How the Level of Reward Awareness Changes the Computational and Electrophysiological Signatures of Reinforcement Learning

Correa, C.M.C.; Noorman, S.; Jiang, J.; Palminteri, S.; Cohen, M.X.; Lebreton, M.; van Gaal, S.

DOI
10.1523/JNEUROSCI.0457-18.2018

Publication date
2018

Document Version
Final published version

Published in
The Journal of Neuroscience

Behavioral/Cognitive

How the Level of Reward Awareness Changes the Computational and Electrophysiological Signatures of Reinforcement Learning

Camile M.C. Correa,[1] Samuel Noorman,[1] Jun Jiang,[3] Stefano Palminteri,[4,5,6] Michael X. Cohen,[7] Maël Lebreton,[2,8]* and Simon van Gaal[1,2,9]*

[1]Department of Psychology, University of Amsterdam, 1018 WT, Amsterdam, The Netherlands, [2]Amsterdam Brain and Cognition (ABC), University of Amsterdam, 1001 NK, Amsterdam, The Netherlands, [3]Department of Basic Psychology, School of Psychology, Third Military Medical University, Chongqing, People's Republic of China, [4]Département d'Études Cognitives, École Normale Supérieure, 75005 Paris, France, [5]Laboratoire de Neurosciences Cognitives, Institut National de la Santé et de la Recherche Médicale, 75005 Paris, France, [6]Université de Recherche Paris Sciences et Lettres, 75006, Paris, France, [7]Radboud University Medical Center, 6525 GA, Nijmegen, The Netherlands, [8]Center for Research in Experimental Economics and Political Decision Making, Amsterdam School of Economics, University of Amsterdam, 1001 NJ Amsterdam, The Netherlands, and [9]Donders Institute for Brain, Cognition and Behavior, Radboud University Nijmegen, 6500 HE, Amsterdam, The Netherlands

The extent to which subjective awareness influences reward processing, and thereby affects future decisions, is currently largely unknown. In the present report, we investigated this question in a reinforcement learning framework, combining perceptual masking, computational modeling, and electroencephalographic recordings (human male and female participants). Our results indicate that degrading the visibility of the reward decreased, without completely obliterating, the ability of participants to learn from outcomes, but concurrently increased their tendency to repeat previous choices. We dissociated electrophysiological signatures evoked by the reward-based learning processes from those elicited by the reward-independent repetition of previous choices and showed that these neural activities were significantly modulated by reward visibility. Overall, this report sheds new light on the neural computations underlying reward-based learning and decision-making and highlights that awareness is beneficial for the trial-by-trial adjustment of decision-making strategies.

Key words: consciousness; decision-making; prediction error; reinforcement learning

Significance Statement

The notion of reward is strongly associated with subjective evaluation, related to conscious processes such as "pleasure," "liking," and "wanting." Here we show that degrading reward visibility in a reinforcement learning task decreases, without completely obliterating, the ability of participants to learn from outcomes, but concurrently increases subjects' tendency to repeat previous choices. Electrophysiological recordings, in combination with computational modeling, show that neural activities were significantly modulated by reward visibility. Overall, we dissociate different neural computations underlying reward-based learning and decision-making, which highlights a beneficial role of reward awareness in adjusting decision-making strategies.

Introduction

How we make decisions depends strongly on the outcomes that have been previously associated with the available courses of action. Actions that often have been linked with rewards (e.g., food, money) are more likely to be repeated than actions that have not

Received Feb. 20, 2018; revised Sept. 18, 2018; accepted Sept. 20, 2018.

Author contributions: C.M.C.C., M.X.C., and S.v.G. designed research; C.M.C.C. and S.N. performed research; M.L. contributed unpublished reagents/analytic tools; C.M.C.C., S.N., J.J., S.P., M.L., and S.v.G. analyzed data; C.M.C.C., M.X.C., M.L., and S.v.G. wrote the paper.

This work was supported by grants from the Netherlands Organization for Scientific Research (NWO VENI 451-11-007) and the European Research Council (ERC starting grant, 715605, consciousness) awarded to S.v.G. M.L. is supported by the Netherlands Organization for Scientific Research (NWO VENI 451-15-015). C.M.C.C is supported by the Brazilian Science Without Borders program. JJ is supported by the National Natural Science Foundation of China (Grant No. 31600874).

*M.L. and S.v.G. are co-senior authors.

The authors declare no competing financial interests.

Correspondence should be addressed to either Simon van Gaal or Maël Lebreton, Nieuwe Achtergracht 129 B, REC G, Room G0.012, Ground floor, Postbus 15900, 1001 NK, Amsterdam, The Netherlands, E-mail: simonvangaal@gmail.com or mael.lebreton@gmail.com.

https://doi.org/10.1523/JNEUROSCI.0457-18.2018

 0270-6474/18/3810338-11$15.00/0

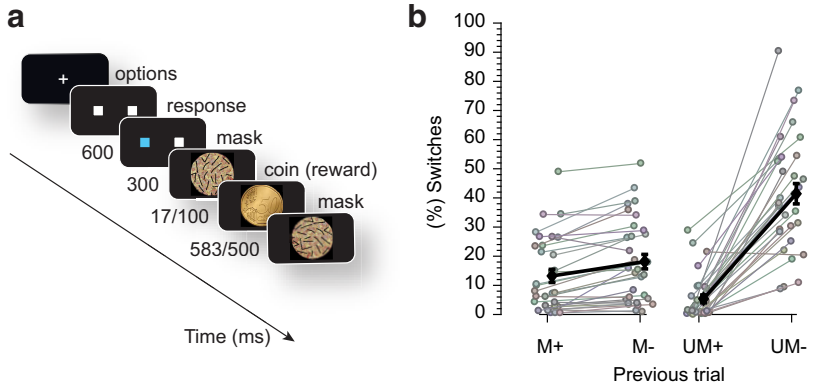

Figure 1. Experimental setup and behavior. ***a***, Two response options (white boxes on the left/right of fixation) were shown on the screen until a response was given. A correct response was rewarded with a 70% probability (50 cent coin) and not rewarded with a 30% probability (1 cent coin). Reward visibility was manipulated by masking. Unmasked (long coin presentation, short backward mask presentation) and masked (short coin presentation, long backward mask presentation) reward trials were mixed within blocks and randomly chosen across trials (each with a 50% probability). Which response option was most rewarded changed every 75–125 trials. ***b***, The percentage of switches, at the group level (in black) and for individual subjects (in gray) after specific trials. M: masked; UM: unmasked; +: reward; −: no-reward; error bars represent ± s.e.m.

been rewarded (or even punished; Dayan and Balleine, 2002; Berridge and Robinson, 2003; Rangel et al., 2008). Generally, the notion of reward is strongly associated with subjective evaluation, related to conscious processes such as "pleasure," "liking," and "wanting" (Berridge and Robinson, 2003). However, how human decision-making changes depending on reward awareness is unclear. Assessing how the level of awareness of information changes or may bias value-based learning and decision-making may prove critical to understanding apparent irrationality observed in human behavior (Kahneman, 2003; Evans, 2008; Weber and Johnson, 2009; Evans and Stanovich, 2013; Newell and Shanks, 2014).

