling task relative to the learning task.

Analyzing amplitudes in the time window of the FRN at electrode site FCz (Figure 3A), we found a significant main effect of both Outcome, $F(1, 29) = 66.15$, $p < .001$, $\eta_p = .70$, and Task, $F(1, 29) = 15.72$, $p < .001$, $\eta_p = .35$. The FRN was increased (i.e., more negative) for loss ($M = 2.22$ μV, $SEM = 0.80$) relative to win feedback ($M = 6.47$ μV, $SEM = 0.72$) and was increased in the gambling task ($M = 3.25$ μV, $SEM = 0.76$) relative to the learning task ($M = 5.44$ μV, $SEM = 0.83$). However, we did not find a significant interaction between Task and Outcome, $F(1, 29) = 0.70$, $p = .410$, $\eta_p = .02$. That is, the ΔFRN represented by the difference between win and loss (Figure 3B and C) was comparable in the learning and gambling tasks.

Figure 4 shows the results for the analysis for the P3. In contrast to the previous analysis, the size of the P3 was quantified as the amplitude of the peak in the time range

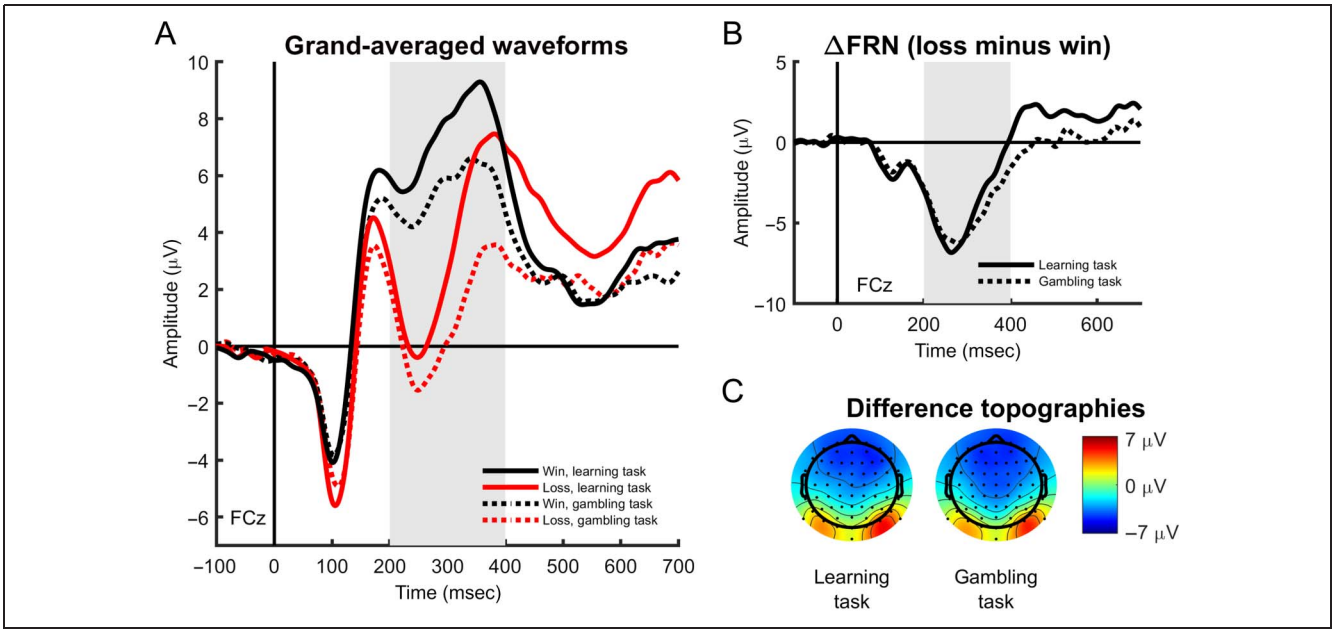

Figure 3. (A) Feedback-locked grand-averaged waveforms at electrode FCz. The gray area indicates the time window for the mean amplitude analysis. (B) ΔFRN, calculated as the difference waveform between win and loss feedback for each task. (C) Topographies of the difference wave between loss and win feedback for each task 200–400 msec after feedback onset.

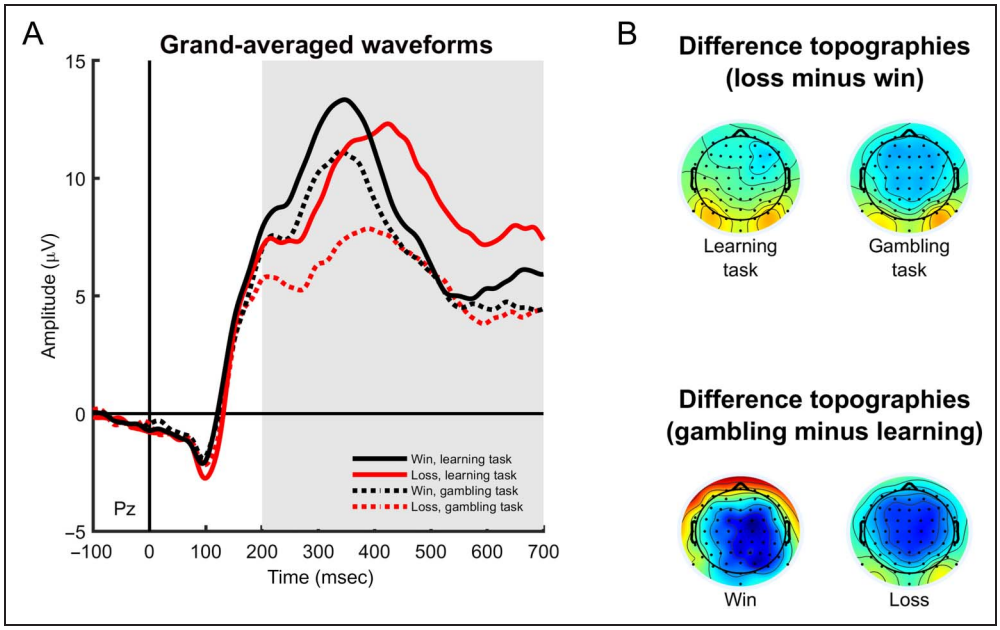

Figure 4. (A) Feedback-locked grand-averaged waveforms at electrode Pz. The gray area indicates the time window for the maximum peak amplitude analysis. (B) Topographies of the peak differences between loss and win feedback for each task. (C) Topographies of the peak differences between the learning and gambling tasks for each feedback. Peak (difference) topographies are compiled by plotting the amplitudes of all electrodes at the latency of the peak amplitudes of electrode Pz, separately for each condition.

of 200–700 msec at electrode Pz, determined individually for each task and participant.[3] This method also allows for analyzing possible latency shifts across tasks, which appear to be evident in Figure 4A. For the analysis of peak amplitudes, we found a marginally significant main effect of Outcome, $F(1, 29) = 4.16$, $p = .051$, $\eta_p = .13$, and a significant main effect of Task, $F(1, 29) = 27.09$, $p < .001$, $\eta_p = .48$, but these effects were qualified by a significant interaction between both variables, $F(1, 29) = 6.34$, $p = .018$, $\eta_p = .18$. Win feedback showed a larger P3 than loss feedback in the gambling task, $t(29) = 3.06$, $p = .005$, but no such difference was obtained in the learning task, $t(29) = 0.21$, $p = .837$. Notably, however, the P3 was larger in the learning task than the gambling task for both win feedback, $t(29) = 3.19$, $p = .003$, and loss feedback, $t(29) = 5.62$, $p < .001$. Although Figure 4A suggests that the P3 peaks for loss feedback occurred slightly later than the peaks for win feedback, no significant effects were obtained for the analysis of peak latencies (all ps $> .16$).

Taken together, these results show that learnability influences feedback processing on the neural level. However, this effect differed according to the stage of feedback processing. Although the amplitude was generally more negative for the gambling task in the time window of the FRN, the ΔFRN (i.e., the increased negativity for losses relative to wins), which has been associated with reinforcement learning and reward prediction error calculation, was not modulated by learnability. In contrast, the P3 was generally larger for the learning task but showed an effect of feedback valence only for the gambling task.

After showing the impact of learnability on feedback processing, we also included Subsequent Choice as a variable for the ANOVA in an exploratory analysis to investigate the relationship between feedback-related brain activity and subsequent behavior. Subsequent choice codes for whether the upcoming choice for the same stimulus pair is the same (stay) or different (switch) to the current trial, thus linking neural feedback processing with WSLS behavior. To analyze a sufficient number of trials (>10), we had to exclude three participants from this additional step. For the time window of the FRN, we found similar effects of Outcome and Task as previously reported. In addition, the interaction between Subsequent Choice and Task condition was significant, $F(1, 26) = 9.23$, $p = .005$, $\eta_p = .26$. Subsequent switching behavior was associated with stronger positive deflections compared to stay behavior only in the learning task, $t(26) = 2.68$, $p = .013$, but not in the gambling task, $t(26) = 0.37$, $p = .716$. Whereas amplitudes in the time window of the FRN revealed only a marginally significant interaction between all three involved factors, $F(1, 26) = 3.00$, $p = .095$, $\eta_p = .10$, amplitudes in the P3 time window clearly indicate the same interactive effect, $F(1, 26) = 7.75$, $p = .001$, $\eta_p = .23$. Only for the learning task, there was a significant interaction between Subsequent Choice and Outcome, $F(1, 26) = 9.56$, $p = .005$, $\eta_p = .27$. Only for wins, subsequent switching ($M = 17.46$ μV, $SEM = 0.67$) indicates higher P3 amplitudes compared to stay behavior ($M = 14.56$ μV, $SEM = 0.50$), $t(26) = 4.10$, $p < .001$. For losses, subsequent switching ($M = 15.12$ μV, $SEM = 0.55$) was not different from stay behavior ($M = 14.90$ μV, $SEM = 0.68$), $t(26) = 0.34$, $p = .735$. There was no comparable effect for the gambling task, $F(1, 26) = 0.12$, $p = .732$, $\eta_p = .00$, indicating that there is no link between subsequent choice and P3 amplitude for the gambling task. In conclusion, the results from this exploratory analysis reveal that subsequent choice can be predicted from neural activity after feedback processing. More specifically, following the idea of WSLS, it seems that P3 reflects the overcoming of such behavioral tendencies in the learning task (Daw et al.,

2006), thus suggesting a link between activity in the P3 window and behavioral adaptation that goes beyond a simple TD account of behavior.

Model-based Analysis

After finding evidence for an influence of learnability on choice behavior and the neural correlates of feedback processing, we applied model-based analysis to reveal the functional mechanisms underlying these effects. In comparison to the previously reported analyses, computational modeling allows us to investigate the mechanistic underpinnings of decision-making and learning by establishing a formally explicit link between model variables and behavioral and neural activity. On the basis of the assumption that this link resembles evidence for TD, we hypothesized to find modulations of learnability, leading to separable patterns in the two tasks.

To validate the identifiability of the proposed mechanisms, we first conducted a model recovery procedure during which we simulated behavioral data from 100 participants for each of the six candidate models. At their core, these models share the rationale for learning but use different variable parameters to explain differences between the learning and gambling tasks. Variable learning rate models assume that instructions on learning and gambling tasks lead to distinct learning rates, variable inverse temperature models assume that the instruction drives distinct inverse temperature, and variable policy mixture models assume that behavior can be captured as a mixture of a pure learning and a pure guessing policy. Simulated data from these models were then subjected to model selection. Bayesian model comparison via PXP revealed that our models were highly recoverable from the simulated data, as shown in Figure 5. Interestingly, the BIC and AIC metrics were unable to recover the models that generated the simulated data, which appears to reflect a bias toward simple models because of the strong weighting of model complexity in these measures. Taken together, these results suggest that model recovery can be subject to considerable biases arising from the specific comparison metrics. Given these findings, we conclude that, for the present task and models, PXP seems to be the most reliable metric for model comparison.

Afterward, we fitted the six candidate models to the empirical data. Table 2 shows BIC, AIC, and PXP for each model. Although the variable policy mixture model with the full updating mechanism has the highest PXP (37%), the overall pattern (see Table 2) indicates that models are similarly prevalent in the population. In line with our model recovery findings, our interpretation based on PXP strongly deviates from the BIC and AIC metrics, which clearly identify the variable learning rate model with full updating as the best model. Although this finding from BIC and AIC seems to have a straightforward interpretation, the previous model recovery procedure already pointed toward a considerable shortcoming of those metrics, namely, the confusion between models involved and the preference for simple compared to complex model. Although BIC and AIC are still widely used in the literature, their application is increasingly criticized, especially with the emergence of alternative metrics that remedy some common shortcomings (e.g., Rigoux et al., 2014; Stephan, Penny, Daunizeau, Moran, & Friston, 2009).

One of those common shortcomings is that BIC and AIC must be interpreted as fixed-effects metrics; that is, only one model accounts for all the data simultaneously. On the other hand, Bayesian model selection, as employed in this study, derives random effects metrics that imply that different models can account for data from

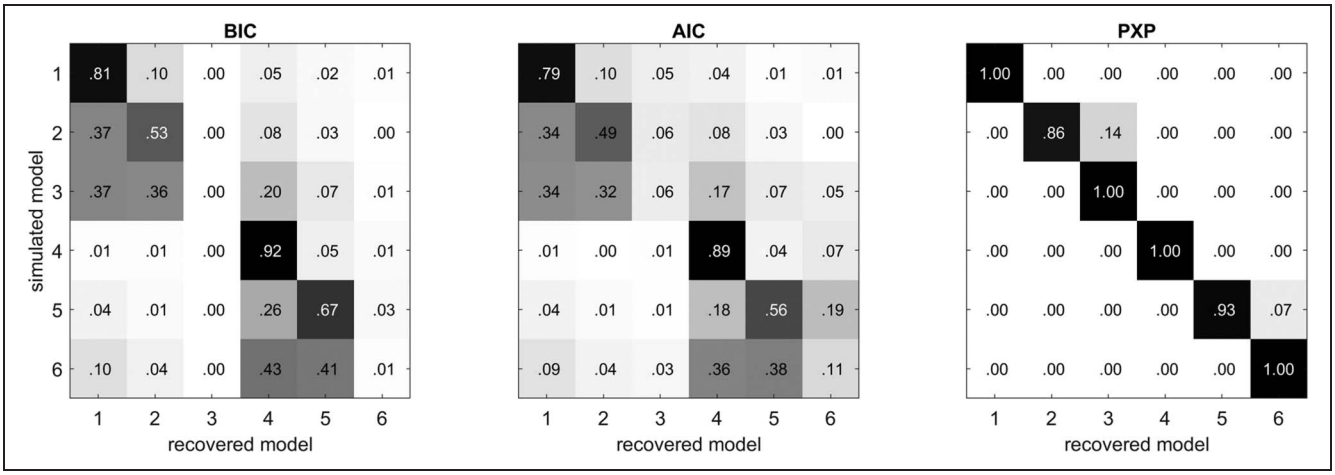

Figure 5. Results of the model recovery procedure. Confusion matrices, separately for the AIC, BIC, and PXP metrics. Values within each row of the confusion matrix sum to 1. For each model, 200 data sets (participants) were simulated. For AIC and BIC, values within each cell indicate the proportion of the simulated data sets recovered as the winning model. For PXP, values indicate the probability that the respective model is more frequent than all other possible models, given equal prevalence in the population. Optimal recovery would yield the identity matrix. Variable learning rate models = 1, 4; variable inverse temperature model = 2, 5; mixture policy model = 3, 6; partial updating = 1–3; and full updating = 4–6.