Rewards have two fundamental roles in the decision-making process. First, in decision situations, expected rewards act as incentives, which determine choices and increase the amount of motor or cognitive effort one is willing to expend to reach a goal (Berridge, 2004; Schmidt et al., 2012). Second, after a decision has been enacted and the action effectuated, the obtained reward, or the absence of reward, drives important learning processes: successful actions are reinforced, while unsuccessful ones are discouraged (Sutton and Barto, 1998). Despite rewards being strongly associated with subjective feelings, notably with emotions and with the notion of expected pleasure (Berridge and Robinson, 2003), recent studies have reported that reward cues that are masked from awareness can still directly influence task performance (Pessiglione et al., 2007; Aarts et al., 2008; Bijleveld et al., 2012; Capa et al., 2013). These results suggest that the first role of reward information—incentivizing decision and effort production—may be processed outside the scope of awareness in the human brain to facilitate human performance (but for results challenging this view, see Bijleveld et al. 2014). On the other hand, little is known about whether and how the second role of rewards (i.e., the propensity to reinforce successful actions) is modulated by awareness.

To address this question, thirty-two participants performed a probabilistic reversal learning task in which we manipulated the visibility of reward using a standard masking technique. Participants were instructed to choose one of two response options, which led probabilistically either to a significant reward (a 50 cent coin, "reward condition") or a negligible one (a 1 cent coin, "no-reward condition"). Response–reward contingencies reversed several times over the course of the experiment, and participants were instructed to select the response option that was most often rewarded (Fig. 1*a*). Masked (M) and unmasked (UM) feedback were mixed within blocks to explore the relative weighting of both types of feedback. We combined EEG measurements with computational modeling to investigate, at the time of reward processing and on a trial-by-trial basis, the neural correlate of the different processes influencing participants' future choices and how those were affected by reward visibility. Thereby, the present work builds on previous studies that have linked reinforcement learning (RL) models to human neural data obtained from both fMRI and EEG measurements (Debener et al., 2005; O'Doherty et al., 2007; Daw et al., 2011; Fischer and Ullsperger, 2013; Hauser et al., 2014; Ullsperger et al., 2014; Fouragnan et al., 2017). In line with previous work, event-related potential (ERP) analyses focus on the feedback-related negativity (FRN) and the P3 component (Holroyd and Coles, 2002; Holroyd et al., 2003). The investigations on the EEG correlates of RL learning concentrate on the following three (computational) variables: the prediction error (signed PE), the level of surprise (unsigned PE), and the switch/stay behavior on the next trial (Cohen and Ranganath, 2007; Fischer and Ullsperger, 2013; Fouragnan et al., 2017; Collins and Frank, 2018). This approach allows us to investigate the impact of reward visibility on different cognitive processes involved in probabilistic reward-guided learning.

Materials and Methods

Participants

Thirty-two students from the University of Amsterdam (8 males, 24 females; mean ± SD age, 22.25 ± 3.1 years) participated in the experiment for course credits or financial compensation. All participants gave their written informed consent before participation, had normal or corrected-to-normal vision, and were naive to the purpose of the experiments. All procedures were executed in compliance with relevant laws and institutional guidelines and were approved by the local ethical committee of the University of Amsterdam.

Task

Stimuli were presented using Presentation software (Neurobehavioral Systems) against a black background at the center of a 20 inch VGA (video graphics array) monitor (frequency, 60 Hz), which was viewed by the participants from a distance of ~80 cm. Participants should fixate at the center of the screen and choose between a left or a right box 15 cm distant from each other by pressing a correspondent left or right chair button (parallel button).The chosen square was illuminated in blue for 600 ms, indicating the participants' response followed by a reward (a 50 cent coin) or a punishment (a 1 cent coin) that could be shown in a visible (100 ms) or masked (17 ms) way. Stimuli were used similarly to those by (Zedelius et al., 2012). A variable intertrial interval, 1500–2500 ms, separated each trial. If participants did not select a target after 1500 ms, a "too late!" message was displayed (Fig. 1*a*).

Sides were rewarded in a 70%/30% fashion. This probability condition was reversed several times during the 1200 trials so that, to decide advantageously, participants had to keep track of eventual "rule changes." We refer to the choices made on the 30% probability side as "incorrect choices," and those made according to the 70% rewarded side as "correct choices." Probabilities were fixed across trials within blocks, which lasted 75–125 trials. The block length had a minimum value, but it was dependent on how fast participants could learn the rule at stake. To assure that