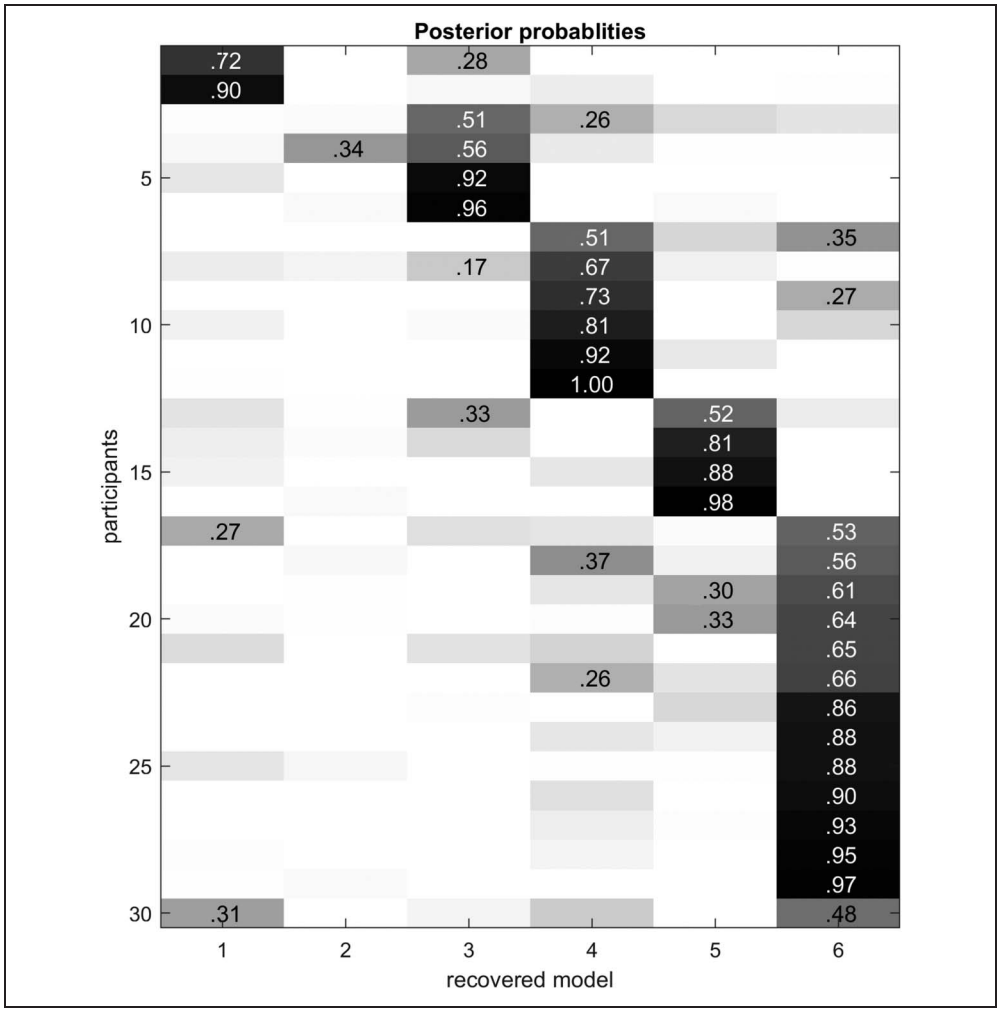

Figure 6. Posterior probabilities for the empirical data, sorted according to the highest probability per model. Variable learning rate models = 1, 4; variable inverse temperature model = 2, 5; mixture policy model = 3, 6; partial updating = 1–3; and full updating = 4–6.

different participants. In line with this assumption and on the basis of the PXP, we extracted for each participant the posterior belief that a specific model generated its data. As shown in Figure 6, the dominating model varied considerably between participants. In line with the finding that PXP did not identify a winning model on the group level, this further supports the existence of strong inter-individual differences in learning strategies.

In a next step, we compared the estimated values of the different parameters between the learning and gambling tasks (Table 3). We found that the variable parameters for all models are significantly different in the

Table 3. Mean Estimated Model Parameters

	Full Updating			*Partial Updating*		
Parameter	α	β	ω	α	β	ω
α_{LEARN}	**0.42** (±0.28)***	0.29 (±0.24)	0.44 (±0.27)	**0.29** (±0.22)***	0.21 (±0.18)	0.32 (±0.23)
α_{GAMB}	**0.10** (±0.17)			**0.07** (±0.15)		
β_{LEARN}	1.29 (±0.89)	**1.67** (±1.18)***	1.46 (±0.97)	1.10 (±0.64)	**1.39** (±0.87)***	1.28 (±0.72)
β_{GAMB}		**0.65** (±0.32)			**0.60** (±0.26)	
ω_{LEARN}	—	—	**0.82** (±0.20)***	—	—	**0.80** (±0.20)***
ω_{GAMB}			**0.23** (±0.23)			**0.21** (±0.23)
p	0.33 (±0.26)	0.33 (±0.26)	0.34 (±0.27)	0.35 (±0.25)	0.34 (±0.25)	0.35 (±0.25)

Values represent the mean, and those in brackets indicate the standard deviation; boldfaced values were estimated separately for the learning and gambling tasks and compared using paired t test.