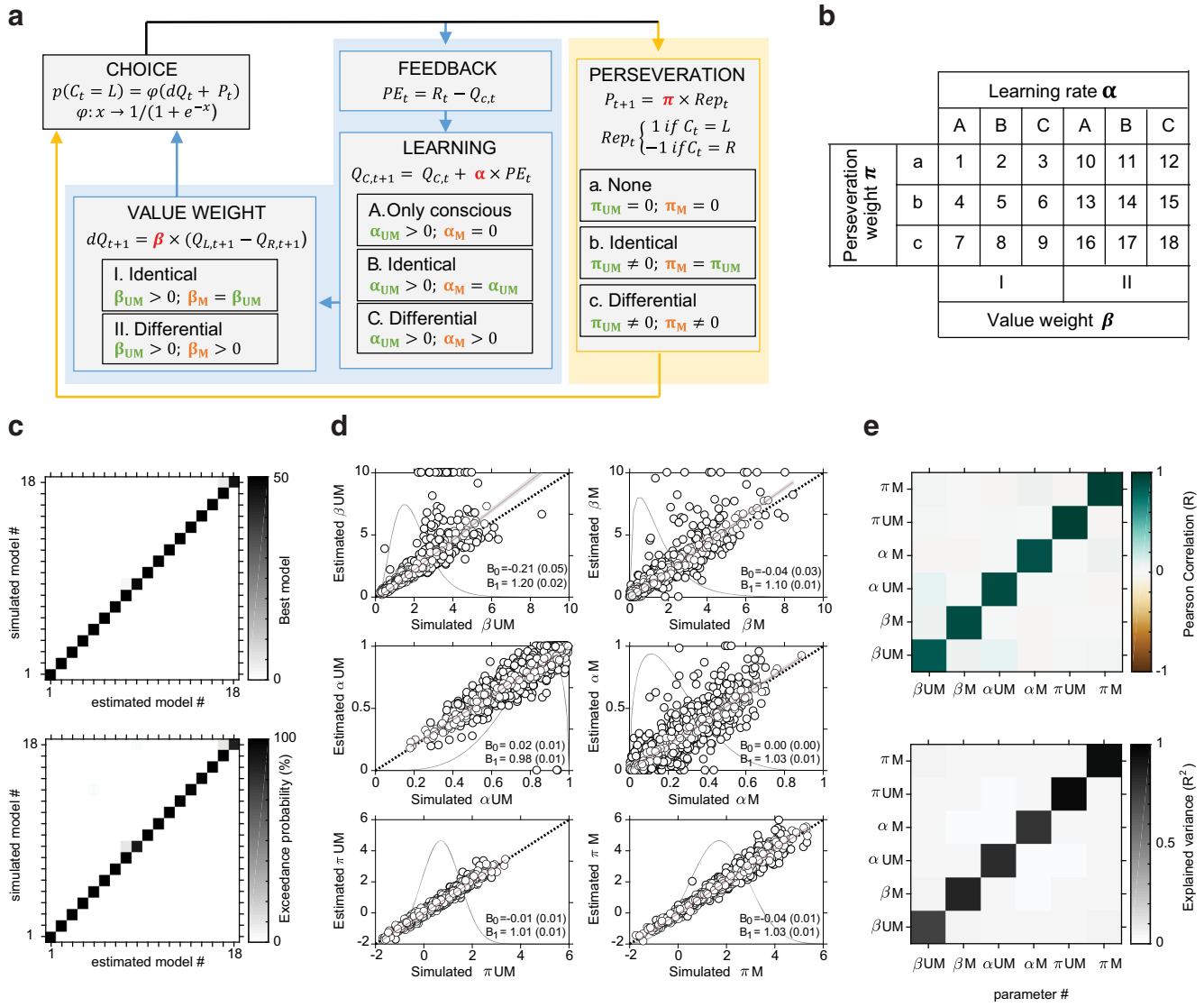

Figure 2. Modeling approach. ***a***, The computational architecture used to build the model space. ***b***, Model space. Eighteen models were built by systematically combining the different options available for the different computational modules. ***c***, Model identifiability analysis. Data from 32 synthetic participants were simulated with each of our 18 models. Bayesian model selection was used to identify the most probable model generating the data, using model exceedance probability. This procedure was repeated 50 times. Overall, all 18 models were correctly identified more than 90% of the time (>45 out of 50 simulations, see top confusion matrix), with an average exceedance probability > 90% (bottom confusion matrix). ***d***, Parameter recovery analysis - general. Overall, data from 1600 synthetic participants (50 simulations × 32 individuals) were simulated with the full model (model 18). The 6 estimated parameters per participants were then regressed against the true parameters used for simulating the data. Results show very good identifiability, with regression intercepts (β_0s) close to 0, regression slopes (β_1s) close to 1 and highly significant (all *p*-values lower than Matlab's precision –i.e. reported as = 0). Each dot represents a synthetic individual. The black dotted lines represent the identity line, the red continuous lines the best linear fits, and the shaded grey areas the 95% confidence interval around the best-linear fit. The grey densities represent the probability distributions used to sample the parameters. ***e***, Parameter recovery analysis – individual simulations. The confusion matrices represent summary statistics of the correlations between parameters, estimated over 32-subjects simulations, and averaged over the 50 simulations. Diagonal: correlations between simulated and estimated parameters. Off diagonal: cross correlation between estimated parameters. Top: Pearson correlation (R). Bottom: explained variance (R^2).

everyone could learn the probabilities, for at least 10 trials in a row they should have been able to choose the "correct side" option for >60% of the last 25 trials, otherwise additional trials could be added until this condition was completed. Self-paced rest breaks were given every 70 trials, presenting participants with the percentage of correct sides they have chosen according to the rule at stake. This break never coincided with the changing probability conditions, and participants were told about that.

In 10% of the trials, a forced choice discrimination question asked "Which coin did you just see?" while displaying a 1 cent or a 50 cent coin. This questions was asked equally often after unmasked and masked coins. Participants were instructed that the probability of the correct response being a 1 cent or 50 cent coin was 50%. It was explained to participants that they would be paid according to their performance at the end of the experiment. Finally, all participants received a bonus of €5 on top of what they had already received. Participants were instructed to choose one of the two targets on each trial, to pay attention to the reward, and to try to win as much money as possible.

Models building blocks

We designed 18 different models, all adapted from a Q-learning model. Our Q-learning included the following three basic modules: learning, choice, and perseveration (Fig. 2*a*).

Learning. The basic idea is that participants learn by trial and error to compute a value *Q* for each option (choosing the left or the right cue). At each trial *t*, after a choice is made and the outcome of the choice R_t is revealed, the *Q* value of the chosen option ($Q_{C,t+1}$) is updated by integrating a so-called prediction error, δ_t, which compares what was expected ($Q_{C,t}$) to the actual outcome, as follows:

$$\delta_t = R_t - Q_{C,t}.$$

This update is typically scaled by a learning rate α, such that:

$$Q_{C,t+1} = Q_{C,t} + \alpha \times \delta_t.$$

Choice. To account for the fact that people try to maximize their expected outcome, but can make errors or explore locally suboptimal options, the choice (C_t) is typically implemented as a softmax function, as follows:

$$P(C_t = a) = (1 + exp(\beta \times (Q_t(a) - Q_t(b))))^{-1},$$

where β is the slope of the logistic choice function—the inverse temperature parameter—which we refer to as the value weight.

Perseveration. To capture the tendency of participants to stick to their previous choices independently of the received reward, we also included a perseveration bias, π_t, in the choice function. This function becomes the following:

$$P(C_t = a) = (1 + exp(\beta \times (Q_t(a) - Q_t(b)) + \pi \times P_t))^{-1},$$

where

$$P_{t+1} = \begin{cases} 1 \ if \ C_t = a \\ -1 \ if \ C_t = b \end{cases},$$

and π governs the weight of the past choice on the present decision, referred to as the perseveration weight.

When both learning and perseveration are present, the relative importance of β and π allow the model capture participants tendency to tradeoff between sampling from learned value (β) versus simply repeating previous choices (π).

Model space

Given that our task incorporates two types of reward—masked versus unmasked—several scenarios are possible for learning and perseveration, which can be accounted for by different models. We first assumed that all models share a common basic block; that is, people learn from unmasked reward. Additionally, people can learn from masked reward, either at the same pace or at a different pace than after unmasked reward. Likewise, the value weight parameter can be identical or different after unmasked versus masked reward. As for the perseveration, it can be absent after both masked and unmasked reward: present and of identical strength, or present with different strengths. Those three learning, two choice-temperature, and three perseveration scenarios were therefore combined, generating 18 possible models in our model space (Fig. 2*a*,*b*).

Parameter optimization

We optimized the free parameters (α values, β values, and π values) of the models by minimizing the negative log likelihood (LLmax) of the participant-observed choices under the model using the fmincon function in Matlab (MathWorks), initialized at multiple starting points of the parameter space.

Model comparison

LLmax values were used to compute the Bayesian information criterion (BIC), for each model, at the individual level [BIC = 2 × (LLmax) + df × log(n_{trial})], and used it to approximate the model evidence (e = −BIC/2). Individual model evidence values were then fed to the mbb-vb-toolbox (http://mbb-team.github.io/VBA-toolbox/) to run a Bayesian model comparison (BMC; Daunizeau et al., 2014). This Bayesian procedure estimates, among other criteria, the exceedance probability (denoted XP) for each model within a set of models, given the data gathered from all participants. XP quantifies the belief that the model is more likely than all the other models of the set. An XP >95% for one model within a set is therefore typically considered as significant evidence in favor of this model being the most likely. In addition, the relative BIC (δBIC; i.e., the BIC for each model relative to best model) can be used to compare models based on the Bayes factor scale proposed by Kass and Raftery (1995).

Model identifiability and parameter recovery

We ran 50 simulations, generating choice patterns for cohorts of 32 synthetic subjects with the 18 different models in our model set. For those simulations, parameters were randomly sampled from probability distributions, which approximate the distribution of parameters estimated from fitting the complete model (i.e., model 18) to the choices of our 32 participants. As is common in the field (Daw et al., 2011; Palminteri et al., 2015), inverse temperature parameters were sampled in Gamma distributions defined by a shape (a) and a scale (b) parameter (UM: a = 4.0; b = 0.5; M: a = 1.5; b = 1.0), and learning rates were sampled in β distributions defined by two parameters, α and β (UM: $\alpha = 5.0$; $\beta = 1.5$; M: $\alpha = 1.5$; $\beta = 5.0$). Finally, perseveration parameters were sampled in normal distributions, characterized by mean (μ) and SD (σ; UM: $\mu = 0.7$; $\sigma = 0.8$; M: $\mu = 1.7$; $\sigma = 1.2$). Task properties and contingencies (e.g., block lengths) used for the simulations were rigorously identical to the 32 instances that participants faced in our experiment.