*** $p < .001$.

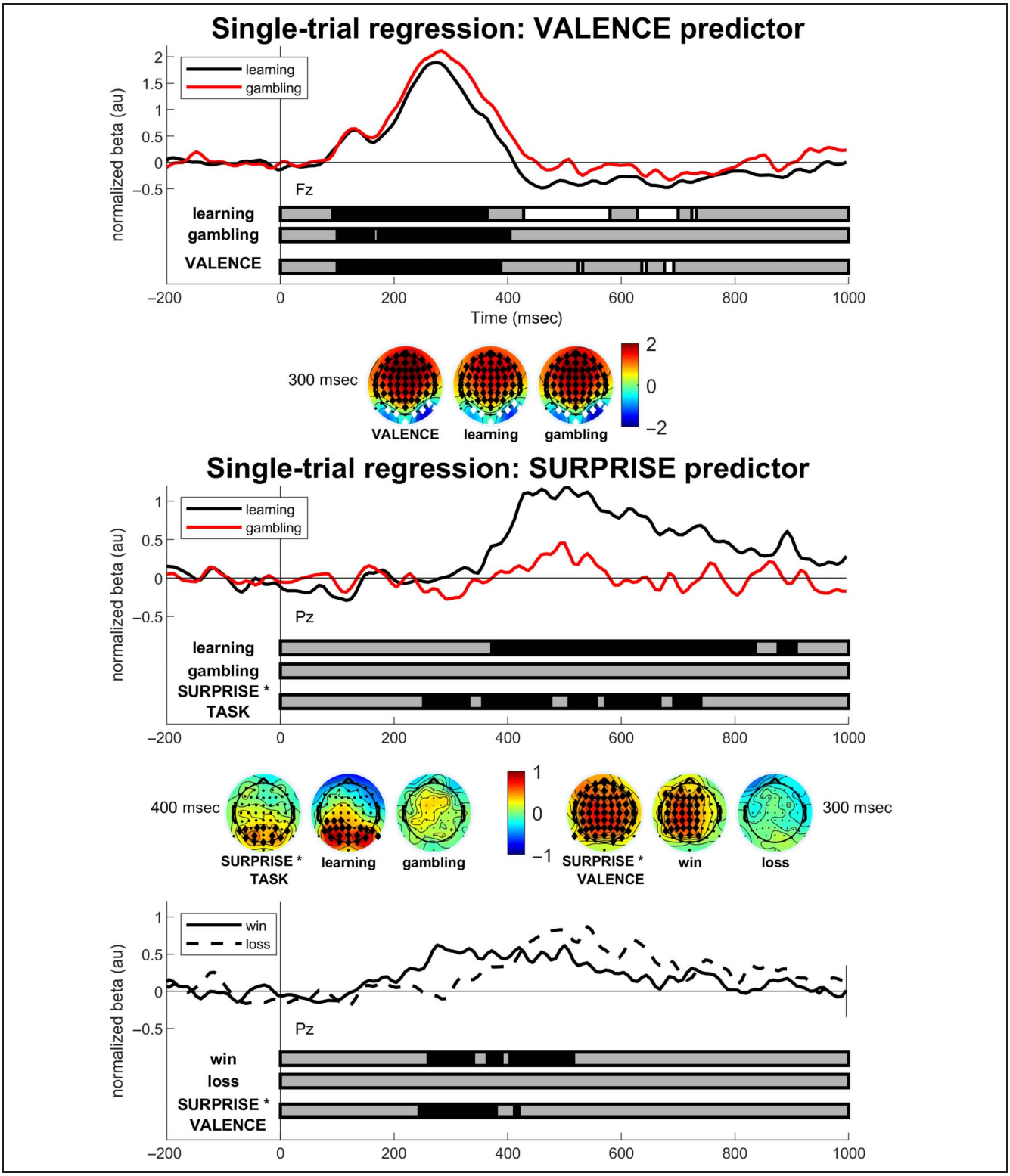

Figure 7. Feedback-locked single-trial regression for the variable mixture policy model with full updating. (A) Normalized regression (beta) values for the valence predictor at electrode site FCz. Gray bars below indicate the time windows that were considered for cluster-based permutation testing. Black bars indicate time windows with significant positive effects, and white bars indicate significant negative effects. "Learning" and "gambling" refer to beta values obtained for the post hoc regression analyses separately for each task. (B) Normalized regression values for the Surprise × Task interaction at electrode site Pz. (C) Normalized regression values for the Surprise × Valence interaction at electrode site Pz. For the topographies of the respective regression weights, black diamonds indicate significant positive clusters. White diamonds indicate significant negative clusters.

learning task compared to the gambling task. Crucially, RL-basedness for the winning mixture model with full updating was higher in learning tasks compared to gambling tasks, $t(29) = 11.88$, $p < .001$, $d = 2.73$. In line with our hypothesis, this indicates that the dominant driving force for behavior in the learning task was action selection under the (reinforcement) learning policy, whereas the dominant driving force for behavior in the gambling task was action selection under the guessing policy.

In a final step, we used the estimated PE from our model to investigate whether valence and surprise processing is modulated by learnability. To this end, we simulated these single-trial values for each model and participant. To account for interindividual differences, we calculated the weighted average PE across models for each participant from the posterior probabilities and correlated these values with neural activity using multiple linear regression analyses. We hypothesized to find a stronger connection between model estimates and neural data in the learning task than in the gambling task for those aspects of the PE that are influenced by learnability.

As the model-based regressor for valence perfectly coincides with the outcome variable in the ERP analyses, we expected to find similar modulations in our single-trial analysis. In line with this hypothesis, we found two significant clusters that distinguish between wins and losses (Figure 7). First, there was a positive correlation between valence and neural activity at frontocentral electrode sites ranging between 92 and 452 msec ($p < .001$). Second, there was a sustained negative correlation at posterior (occipital) electrode sites ranging between 60 and 1000 msec ($p = .009$). Crucially, these valence-driven neural patterns were not modulated between tasks ($p > .083$), suggesting that learnability does not modulate the processing of valence information.

In contrast to valence information, surprise processing was clearly modulated between the two tasks, as indicated by a significant interaction between Task and Surprise at posterior electrode sites in a time window between 244 and 748 msec ($p = .036$). On the basis of the visual inspection of the topographies, we selected the time course at electrode site Pz, where the significant positive correlation in the learning task ranged between 372 and 908 msec ($p = .001$). For the gambling task, no comparable correlation was evident (all ps $\geq .592$). In line with our hypothesis, this indicates that learnability impacts TD learning on the neural level, leading to a suppression of surprise processing if the task is not learnable, as is the case in the gambling task.

Interestingly, we also found a significant interaction between Valence and Surprise ($p < .008$), suggesting that these two aspects of the PE are interrelated on the neural level. The effect covered from 212 to 1000 msec but was most pronounced over central electrode sites at around 300 msec. Whereas the correlation between surprise and neural activity was evident when valence was positive ($p = .021$), no such correlation was evident for negative valence ($p = .060$).

Please note that the qualitative and quantitative patterns of the single-trial regression from the Laplace-informed posterior probabilities can be replicated for each individual model as well as the BIC-informed posterior probabilities. In line with suggestions from the literature (Wilson & Niv, 2015), model-based neural results are often robust across a wide variety of computational models and not specifically conditioned on the choices for computational modeling.

DISCUSSION

TD learning revolves around the concept of a PE, which carries information about the valence and surprise of a specific outcome. In this study, we investigated the influence of learnability on the neural underpinnings of these constituent aspects of the PE by contrasting a learning variant and a gambling variant of a simple two-armed bandit task. The learnable structure of the learning task was realized as a learnable link between actions and subsequent feedback, whereas the random structure of the gambling task was realized as the absence of such a learnable link. On the basis of the idea of a goal-directed agent, we hypothesized that the difference in learnability should lead to a corresponding modulation of the behavioral and neural patterns of (TD) learning, and we asked which level of processing could be affected by this modulation.

Using computational modeling, we successfully extracted the neural footprints of both valence and surprise information in our EEG data. On the one hand, information about the surprise of an outcome was reflected in a pronounced central cluster. Because of its spatial and temporal distribution, this effect cannot be easily assigned to a specific component of the human ERPs. Crucially, however, differences between task conditions were evident over posterior sites that strongly resemble the P3. As indexed by the absence of any correlation between surprise and neural activity in the gambling task, its expression appeared to be largely suppressed when feedback processing was random and thus irrelevant for behavioral adaptation. On the other hand, information about the valence of an outcome was reflected by an early frontocentral cluster of activation that is spatially and temporally very similar to the ΔFRN. In contrast to surprise, the manipulation of learnability had no influence on valence calculation. The fully random feedback in the gambling task led to the same valence effect as the feedback in the learning task, although learning was strongly reduced in the former. In summary, these findings support the idea that the different aspects of the PE are differentially modulated by the learnability of the environment. Whereas early processing of valence in the time range of the FRN remains intact across the two task conditions, subsequent processing of surprise could be suppressed for environments with a random structure, that is, during the gambling task.

Interestingly, our results of the single-trial analysis also contribute to answering the question on how information about valence and surprise interact on the neural level. In the cognitive neuroscience literature, there is an ongoing debate on this issue. Following the earliest account of the FRN (Holroyd & Coles, 2002), activity in the time window of the FRN should reflect surprise only for negative outcomes (negative and signed PE). On the basis of new evidence, which showed that the neural response to feedback was mainly driven by positive feedback, this account was subsequently altered and updated (Holroyd et al., 2009; Holroyd, Pakzad-Vaezi, & Krigolson, 2008). Now, a reward positivity is assumed to be elicited after positive outcomes (positive and signed PE). Alternatively, others have suggested that PEs are reflected in the brain irrespective of the valence of the outcome (unsigned PE or surprise; Alexander & Brown, 2011, 2015). Although evidence from a meta-analysis suggests that the FRN is sensitive to an unsigned PE (Sambrook & Goslin, 2015), only few studies actually estimated PEs using computational modeling (Burnside et al., 2019; Sambrook et al., 2018; Fischer & Ullsperger, 2013; Walsh & Anderson, 2011). In our study, surprise was reflected in the neural data more strongly for positive than negative outcomes, thus adding to the idea of a positive and signed PE signal in the brain, which however is not reflected in an FRN or reward positivity, but rather in a more posterior brain activity.