Then, we ran our BMC analysis on those 50 × 18 different simulations and checked that all models are identifiable (i.e., can be correctly estimated as the most probable model in the set of 18 models by the BMC approach when they were actually used to generate the data). This first analysis intends to verify that nothing in the design of the model set, the parameter estimation, or the model comparison approach, unduly advantages model 18 (e.g., that it is the most complex model), leading to mistakenly overestimate the probability that model 18 explains our participants' choices in lieu of other models. Next, because our models are nested, we assessed the parameter recovery in the full-model case (model 18): we computed the Pearson correlation between the parameters used to generate the data, and the parameters estimated by the maximum-likelihood fitting procedure. Additionally, we estimated the correlation between estimated parameters.

Parameters and model recovery

All 18 models are correctly identified >90% of the time, with an average XP of >90% (Fig. 2*c*). A closer look at the parameters estimated from the 1200 trials over the 50 simulations run with model 18 (the most complex model, in which all other model are nested) show that parameters are also very well recovered, with regression intercepts (β_0 values) close to 0, and regression slopes (β_1 values) close to 1 and highly significant [all p values lower than the precision Matlab are reported as equal to 0; Fig. 2*d*]. At the scale of a single simulation, the correlation between simulated and estimated parameters over 32 synthetic participants was very significant (averaged Pearson correlation = 0.92, averaged $R^2 = 0.85$; Fig. 2*e*, diagonals), while no cross-correlation was observed between parameters (all R^2 values <0.06; Fig. 2*e*, off-diagonals).

EEG measurements

EEG data were recorded and sampled at 512 Hz using a BioSemi ActiveTwo System. A total of 64 four scalp electrodes was measured, as well as 4 electrodes for horizontal and vertical eye movements (each referenced to their counterpart) and 2 reference electrodes on the ear lobes (the average was used for referencing). After acquisition, standard preprocessing steps were performed in the EEGLAB toolbox in Matlab. Data were bandpass filtered from 0.5 to 40 Hz off-line for ERP analyses. Epochs ranging from 1.8 s before to 2 s after reward presentation were extracted. Linear baseline correction was applied to these epochs using a −200 to 0 ms window. The resulting trials were visually inspected, and those containing artifacts were removed manually. Moreover, electrodes that consistently contained artifacts were interpolated. Finally, using independent component analysis, artifacts caused by blinks and other events not related to brain activity were removed from the EEG data.

ERP analyses

We focused on ERP components related to reward outcome processing with different latencies and topographical distributions. To zoom in on these specific components a central region of interest (ROI) was defined as comprising 15 midline electrodes (Fz, F1, F2, FC1, FCz, FC2, Cz, C1, C2, CPz, CP1, CP2, Pz, P1, and P2), where both the relevant components can be observed (frontocentral FRN and centroparietal P3; Cohen et al., 2007, 2011; Chase et al., 2011; Ullsperger et al., 2014). Selecting a predefined ROI limits the number of comparisons that need to be per-

formed, but we note that the results were robust and were not dependent on the specific sets of electrodes used as an ROI (see Fig. 4). We investigated the effect of reward outcome separately for masked and unmasked trials. To correct for multiple comparisons due to the number of time points tested, p values were false discovery rate (FDR) corrected at an α-level of 0.05. All statistical analyses were performed in Matlab (MathWorks). Based on this ERP analysis, three time windows of interest were selected for follow-up analyses in which we related model parameters to single-trial EEG responses.

Single-trial regression analyses

Multiple regressions of ERP amplitude on three model parameters were conducted. For each subject, each electrode, and each time point, the three parameters (PE, |PE|, switch/repeat on the next trial) were entered as predictor variables, and the ERP amplitudes as observations in the regression model. We checked that the correlations between the time series of the three predictors was low (absolute value of Pearson's R averaged over subjects, <0.2), resulting in low-multicollinearity indices [variance inflation factors (VIFs): $VIF_{PE} = 1.0596 \pm 0.0099$; $VIF_{|PE|} = 1.0524 \pm 0.0147$; $VIF_{switch/repeat} = 1.0712 \pm 0.0145$]. β-Coefficients assigned to each predictor column, which reflect the regression weights between each of the three parameters and ERP amplitude, were estimated at the individual level, separately for each electrode and time point. The significance of the predictors was assessed at the population level using random effects (t tests) on the regression coefficients averaged across the predefined time windows (100–300, 300–500, and 500–800 ms) and the predefined ROI.

Code availability

The codes used to analyze data from the current study are available from the corresponding author upon reasonable request.

Data availability

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Results

Behavior

Participants were able to perform the task well, and they accurately tracked probability reversals (mean correct response = 71.3 ± 1.51%). To assess the reward discriminability in the M and UM conditions, we computed participants' d', an unbiased measure of stimulus visibility, from the forced-choice discrimination trials that were presented throughout the task (10% of all trials, hence 120 trials in total). Although the overall discriminability was low in the masked condition, both masked and unmasked conditions exhibited above-chance accuracy in this discrimination test (UM: 96 ± 1.15% correct, $d' = 3.97 \pm 0.14$, $t_{(31)} = 28.38$, $p < 0.001$; M: 55.7 ± 1.13% correct, $d' = 0.35 \pm 0.07$, $t_{(31)} = 4.91$, $p < 0.001$). Given that chance-level performance on such a forced-choice discrimination task is a typical criterion used to show that participants are unable to perceive a stimulus consciously (Sandberg et al., 2010; Overgaard and Sandberg, 2012), this result implies that we cannot consider that the masked reward was nonconscious in all participants and for all trials.

Having established that participants performed the task correctly, we turned to a typical behavioral analysis of learning. Following previous studies (Chase et al., 2011; den Ouden et al., 2013), we computed switch rates of participants after positive and negative outcomes, in both unmasked and masked conditions. Critically, participants switched their response more often after no reward than after reward, and did so in both the unmasked and masked conditions (UM: difference 36.06 ± 0.59%, $t_{(31)} = 10.76$, $p < 0.001$; M: difference 4.90 ± 0.15%, $t_{(31)} = 5.65$, $p < 0.001$). The fact that participants tended to switch their choices significantly more after no reward (1 cent) versus reward (50 cents) is generally interpreted as evidence for learning. It would therefore be tempting to conclude that our participants significantly learned from both unmasked and masked rewards. However, this interpretation of switch patterns may not be devoid of statistical confounds, especially in designs where conditions (in this case, masked and unmasked) are intermixed. Indeed, this pattern of results could easily be produced by participants learning the value of options from unmasked rewards and deriving all choices from those values (i.e., in the total absence of learning from masked reward). This is why we turned to model-based behavioral analyses that are devoid of this statistical confound, aiming at showing that learning from masked reward outcomes is still present when these issues are taken into account.

Computational modeling

A simple δ rule was used to capture how individuals updated the value of the chosen options after receiving reward. Following classical associative learning algorithms, the extent to which previous reward is integrated in the future option value was controlled by a learning rate, α. Choices were derived from a logistic (softmax) choice function on the difference between option values. The slope of this choice function, typically referred to as choice temperature, was defined as the value weight β. Although very popular and accounting for a wide range of behavior, this learning mechanism might not account for the full choice pattern of participants in our task; indeed, within blocks, our participants might identify the best option and therefore start disregarding the feedback, putting more weights on their priors. To account for this behavior, we added a perseveration module to our computational model. Perseveration, defined as the tendency to repeat a choice regardless of the previous outcome, was integrated as an additional "bias" in the choice function, which regulated the probability of choosing the same option as that in the previous trial (Rutledge et al., 2009; Seymour et al., 2012; den Ouden et al., 2013; Voon et al., 2015). The extent to which perseveration contributed to the final choice was determined by a perseveration weight, π (Fig. 2*a*; see Materials and Methods). We then systematically explored how masked versus unmasked reward impacted those different modules, by creating sets of models allowing, or not allowing, parameters to differ between those two conditions (see Materials and Methods; Fig. 2*b*). We thereby built 18 different models, which were subsequently fit to the behavior, using a maximum likelihood procedure. A model recovery (Fig. 2*c*) and a parameter recovery (Fig. 2*d*,*e*) analysis confirmed that our modeling approach is suitable to address our questions of interests (Palminteri et al., 2017; see Materials and Methods).

Regarding our participants' data, a Bayesian model comparison approach identified model 18 as the best among our designs to explain the behavior (XP > 80%; Fig. 2*c*). The best fitting model differentiates learning rate, value weight, and perseveration weight parameters after unmasked and masked reward. Importantly, because our model space included models explicitly omitting learning from masked reward (Fig. 2*b*), this model comparison result demonstrates the existence of learning from masked reward, even when perseveration effects are taken into account.

Participant-level data reveals that the best fitting model gives a very good account of participant's learning and switch behavior (average likelihood per trial = 78.70 ± 2.11%; Fig. 3*a* for three representative participants, s10, s20, and s30). We then turned to the analysis of the best fitting model parameters (Fig. 3*b*). Learning rates appeared to be higher after unmasked than masked reward ($\alpha_{UM} = 0.67 \pm 0.03$; $\alpha_{M} = 0.19 \pm 0.02$; $t_{(31)} = 17.01$, $p <$

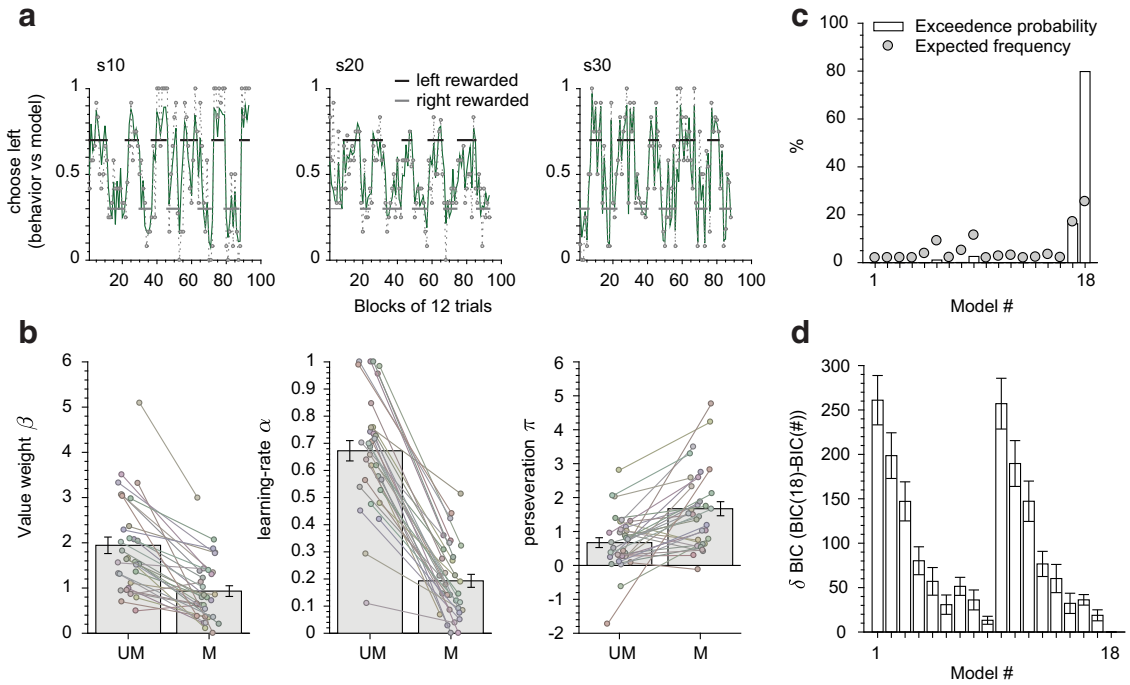

Figure 3. ***a***, Time course of the learning task by three representative participants (participant numbers 10, 20 and 30). The x-axis represents blocks of trials during the experiment and the y-axis represents the local fraction of left-hand responses selected by the participant. Thick black and gray lines represent the reward probability in the different blocks (75–125 trials). Gray-dotted lines represent the local fraction of left-hand responses. Green thick line represent the local probability of left-hand responses predicted by the computational model. Both behavioral choices and model predictions are averaged over 12 trials bins, and aligned on block transitions. ***b***, Model parameters for masked and unmasked conditions. Left: value weight. Middle: learning rate. Right: perseveration weight. M: masked reward, UM: unmasked reward. Histograms and error bars represent mean ± s.e.m. Connected dots represent individual parameters. ***c***, Model comparison. Results of a Bayesian model comparison analysis on our participants' data. White histograms indicate the exceedance probability of each model, and grey dots their expected frequencies. ***d***, Relative BIC. Bayesian Information Criterion (BIC) of each model, compared to the best fitting model BIC (model 18). BICs are computed at the individual level (random effects). Histogram and error bars represent mean ± s.e.m.

0.001), and so did value weights ($\beta_{UM} = 1.94 \pm 0.18$; $\beta_M = 0.93 \pm 0.12$; $t_{(31)} = 7.24$, $p < 0.001$). However, the opposite was found for the weight put on previous choices ($\pi_{UM} = 0.67 \pm 0.15$; $\pi_M = 1.67 \pm 0.21$; $t_{(31)} = -4.72$, $p < 0.001$; Fig. 3*b*).

These results lead to several crucial insights concerning reward learning. First, they demonstrate the existence of robust learning from masked rewards. Second, they clearly illustrate changes, due to reward visibility, in the trade-off between the tendency to base choices on the learned options' values, and the tendency to repeat previous choices regardless of previous outcome. This thus suggests that the reliance on the longer-term priors, based on the accumulation of recent choices, is increased when the outcome on the current trial is masked and therefore unreliable.

Finally, we ran independent linear regressions with each of the individual parameters from the model (six parameters in total) as independent variables and overall performance (percentage correct) as the dependent variable to explore what model parameters correlate with individual performance. Results show that inverse temperatures (β_{UM}: $\beta = 0.060$, $p < 0.001$; β_M: $\beta = 0.076$, $p < 0.001$) and perseveration parameters (π_{UM}: $\beta = 0.043$, $p = 0.016$; π_M: $\beta = 0.046$, $p < 0.001$) are positively correlated with performance, while learning rates (α_{UM}: $\beta = -0.210$, $p = 0.0016$; α_M: $\beta = -0.229$, $p = 0.036$) are negatively correlated with performance.

ERPs and model-based EEG results

Having established, thanks to the manipulation of reward visibility, a clear computational dissociation between the contributions of learning versus choice perseveration to the behavior of our participants, we next aimed at dissociating the neural signatures of those components by leveraging electrophysiological recordings. To first identify the electrophysiological time windows of interest, we performed an ERP analysis of reward-related activity, contrasting reward versus no-reward outcomes, at our central region of interest, which was based on previous studies (Cavanagh et al., 2010; Cohen et al., 2011; Ullsperger et al., 2014; see Materials and Methods).

Our analysis of event-related potentials revealed three significant events in the neural signal evoked by fully conscious (unmasked) outcomes: an early FRN at frontocentral electrodes ("early" event), which was followed by a second, more centrally distributed negative component ("middle" event), and a final parietal P3 component ("late" event; Fig. 4*a*; FDR corrected across time, $p < 0.05$). Crucially, while masked outcomes also elicited an early frontocentral FRN, neither the second negative ERP component nor the P3 component could be observed in the masked condition (FDR corrected across time, $p < 0.05$; Fig. 4*b*).