Regarding our central manipulation of learnability, the analysis of ERPs revealed a pattern of results that showed not only commonalities but also differences as compared to that of the model-based analysis. Although the overall amplitude in the FRN time window was influenced by learnability, the ΔFRN (i.e., the increased negativity for negative feedback as compared to positive feedback) did not differ between the learning task and the gambling task, suggesting that the frontocentral cluster reflecting valence information in the model-based analysis is linked to the ΔFRN. In contrast to the FRN, P3 amplitudes revealed a modulation by learnability. Surprisingly, the difference in P3 amplitudes between win and loss outcomes was evident in the gambling task but was absent for the learning task. This pattern could reflect that, because of the yoking of feedback sequences, positive feedback was unexpectedly more frequent in the gambling task, thus inducing an additional expectancy effect in the P3 data.

The comparison of results from the ERP and model-based analyses demonstrates that a direct mapping of ERP components to distinct (cognitive) processes is insufficient and premature. The observation that the gambling task shows a generally reduced P3 (possibly reflecting reduced learning) but an increased P3 valence effect suggests that more than one process is reflected in this component. The same could be inferred from the differential effect of learnability on overall FRN amplitudes in general and on the ΔFRN in particular. Additional leverage for the idea of multiple parallel neural processes reflected in the distinct ERP components comes from our exploratory analysis, suggesting that the P3 reflects an overcoming of behavioral tendencies imposed by TD learning. Independent support for this idea comes from different lines of research. For example, a recent study showed that whereas the link between the FRN and behavioral adaptation was driven by reinforcement learning, the link between P3 and behavioral adaptation was triggered by explicit rules that go beyond patterns derived from reinforcement learning (Chase, Swainson, Durham, Benham, & Cools, 2011). On the basis of the idea that action selection involves switching between exploitative and explorative modes (Daw et al., 2006), the P3 could provide a neural correlate for a switch toward explorative behavior.

On a methodological level, this study highlights a crucial advantage of model-based analyses, which enabled us to extract functionally meaningful activity from the neural data by combining the strength of multiple methods (computational modeling, regression, and cluster-based permutation testing). Computational modeling is particularly beneficial when theoretical variables cannot be controlled sufficiently by experimental design, as is the case with surprise in volatile environments and learning in general. Even under such challenging conditions, we were thus able to derive predictions about neural patterns. In contrast to the discrete nature of the traditional ERP approach, the use of a regression approach in our study allowed us to refrain from splitting surprise into bins (e.g., median or mean) but analyze it instead as a continuous variable. As the regression method has gained popularity over the last few years, there is now not only a huge body of literature for further discussion but also comprehensive introduction of this matter (Wilson & Niv, 2015; Mars et al., 2012; Gläscher & O'Doherty, 2010; O'Doherty, Hampton, & Kim, 2007). Whereas the traditional ERP approach always relies on ad-hoc defined components that often consist of a mixture of different activities (e.g., theta band and delta band in the FRN time window; Bernat, Nelson, & Baskin-Sommers, 2015; Foti, Weinberg, Bernat, & Proudfit, 2015; Harper, Malone, & Bernat, 2014) and are often a subject to debates regarding their quantification and functional meaning (Krigolson, 2018; Picton et al., 2000), cluster-based permutation testing is not bound to strong constraints in both temporal and spatial domains, making it a valuable tool for both confirmatory and explorative analyses (for a helpful discussion on the advantages and pitfalls, see Sassenhagen & Draschkow, 2019; Maris & Oostenveld, 2007). Although each of the methods described above makes a specific contribution to the main question under scrutiny, only their combination allowed us to gain novel and critical insights into feedback processing in the human brain, which would not have been possible using traditional methods. Hence, their combination allowed us to foster our understanding of the putative underpinnings of decision-making and learning.

Our study design and results are complemented by a recent study contrasting feedback processing under an

active and observational learning condition (Burnside et al., 2019). Using model-based single-trial regression analyses, the authors showed that neural activity in the time window of the FRN is driven by a PE and is clearly modulated between task conditions. Neural activity in the time window of the P3, however, is driven only by outcome and is not modulated between task conditions. Although these results are in stark contrast to the new findings reported above, it is noteworthy to take a closer look at the paradigm used in that previous study and underlying these effects. In the active condition, participants played a probabilistic but stationary bandit task, whereas in the observational condition, they merely watched another person. Whereas the active condition shares some similarities with our learning condition, outcome information in the observation condition is orthogonally different to our gambling condition. On the one hand, observational learning still necessitates learning from (observed) action, but without the outcome being relevant for the agent's goal achievement. Gambling, on the other hand, does not necessitate learning, although the outcome is still relevant for goal achievement. These differences in task design are mirrored on the neural level and indicate the impact of top–down processing selectively modulating TD learning based on environmental demands.

The question emerges whether, rather than the selective suppression of surprise processing, differences in visual attention to stimuli and feedback could have produced the present results. Recent studies highlighted the importance of attention for reinforcement learning using a multidimensional bandit task, in which only one of multiple stimulus dimensions was relevant for feedback (Leong, Radulescu, Daniel, DeWoskin, & Niv, 2017; Niv et al., 2015). Critically, however, they found attention to unselectively constrain reinforcement learning, whereas only the surprise processing was selectively modulated in our unidimensional bandit task. This could suggest that our results do not primarily reflect differences in attention between learning and gambling tasks. In fact, our task design, especially the implementation of catch trials, was chosen to control for attentional differences between tasks. Although attention to feedback was slightly weaker in the gambling task, as indicated by both catch trial accuracy and subjective ratings, the same measures also showed that attention was clearly allocated to feedback even in the gambling task, and ERP data as well as model-based analysis revealed that the effect of valence was identical between tasks. The interpretation that attention plays only a minor role for our results is further supported by our computational modeling results. In summary, our findings preclude a simple alternative explanation by mere attentional differences.

The accuracy of model-based analyses strongly relies on multiple factors such as the model selection procedure and the validity of both the applied computational model and experimental manipulation (Palminteri, Wyart, & Koechlin, 2017; Nassar & Frank, 2016; Wilson & Niv, 2015; Rigoux et al., 2014; Nassar & Gold, 2013; Stephan et al., 2009). Therefore, we first evaluated the identifiability of our computational models using a model recovery procedure. This procedure indicated that recoverability differed considerably across metrics. Consequently, we relied model comparison on the metric that is most convincing given the model recovery procedure (i.e., the PXP). Independent of model comparison, parameter estimates from each model revealed a significant difference between instructed task types, further validating our experimental manipulation.

The obtained context-dependent arbitration between policies/parameters is highly adaptive. In the learnable environment, behavior was dominated by reinforcement learning principles and showed clear hallmarks of behavioral adaptation. Otherwise, in the random environment, behavior was dominated by stochastic choice (although some behavioral adaptation was still observable). Irrespective of the true underlying mechanism (decreased learning rate, increased temperature, or decreased RL-basedness), stochastic choice is often implemented in computational models of learning. Its purpose is mostly to serve as a baseline or dummy condition that is compared to biologically plausible choice policies (e.g., Doll, Duncan, Simon, Shohamy, & Daw, 2015; Steingroever, Wetzels, & Wagenmakers, 2014; Worthy & Maddox, 2014; Worthy, Hawthorne, & Otto, 2013). The present findings however suggest that stochastic choice behavior can be utilized in a goal-directed way as well. In line with recent animal findings (Tervo et al., 2014) and on the basis of the idea of a cost–benefit arbitration (Kool, Gershman, & Cushman, 2017, 2018), we suggest that stochastic choice behavior maximizes outcomes, while simultaneously minimizing computational costs, for example, by withholding resources budgeted for the processing of surprise or the updating of expectations.