To relate the contributions of the different computational modules identified in our best fitting model (Fig. 2, model 18) to electrophysiological signatures of outcome-guided decision-making, we then turned to a model-based analysis of the EEG signal. In each participant, at each electrode and at each time point, we estimated a multiple regression with the trialwise time series of electrophysiological activity as the dependent variable, and the trialwise time series of latent variables as independent variables (see Materials and Methods). Three such independent variables, derived from our best fitting model, were included in this multiple regression: the signed prediction error; the unsigned prediction error (typically interpreted as a measure of surprise;

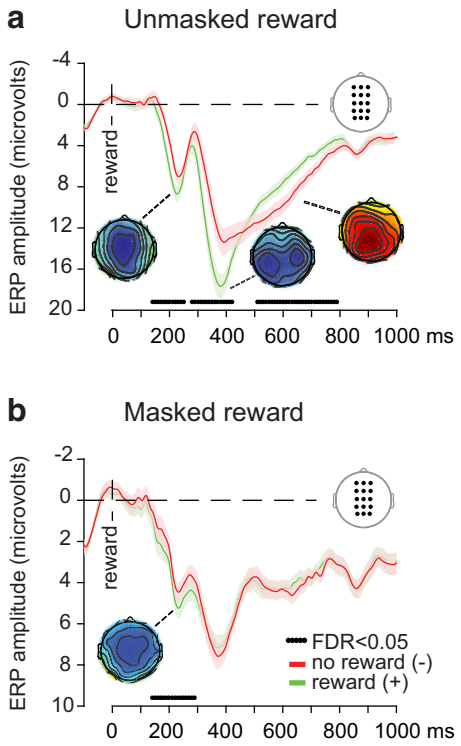

Figure 4. ERP results. ERPs for no-reward (red lines) and reward (green lines) for unmasked ***a***, and masked conditions. ***b***, Time = 0 ms is reward presentation. The lower dotted black lines indicate significant time-windows, FDR corrected across the entire ERP time-window ($p < 0.05$). Topographical distribution maps of the reward valence effect (no-reward minus reward, − vs +) were taken from the three broad time-windows (100–300 ms, 300–500 ms and 500–800 ms; scaling maps unmasked reward from left to right: [−2:2], [−5:5], [−2:2]; scaling maps masked reward: [−2:2]). Error bars represent ± s.e.m.

Pearce and Hall, 1980; Cavanagh and Frank, 2014); and a variable indexing whether participants switched or repeated their choice from the previous trial to the next trial, which is directly related to the perseveration process (switch/stay behavior). Previous research has shown the existence of temporally overlapping but spatially separate contributions of the signed prediction error, reflecting the valence of the prediction error (positive or negative) and the unsigned prediction error (the absolute degree of expectation violation also referred to as surprise) to reward learning (Fouragnan et al., 2017).

In our model-based analyses, we focus on the three contiguous time windows in which the model-free effects were most pronounced (early, 100–300 ms; middle, 300–500 ms; late, 500–800 ms). The signed PE regression results showed two clear peaks strongly overlapping in time with the early two ERP components that were revealed in the model-free ERP analysis (Fig. 5*a*). For both masking conditions, the signed prediction error was encoded in the early FRN (early time window: UM: $t_{(31)} = 6.8$, $p < 0.001$; M: $t_{(31)} = 4.2$, $p < 0.001$; difference: $t_{(31)} = 3.0$, $p = 0.005$). Similar results were obtained for the mid-latency negativity (middle time window: UM: $t_{(31)} = 11.2$, $p < 0.001$; M: $t_{(31)} = 3.0$, $p = 0.005$; difference: $t_{(31)} = 8.1$, $p < 0.001$). In contrast, the later P3 component appeared to reach significance only in the masked outcome condition, although both conditions did not differ significantly (late time window: UM: $t_{(31)} = 0.85$, $p = 0.40$; M: $t_{(31)} = 4.1$, $p < 0.001$; Fig. 5*a*).

Analyses of the unsigned prediction error signals (i.e., the level of surprise) revealed a rather different pattern of results. For both masked and unmasked reward, and in line with previous findings (Mars et al., 2008; Fischer and Ullsperger, 2013; Fouragnan et al., 2017), this variable was represented in the later P3-like component [time window 300–500 ms: UM: $t_{(31)} = 5.5$, $p < 0.001$; M: $t_{(31)} = 1.8$, $p = 0.08$; time window 500–800 ms: UM: $t_{(31)} = 8.4$, $p < 0.001$; M: $t_{(31)} = 2.2$, $p = 0.03$; Fig. 5*b* (note that headmaps are shown for the middle and late windows combined, 300–800 ms)]. In both time windows, the effects were stronger for unmasked than masked rewards (all p values <0.001). No significant effects were observed in the early time window (all p values >0.3).

Finally, we observed a strong relation between switch/stay behavior on the next trial, closely related to the perseveration parameter in the modeling approach, and a broad central positivity (Fig. 5*c*). This effect was already present from the early time window onward and was always present regardless of reward visibility [time window 100–300 ms: UM: $t_{(31)} = 2.9$, $p = 0.006$; M: $t_{(31)} = 2.9$, $p = 0.006$; difference: $t_{(31)} = -0.8$, $p = 0.4$; time window 300–500 ms: UM: $t_{(31)} = 5.1$, $p < 0.001$; M: $t_{(31)} = 5.6$, $p < 0.001$; difference: $t_{(31)} = 0.5$, $p = 0.6$; time window 500–800 ms: UM: $t_{(31)} = 7.1$, $p < 0.001$; M: $t_{(31)} = 3.8$, $p < 0.001$; difference: $t_{(31)} = 2.2$, $p = 0.034$; Fig. 5*c* (note that headmaps are shown for the middle and late windows combined, 300–800 ms)]. Interestingly, these effects were very similar for masked and unmasked rewards until ~500 ms after stimulus presentation, and significant visibility-related differences only started to emerge in the late time window. Thus, a larger parietal positive component was associated with an increased likelihood of switching the response option on the next trial. This last analysis not only replicates previous findings about the electrophysiological signature of model-free switching behavior after fully conscious reward (Chase et al., 2011; Fischer and Ullsperger, 2013), but also extends them to the case where reward visibility is very low.

Finally, we ran independent linear regressions with each of the individual EEG regressor weights shown in the bar plots of Figure 5 (PE, surprise, switching), for masked and unmasked feedback, for each of the three time windows of interest, as independent variables and overall performance (percentage correct) as the dependent variable, to explore what neural mechanisms correlate with individual performance (18 regressions in total, Bonferroni corrected). Results show that only the middle and late EEG-switching effects from the unmasked feedback (Fig. 5*c*) were positively correlated with performance (both p values <0.0005).

Discussion

We combined a reinforcement learning task, a masking procedure, computational modeling and EEG recordings to investigate the impact of reward visibility on different cognitive processes involved in probabilistic reward-guided learning. In behavioral analyses, we observed that participants switched their responses after unmasked and masked unfavorable outcomes (no-reward) more often than after favorable outcomes (reward; note that masked feedback is not considered "unconscious" here). This pattern of behavior is typically interpreted as evidence for learning. Next, we combined computational modeling with a model comparison approach. We designed a set of 18 models, built on mixtures of unmasked and masked modules, accounting for reward-based learning and choice perseveration. Reward-based learning was simply operationalized as prediction error-based learning, in line with popular model-free reinforcement learning algorithms (Sutton and Barto, 1998; Dayan and Balleine, 2002; Berridge, 2004; den Ouden et al., 2013). We then systematically compared the ability of these models to explain our participants' behavior with a rigorous Bayesian model comparison approach

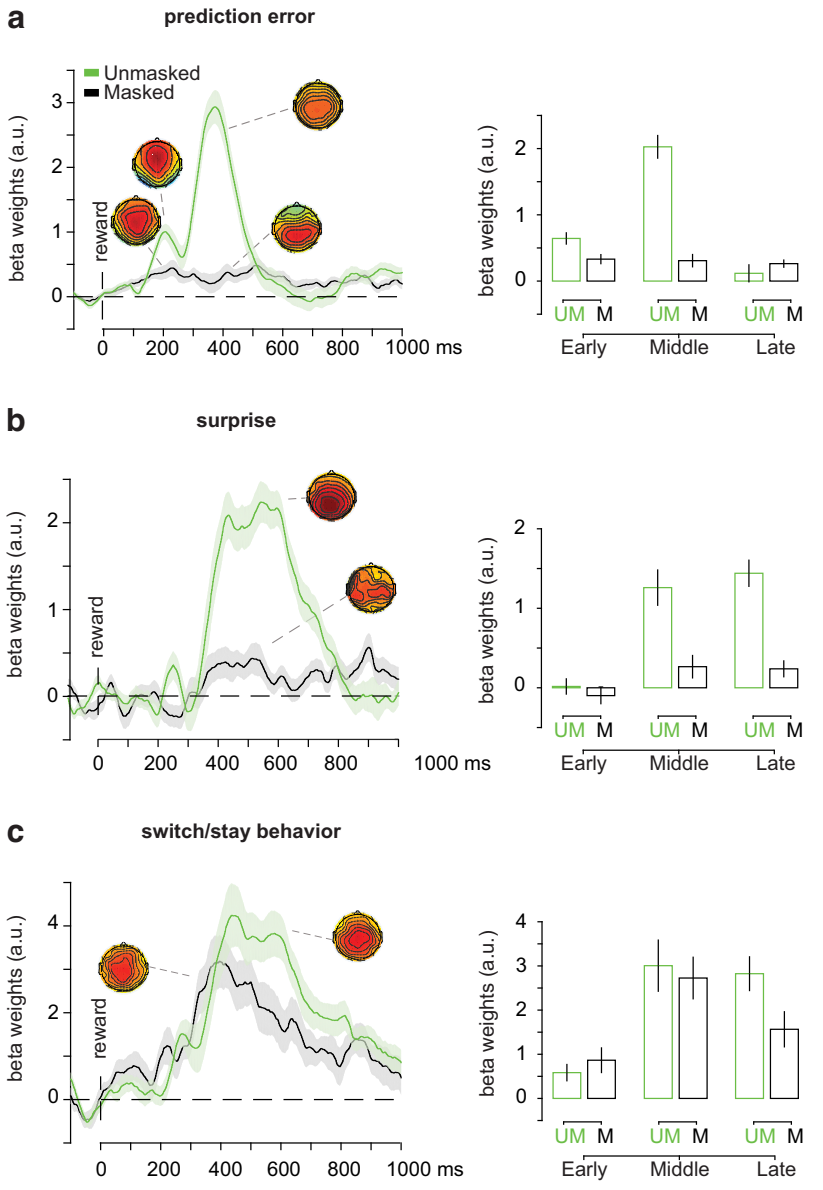

Figure 5. Model-based EEG analysis. ***a***, The time courses of regression weights of the signed PE regressed on the reward-locked EEG signal derived from a central ROI. Effects are plotted separately for unmasked (green) and masked (black) reward outcomes. Shaded areas indicate the s.e.m. Topographical maps show the regression weights during the relevant time windows. Both unmasked and masked reward showed early and mid-latency EEG-PE covariations which are shown in b. Note that the polarities of these components are reversed compared to the ERP results, which in accordance with our expectations, because these ERP modulations are all associated with negative PE values, leading to a reversal of the polarities (maps: 100–300 ms and 300–500 ms; scaling: early masked = [−0.5:0.5], mid-latency masked = [−0.5:0.5], early unmasked = [−1:1], middle unmasked = [−3:3]). Bar plots of the signed PE effect for the three time-windows of interest. ***b***, The time courses of regression weights of the unsigned PE, or the level of surprise, regressed on reward-locked EEG signal derived from a central ROI. Both unmasked and masked rewards showed late EEG-surprise covariations (maps: 300–800 ms; scaling: masked = [−0.5:0.5], unmasked = [−2:2]). Bar plots of the surprise effect. ***c***, The time courses of regression weights of switch/stay behavior regressed on the reward-locked EEG signal derived from a central ROI. Both unmasked and masked reward showed late EEG-switch/stay behavior covariations (maps: 300–800 ms; scaling: masked = [−3:3], unmasked = [−3:3]). Bar plots of the switch/stay behavior effect. Error bars represent ± s.e.m. M: masked reward, UM: unmasked reward.

(Daunizeau et al., 2014). In our model set, which comprised models with and without learning modules from masked feedback, a model including both the masked and unmasked learning modules was identified as the best model. This approach operationalized a clear testing of learning from masked outcomes and provided clear evidence toward the existence of such learning. Our best fitting model also included modules for perseveration after masked and unmasked rewards.

An analysis of the best fitting model parameters revealed that learning rates were significantly positive for both visibility modules, although smaller for the masked feedback module. This confirms that participants indeed used both unmasked and masked (although to a lesser extent) reward outcome to inform further decisions. Our results show that the perseveration parameter was also significantly positive for both the visibility modules, although perseveration was smaller for the fully conscious module. This indicates that participants were biased toward repeating previous choices, independently of the outcome of their decisions, an actually frequent observation in human and nonhuman reinforcement learning tasks (Lau and Glimcher, 2005; Schönberg et al., 2007; Rutledge et al., 2009; Seymour et al., 2012; den Ouden et al., 2013). Although often given a low-level interpretation and a connotation of suboptimality (Voon et al., 2015), perseveration can also constitute the implementation of higher-level behavior: in our task, it is likely that, within a block, participants identified the "good" option based on the integration of information over a long sequence of trials, and therefore decided to ignore irrelevant negative reward by basing their choices only on their prior. After masked reward, participants persevered more than after fully conscious reward, revealing that participants stuck to their decision strategy, based on the integration of information over a longer sequence of trials, when full conscious awareness of the outcome was (often) lacking.