The idea that learnability exerts its influence by altering the trade-off between alternative internal task representations/parameters is compatible with suggestions in the existing literature. It has already been assumed that learning tasks are characterized by a reliance on external feedback, whereas gambling tasks are characterized by a reliance on internal expectations (Holroyd et al., 2009; Hajcak, Moser, Holroyd, & Simons, 2007; Holroyd, Hajcak, & Larsen, 2006). Similar ideas received attention also within other domains of reinforcement learning research (Daw & O'Doherty, 2013; Niv, 2009; Dayan & Niv, 2008). A frequent assumption is that reinforcement learning is achieved by the interplay between a model-free system that learns from experience and a model-based system in which internal task models drive behavior in a more top–down way. According to one proposal, arbitration between these different systems for decision-making is based on the relative uncertainty of their respective estimates with the less uncertain estimate predominantly driving behavior (Daw, Niv, & Dayan, 2005). Although such uncertainty-based arbitration is implemented on a

trial-to-trial basis (Lee, Shimojo, & O'Doherty, 2014), related evidence suggests that instructions can influence the arbitration between learning systems via pFC loops that modulate reinforcement learning in the striatum (Doll, Hutchison, & Frank, 2011; Doll, Jacobs, Sanfey, & Frank, 2009). A similar neural pathway between pFC and the striatum may account for the putative top–down modulation of surprise in this study.

Taken together, our results show that task learnability can modulate reinforcement learning by means of two mechanisms. On the one hand, this modulation occurs via the selective suppression of surprise when a task is unlearnable, leaving the evaluation of valence unaffected. On the other hand, we suggest that behavioral adaptation on the basis of task learnability is driven by a flexible cost–benefit arbitration.

Acknowledgments

This work was supported by funding from the National Science Centre of Poland (2015/19/B/HS6/01259) and the Polish National Agency for Academic Exchange (Bekker Programme signature: PPN/BEK/2018/1/00257), awarded to W. W. Moreover, this work was supported by research grants (G024716N and G048119N) from the Research Foundation Flanders (FWO), awarded to G. P. and M. C. S. The authors declare no conflict of interests.

Reprint requests should be sent to Franz Wurm, Cognitive Psychology Unit, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands, or via e-mail: f.r.wurm@fsw.leidenuniv.nl.

Author Contributions

F. W., W. W., B. E., M. C. S., G. P., and M. S. designed the study; W. W. and M. C. S. collected the data; F. W. analyzed the data; and F. W., W. W., B. E., M. C. S., G. P., and M. S. wrote the article. All authors approved the final version of the article before submission.

Funding Information

Mario Carlo Severo and Gilles Pourtois: Fonds Wetenschappelijk Onderzoek (https://dx.doi.org/10.13039/501100003130), grant number: G024716N, G048119N. Wioleta Walentowska: Narodowa Agencja Wymiany Akademickiej (https://dx.doi.org/10.13039/501100014434), grant number: PPN/BEK/2018/1/00257. Wioleta Walentowska: National Science Centre of Poland, grant number: 2015/19/B/HS6/01259.

Diversity in Citation Practices

A retrospective analysis of the citations in every article published in this journal from 2010 to 2020 has revealed a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .408, W(oman)/M = .335, M/W = .108, and W/W = .149, the comparable proportions for the articles that these authorship teams cited were M/M = .579, W/M = .243, M/W = .102, and W/W = .076 (Fulvio et al., *JoCN*, 33:1, pp. 3–7). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance.

Notes

1. Please note that there is no clear consensus about the labeling of the FRN in the literature. To prevent confusion, we refer to the ΔFRN as the difference between positive and negative outcomes. For an insightful discussion on the naming and quantification of the FRN component, refer to a recent methodological review by Krigolson (2018).
2. In contrast to the Laplace-informed Bayesian model selection, BIC-informed Bayesian model selection exhibits similar biases in model recovery as the BIC and AIC metric, namely, favoring simpler models over complex models (in terms of number of free parameters).
3. Please note that a similar pattern of results was obtained when we quantified the P3 as mean amplitude in a time window between 200 and 500 msec at electrode Pz.

REFERENCES

Alexander, W. H., & Brown, J. W. (2011). Medial prefrontal cortex as an action–outcome predictor. *Nature Neuroscience*, *14*, 1338–1344. https://doi.org/10.1038/nn.2921, PubMed: 21926982

Alexander, W. H., & Brown, J. W. (2015). Hierarchical error representation: A computational model of anterior cingulate and dorsolateral prefrontal cortex. *Neural Computation*, *27*, 2354–2410. https://doi.org/10.1162/NECO_a_00779, PubMed: 26378874

Balleine, B. W. (2005). Neural bases of food-seeking: Affect, arousal and reward in corticostriatolimbic circuits. *Physiology & Behavior*, *86*, 717–730. https://doi.org/10.1016/j.physbeh.2005.08.061, PubMed: 16257019

Bell, A. J., & Sejnowski, T. J. (1995). An information–maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*, 1129–1159. https://doi.org/10.1162/neco.1995.7.6.1129, PubMed: 7584893

Bernat, E. M., Nelson, L. D., & Baskin-Sommers, A. R. (2015). Time–frequency theta and delta measures index separable components of feedback processing in a gambling task. *Psychophysiology*, *52*, 626–637. https://doi.org/10.1111/psyp.12390, PubMed: 25581491

Botvinick, M. M., Niv, Y., & Barto, A. G. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, *113*, 262–280. https://doi.org/10.1016/j.cognition.2008.08.011, PubMed: 18926527

Burnside, R., Fischer, A. G., & Ullsperger, M. (2019). The feedback-related negativity indexes prediction error in active but not observational learning. *Psychophysiology*, *56*, e13389. https://doi.org/10.1111/psyp.13389, PubMed: 31054155

Chase, H. W., Swainson, R., Durham, L., Benham, L., & Cools, R. (2011). Feedback-related negativity codes prediction error

but not behavioral adjustment during probabilistic reversal learning. *Journal of Cognitive Neuroscience*, *23*, 936–946. https://doi.org/10.1162/jocn.2010.21456, PubMed: 20146610
Cohen, M. X., & Ranganath, C. (2007). Reinforcement learning signals predict future decisions. *Journal of Neuroscience*, *27*, 371–378. https://doi.org/10.1523/JNEUROSCI.4421-06.2007, PubMed: 17215398
D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). BOLD responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, *319*, 1264–1267. https://doi.org/10.1126/science.1150605, PubMed: 18309087
Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In M. R. Delgado, E. A. Phelps, & T. W. Robbins (Eds.), *Decision making, affect, and learning: Attention & performance XXIII* (Vol. 23, pp. 3–38). New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199600434.003.0001
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. https://doi.org/10.1016/j.neuron.2011.02.027, PubMed: 21435563
Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*, 1704–1711. https://doi.org/10.1038/nn1560, PubMed: 16286932
Daw, N. D., & O'Doherty, J. P. (2013). Multiple systems for value learning. In P. W. Glimcher & E. Fehr (Eds.), *Neuroeconomics: Decision making and the brain* (2nd ed., pp. 393–410). San Diego, CA: Elsevier. https://doi.org/10.1016/B978-0-12-416008-8.00021-8
Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*, 876–879. https://doi.org/10.1038/nature04766, PubMed: 16778890
Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*, 185–196. https://doi.org/10.1016/j.conb.2008.08.003, PubMed: 18708140
Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009, PubMed: 15102499
Dickinson, A., & Balleine, B. W. (2002). The role of learning in the operation of motivational systems. In C. R. Gallistel (Ed.), *Steven's handbook of experimental psychology: Learning, motivation and emotion* (Vol. 3, pp. 497–534). New York: Wiley. https://doi.org/10.1002/0471214426.pas0312
Di Gregorio, F., Ernst, B., & Steinhauser, M. (2019). Differential effects of instructed and objective feedback reliability on feedback-related brain activity. *Psychophysiology*, *56*, e13399. https://doi.org/10.1111/psyp.13399, PubMed: 31131923
Doll, B. B., Duncan, K. D., Simon, D. A., Shohamy, D., & Daw, N. D. (2015). Model-based choices involve prospective neural activity. *Nature Neuroscience*, *18*, 767–772. https://doi.org/10.1038/nn.3981, PubMed: 25799041
Doll, B. B., Hutchison, K. E., & Frank, M. J. (2011). Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *Journal of Neuroscience*, *31*, 6188–6198. https://doi.org/10.1523/JNEUROSCI.6486-10.2011, PubMed: 21508242
Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research*, *1299*, 74–79. https://doi.org/10.1016/j.brainres.2009.07.007, PubMed: 19595993
Ernst, B., & Steinhauser, M. (2017). Top–down control over feedback processing: The probability of valid feedback affects feedback-related brain activity. *Brain and Cognition*, *115*, 33–40. https://doi.org/10.1016/j.bandc.2017.03.008, PubMed: 28407527
Ernst, B., & Steinhauser, M. (2018). Effects of feedback reliability on feedback-related brain activity: A feedback valuation account. *Cognitive, Affective, & Behavioral Neuroscience*, *18*, 596–608. https://doi.org/10.3758/s13415-018-0591-7, PubMed: 29626297
Fischer, A. G., & Ullsperger, M. (2013). Real and fictive outcomes are processed differently but converge on a common adaptive mechanism. *Neuron*, *79*, 1243–1255. https://doi.org/10.1016/j.neuron.2013.07.006, PubMed: 24050408
Foti, D., Weinberg, A., Bernat, E. M., & Proudfit, G. H. (2015). Anterior cingulate activity to monetary loss and basal ganglia activity to monetary gain uniquely contribute to the feedback negativity. *Clinical Neurophysiology*, *126*, 1338–1347. https://doi.org/10.1016/j.clinph.2014.08.025, PubMed: 25454338
Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *Neuroimage*, *34*, 220–234. https://doi.org/10.1016/j.neuroimage.2006.08.035, PubMed: 17055746
Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6. https://doi.org/10.1016/j.jmp.2016.01.006
Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2015). Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience*, *15*, 523–536. https://doi.org/10.3758/s13415-015-0347-6, PubMed: 25801925
Gläscher, J. P., & O'Doherty, J. P. (2010). Model-based approaches to neuroimaging: Combining reinforcement learning theory with fMRI data. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*, 501–510. https://doi.org/10.1002/wcs.57, PubMed: 26271497
Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, *48*, 1711–1725. https://doi.org/10.1111/j.1469-8986.2011.01273.x, PubMed: 21895683
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. https://doi.org/10.1111/j.1469-8986.2007.00567.x, PubMed: 17666029
Harper, J., Malone, S. M., & Bernat, E. M. (2014). Theta and delta band activity explain N2 and P3 ERP component activity in a go/no-go task. *Clinical Neurophysiology*, *125*, 124–132. https://doi.org/10.1016/j.clinph.2013.06.025, PubMed: 23891195
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. https://doi.org/10.1037/0033-295X.109.4.679, PubMed: 12374324
Holroyd, C. B., Hajcak, G., & Larsen, J. T. (2006). The good, the bad and the neutral: Electrophysiological responses to feedback stimuli. *Brain Research*, *1105*, 93–101. https://doi.org/10.1016/j.brainres.2005.12.015, PubMed: 16427615
Holroyd, C. B., Krigolson, O. E., Baker, R., Lee, S., & Gibson, J. (2009). When is an error not a prediction error? An electrophysiological investigation. *Cognitive, Affective, & Behavioral Neuroscience*, *9*, 59–70. https://doi.org/10.3758/CABN.9.1.59, PubMed: 19246327
Holroyd, C. B., Pakzad-Vaezi, K. L., & Krigolson, O. E. (2008). The feedback correct-related positivity: Sensitivity of the event-related brain potential to unexpected positive

feedback. *Psychophysiology*, *45*, 688–697. https://doi.org/10.1111/j.1469-8986.2008.00668.x, PubMed: 18513364
Jepma, M., Brown, S. B. R. E., Murphy, P. R., Koelewijn, S. C., de Vries, B., van den Maagdenberg, A. M., et al. (2018). Noradrenergic and cholinergic modulation of belief updating. *Journal of Cognitive Neuroscience*, *30*, 1803–1820. https://doi.org/10.1162/jocn_a_01317, PubMed: 30063180
Jepma, M., Murphy, P. R., Nassar, M. R., Rangel-Gomez, M., Meeter, M., & Nieuwenhuis, S. (2016). Catecholaminergic regulation of learning rate in a dynamic environment. *PLoS Computational Biology*, *12*, e1005171. https://doi.org/10.1371/journal.pcbi.1005171, PubMed: 27792728
Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Computational Biology*, *7*, e1002055. https://doi.org/10.1371/journal.pcbi.1002055, PubMed: 21637741
Kolossa, A., Kopp, B., & Fingscheidt, T. (2015). A computational analysis of the neural bases of Bayesian inference. *Neuroimage*, *106*, 222–237. https://doi.org/10.1016/j.neuroimage.2014.11.007, PubMed: 25462794
Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost–benefit arbitration between multiple reinforcement-learning systems. *Psychological Science*, *28*, 1321–1333. https://doi.org/10.1177/0956797617708288, PubMed: 28731839
Kool, W., Gershman, S. J., & Cushman, F. A. (2018). Planning complexity registers as a cost in metacontrol. *Journal of Cognitive Neuroscience*, *30*, 1391–1404. https://doi.org/10.1162/jocn_a_01263, PubMed: 29668390
Kopp, B., Seer, C., Lange, F., Kluytmans, A., Kolossa, A., Fingscheidt, T., et al. (2016). P300 amplitude variations, prior probabilities, and likelihoods: A Bayesian ERP study. *Cognitive, Affective, & Behavioral Neuroscience*, *16*, 911–928. https://doi.org/10.3758/s13415-016-0442-3, PubMed: 27406085
Krigolson, O. E. (2018). Event-related brain potentials and the study of reward processing: Methodological considerations. *International Journal of Psychophysiology*, *132*, 175–183. https://doi.org/10.1016/j.ijpsycho.2017.11.007, PubMed: 29154804
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. https://doi.org/10.1901/jeab.2005.110-04, PubMed: 16596980
Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience*, *35*, 287–308. https://doi.org/10.1146/annurev-neuro-062111-150512, PubMed: 22462543
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. https://doi.org/10.1016/j.neuron.2013.11.028, PubMed: 24507199
Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V., & Niv, Y. (2017). Dynamic interaction between reinforcement learning and attention in multidimensional environments. *Neuron*, *93*, 451–463. https://doi.org/10.1016/j.neuron.2016.12.040, PubMed: 28103483
Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*, 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024, PubMed: 17517438
Mars, R. B., Debener, S., Gladwin, T. E., Harrison, L. M., Haggard, P., Rothwell, J. C., et al. (2008). Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. *Journal of Neuroscience*, *28*, 12539–12545. https://doi.org/10.1523/JNEUROSCI.2925-08.2008, PubMed: 19020046
Mars, R. B., Shea, N. J., Kolling, N., & Rushworth, M. F. S. (2012). Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. *Quarterly Journal of Experimental Psychology*, *65*, 252–267. https://doi.org/10.1080/17470211003668272, PubMed: 20437297
Miltner, W. H. R., Braun, C. H., & Coles, M. G. H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a "generic" neural system for error detection. *Journal of Cognitive Neuroscience*, *9*, 788–798. https://doi.org/10.1162/jocn.1997.9.6.788, PubMed: 23964600
Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*, 1936–1947. https://doi.org/10.1523/JNEUROSCI.16-05-01936.1996, PubMed: 8774460
Nassar, M. R., Bruckner, R., & Frank, M. J. (2019). Statistical context dictates the relationship between feedback-related EEG signals and learning. *eLife*, *8*, e46975. https://doi.org/10.7554/eLife.46975, PubMed: 31433294
Nassar, M. R., & Frank, M. J. (2016). Taming the beast: Extracting generalizable knowledge from computational models of cognition. *Current Opinion in Behavioral Sciences*, *11*, 49–54. https://doi.org/10.1016/j.cobeha.2016.04.003, PubMed: 27574699
Nassar, M. R., & Gold, J. I. (2013). A healthy fear of the unknown: Perspectives on the interpretation of parameter fits from computational models in neuroscience. *PLoS Computational Biology*, *9*, e1003015. https://doi.org/10.1371/journal.pcbi.1003015, PubMed: 23592963
Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*, 139–154. https://doi.org/10.1016/j.jmp.2008.12.005
Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., et al. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*, 8145–8157. https://doi.org/10.1523/JNEUROSCI.2978-14.2015, PubMed: 26019331
O'Doherty, J. P., Hampton, A., & Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. *Annals of the New York Academy of Sciences*, *1104*, 35–53. https://doi.org/10.1196/annals.1390.022, PubMed: 17416921
Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Computational Biology*, *13*, e1005684. https://doi.org/10.1371/journal.pcbi.1005684, PubMed: 28800597
Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in Cognitive Sciences*, *21*, 425–433. https://doi.org/10.1016/j.tics.2017.03.011, PubMed: 28476348
Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*, 1042–1045. https://doi.org/10.1038/nature05051, PubMed: 16929307
Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Jr., et al. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology*, *37*, 127–152. https://doi.org/10.1111/1469-8986.3720127, PubMed: 10731765
Polich, J. (2020). 50+ years of P300: Where are we now? *Psychophysiology*, *57*, e13616. https://doi.org/10.1111/psyp.13616, PubMed: 32525221
Pontifex, M. B., Hillman, C. H., & Polich, J. (2009). Age, physical fitness, and attention: P3a and P3b. *Psychophysiology*, *46*, 379–387. https://doi.org/10.1111/j.1469-8986.2008.00782.x, PubMed: 19170947

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—Revisited. *Neuroimage*, *84*, 971–985. https://doi.org/10.1016/j.neuroimage.2013.08.065, PubMed: 24018303

Sailer, U., Fischmeister, F. P. S., & Bauer, H. (2010). Effects of learning on feedback-related brain potentials in a decision-making task. *Brain Research*, *1342*, 85–93. https://doi.org/10.1016/j.brainres.2010.04.051, PubMed: 20423704

Sambrook, T. D., & Goslin, J. (2014). Mediofrontal event-related potentials in response to positive, negative and unsigned prediction errors. *Neuropsychologia*, *61*, 1–10. https://doi.org/10.1016/j.neuropsychologia.2014.06.004, PubMed: 24946315

Sambrook, T. D., & Goslin, J. (2015). A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. *Psychological Bulletin*, *141*, 213–235. https://doi.org/10.1037/bul0000006, PubMed: 25495239

Sambrook, T. D., Hardwick, B., Wills, A. J., & Goslin, J. (2018). Model-free and model-based reward prediction errors in EEG. *Neuroimage*, *178*, 162–171. https://doi.org/10.1016/j.neuroimage.2018.05.023, PubMed: 29758337

San Martín, R. (2012). Event-related potential studies of outcome processing and feedback-guided learning. *Frontiers in Human Neuroscience*, *6*, 304. https://doi.org/10.3389/fnhum.2012.00304, PubMed: 23162451

Sassenhagen, J., & Draschkow, D. (2019). Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology*, *56*, e13335. https://doi.org/10.1111/psyp.13335, PubMed: 30657176

Schiffer, A.-M., Siletti, K., Waszak, F., & Yeung, N. (2017). Adaptive behaviour and feedback processing integrate experience and instruction in reinforcement learning. *Neuroimage*, *146*, 626–641. https://doi.org/10.1016/j.neuroimage.2016.08.057, PubMed: 27577720

Schonberg, T., O'Doherty, J. P., Joel, D., Inzelberg, R., Segev, Y., & Daw, N. D. (2010). Selective impairment of prediction error signaling in human dorsolateral but not ventral striatum in Parkinson's disease patients: Evidence from a model-based fMRI study. *Neuroimage*, *49*, 772–781. https://doi.org/10.1016/j.neuroimage.2009.08.011, PubMed: 19682583

Schultz, W. (2016). Dopamine reward prediction error coding. *Dialogues in Clinical Neuroscience*, *18*, 23–32. https://doi.org/10.31887/DCNS.2016.18.1/wschultz, PubMed: 27069377

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599. https://doi.org/10.1126/science.275.5306.1593, PubMed: 9054347

Seer, C., Lange, F., Boos, M., Dengler, R., & Kopp, B. (2016). Prior probabilities modulate cortical surprise responses: A study of event-related potentials. *Brain and Cognition*, *106*, 78–89. https://doi.org/10.1016/j.bandc.2016.04.011, PubMed: 27266394

Severo, M. C., Paul, K., Walentowska, W., Moors, A., & Pourtois, G. (2020). Neurophysiological evidence for evaluative feedback processing depending on goal relevance. *Neuroimage*, *215*, 116857. https://doi.org/10.1016/j.neuroimage.2020.116857, PubMed: 32304885

Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2014). Absolute performance of reinforcement-learning models for the Iowa Gambling Task. *Decision*, *1*, 161–183. https://doi.org/10.1037/dec0000005

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, *46*, 1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025, PubMed: 19306932

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). Cambridge, MA: MIT Press.

Szekely, A., D'Amico, S., Devescovi, A., Federmeier, K., Herron, D., Iyer, G., et al. (2005). Timed action and object naming. *Cortex*, *41*, 7–25. https://doi.org/10.1016/S0010-9452(08)70174-6

Tervo, D. G. R., Proskurin, M., Manakov, M., Kabra, M., Vollmer, A., Branson, K., et al. (2014). Behavioral variability through stochastic choice and its gating by anterior cingulate cortex. *Cell*, *159*, 21–32. https://doi.org/10.1016/j.cell.2014.08.037, PubMed: 25259917

Walentowska, W., Moors, A., Paul, K., & Pourtois, G. (2016). Goal relevance influences performance monitoring at the level of the FRN and P3 components. *Psychophysiology*, *53*, 1020–1033. https://doi.org/10.1111/psyp.12651, PubMed: 27091565

Walsh, M. M., & Anderson, J. R. (2011). Modulation of the feedback-related negativity by instruction and experience. *Proceedings of the National Academy of Sciences, U.S.A.*, *108*, 19048–19053. https://doi.org/10.1073/pnas.1117189108, PubMed: 22065792

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. https://doi.org/10.1016/j.neubiorev.2012.05.008, PubMed: 22683741

Wilson, R. C., & Niv, Y. (2015). Is model fitting necessary for model-based fMRI? *PLoS Computational Biology*, *11*, e1004237. https://doi.org/10.1371/journal.pcbi.1004237, PubMed: 26086934

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

Worthy, D. A., Hawthorne, M. J., & Otto, A. R. (2013). Heterogeneity of strategy use in the Iowa gambling task: A comparison of win–stay/lose–shift and reinforcement learning models. *Psychonomic Bulletin & Review*, *20*, 364–371. https://doi.org/10.3758/s13423-012-0324-9, PubMed: 23065763

Worthy, D. A., & Maddox, W. T. (2014). A comparison model of reinforcement-learning and win–stay–lose–shift decision-making processes: A tribute to W. K. Estes. *Journal of Mathematical Psychology*, *59*, 41–49. https://doi.org/10.1016/j.jmp.2013.10.001, PubMed: 25214675

Wurm, F., Ernst, B., & Steinhauser, M. (2020). The influence of internal models on feedback-related brain activity. *Cognitive, Affective, & Behavioral Neuroscience*, *20*, 1070–1089. https://doi.org/10.3758/s13415-020-00820-6, PubMed: 32812148