Regarding electrophysiological signatures of reinforcement learning, we observed three neural events evolving over time that were modulated by unmasked outcomes (reward vs no reward): an early frontocentral FRN, a mid-latency central negativity, and a late centroparietal P3 component. Crucially, only the frontocentral FRN, which peaked ~200 ms after outcome presentation, was also modulated by masked outcomes. Many studies have reported that this signal, closely related to the response-locked error-related negativity and originating from the medial frontal cortex (MFC; Debener et al., 2005; Hauser et al., 2014), distinguishes positive from negative outcomes (Holroyd et al., 2003; Hajcak et al. 2006; Cohen et al. 2007; Cavanagh et al. 2010; Chase et al. 2011; Pfabigan et al. 2011; Fouragnan et al. 2017) in reinforcement learning tasks (Holroyd and Coles, 2002). This response may reflect a "fast alarm" signal (or alertness re-

sponse; Fouragnan et al., 2017) that indicates the value of the incoming evidence, which is then accumulated in later stages of the decision-making process (Chase et al., 2011; Ullsperger et al., 2014; Fouragnan et al., 2017), possibly reflected in the P3 ERP component (O'Connell et al., 2012). The late parietal P3 ERP component was observed only after fully conscious (unmasked) reward. This signal has been reported to predict behavioral adaptation and the associated update of new stimulus–response associations in memory (Chase et al., 2011; Ullsperger et al., 2014). The P3 has also been related to decision formation and evidence accumulation processes during perceptual decision-making (Zylberberg et al., 2011; O'Connell et al., 2012; Fischer and Ullsperger, 2013; Ullsperger et al., 2014). Further, our ERP results fit nicely with current theoretical models of conscious and unconscious processes (Lamme, 2006; van Gaal and Lamme, 2012; Dehaene et al., 2014). Within these frameworks, the FRN may reflect a fast feedforward and nonconscious high-level response, whereas the P3 may reflect more conscious and longer-lasting neural responses, potentially dependent on recurrent interactions between distant brain regions (Dehaene and Changeux, 2011).

Although those first EEG analyses outlined important dissociations between learning from reward at different levels of awareness, it is rather difficult to connect these neural signals to precise cognitive processes, using cross-trial averaging and traditional contrast-based ERP methods (Debener et al., 2005; Cohen and Cavanagh, 2011; Pernet et al., 2011; Pfabigan et al., 2011). We therefore ran additional regression analyses in combination with computational modeling to investigate whether single-trial measures of reinforcement learning were influenced by the visibility of probabilistic rewards (Cavanagh et al., 2011; Cohen and Cavanagh, 2011; Pernet et al., 2011). We focused our investigations on the EEG correlates of the following three main computational variables: the prediction error (signed PE), the level of surprise (unsigned PE), and switch/stay behavior on the next trial. This analysis revealed a striking similarity of neural PE correlates after both unmasked and masked reward outcomes, although weaker for the latter. Both the early and mid-latency negative ERP components were associated with PE computation (Fouragnan et al. 2017), whereas the parietal P3 was not. These findings support previous results showing that the FRN reflects signed PE signals (Holroyd and Coles, 2002; Overbeek et al., 2005), likely emerging from dopaminergic projections to the MFC (Schultz, 2007; Jocham et al., 2011; Park et al., 2012; Walsh and Anderson, 2012), although the early response especially has also been linked to noradrenergic and serotonergic modulations (for review, see Fouragnan et al. 2015).

Interestingly, whereas the two early neural events coded for a signed PE signal, the later P3 component was particularly modulated by the unsigned PE, reflecting the level of surprise. Although this corroborates similar results obtained with different techniques and methods (Mars et al., 2008; Fouragnan et al., 2017), we crucially show here that the level of surprise is also encoded in parietal EEG fluctuations elicited by masked reward outcomes. Finally, the EEG switch/repeat correlations that we report here are in line with those of previous studies showing that trial-by-trial switch behavior can be observed at parietal channels as a late positive P3 component (Chase et al., 2011; Fischer and Ullsperger, 2013). In a previous study (Fischer and Ullsperger, 2013) in which the authors combined computational modeling and RL, it has been shown that this neural event did not differ when participants received actual reward about their choice or merely fictive reward. Here we show that this effect likely represents decision strategies that are formed over longer timescales. Overall, these results show that several cognitive processes important for reward-based learning, namely PE computation, surprise, and switch/stay implementation, are processed in the human brain, and that these cognitive processes are temporally and spatially dissociated in time (Fouragnan et al. 2017).

Future directions, open questions, and limitations

Although several crucial question about the role of feedback awareness in reward-based learning were addressed here, several interesting questions remain unanswered. First, the current task design did not allow us to analyze what neural processes may drive "correct switching behavior" versus switching behavior in general, due to the low number of block reversals and therefore the low number of possible correct switch trials (maximum, 11 trials/subject). Future studies may address this issue by incorporating more volatile reward environments, containing more block reversals (and therefore correct switches), to address this issue (Behrens et al., 2007). Another open question relates to the isolation of the neural and cognitive processes underlying the early versus mid-latency frontal ERP negativities. Previous studies have typically observed only one frontal negativity (the FRN), instead of two (Cohen et al., 2011; for review, see Cavanagh and Frank, 2014). At present, it remains unclear why this is the case, and future work is necessary to unravel the task specifics that may drive these differences between studies. The combination of both EEG and fMRI, as performed previously (Debener et al., 2006; Hauser et al., 2014; Fouragnan et al., 2017), may contribute to this endeavor. Finally, future studies are crucial to explore what factors may drive that the model-based single-trial regressions yielded weaker (but often still significant) effects for the masked condition compared with the unmasked condition. An interesting option may be that on a subset of trials masked feedback could have been completely missed by the system, such that no prediction error could be generated (and represented in the EEG).

References

Aarts H, Custers R, Marien H (2008) Preparing and motivating behavior outside of awareness. Science 319:1639. CrossRef Medline

Behrens TE, Woolrich MW, Walton ME, Rushworth MF (2007) Learning the value of information in an uncertain world. Nat Neurosci 10:1214–1221. CrossRef Medline

Berridge KC (2004) Motivation concepts in behavioral neuroscience. Physiol Behav 81:179–209. CrossRef Medline

Berridge KC, Robinson TE (2003) Parsing reward. Trends Neurosci 26:507–513. CrossRef Medline

Bijleveld E, Custers R, Van der Stigchel S, Aarts H, Pas P, Vink M (2014) Distinct neural responses to conscious versus unconscious monetary reward cues. Hum Brain Mapp 35:5578–5586. CrossRef Medline

Bijleveld E, Custers R, Aarts H (2012) Adaptive reward pursuit: how effort requirements affect unconscious reward responses and conscious reward decisions. J Exp Psychol Gen 141:728–742. CrossRef Medline

Capa RL, Bouquet CA, Dreher JC, Dufour A (2013) Long-lasting effects of performance-contingent unconscious and conscious reward incentives during cued task-switching. Cortex 49:1943–1954. CrossRef Medline

Cavanagh JF, Frank MJ (2014) Frontal theta as a mechanism for cognitive control. Trends Cogn Sci 18:414–421. CrossRef Medline

Cavanagh JF, Frank MJ, Klein TJ, Allen JJ (2010) Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. Neuroimage 49:3198–3209. CrossRef Medline

Cavanagh JF, Wiecki TV, Cohen MX, Figueroa CM, Samanta J, Sherman SJ, Frank MJ (2011) Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. Nat Neurosci 14:1462–1467. CrossRef Medline

Chase HW, Swainson R, Durham L, Benham L, Cools R (2011) Feedback-related negativity codes prediction error but not behavioral adjustment

during probabilistic reversal learning. J Cogn Neurosci 23:936–946. CrossRef Medline

Cohen MX, Cavanagh JF (2011) Single-trial regression elucidates the role of prefrontal theta oscillations in response conflict. Front Psychol 2:30. CrossRef Medline

Cohen MX, Ranganath C (2007) Reinforcement learning signals predict future decisions. J Neurosci 27:371–378. CrossRef Medline

Cohen MX, Elger CE, Ranganath C (2007) Reward expectation modulates feedback-related negativity and EEG spectra. Neuroimage 35:968–978. CrossRef Medline

Cohen MX, Wilmes K, Vijver Iv (2011) Cortical electrophysiological network dynamics of feedback learning. Trends Cogn Sci 15:558–566. CrossRef Medline

Collins AGE, Frank MJ (2018) Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. Proc Natl Acad Sci U S A 115:2502–2507. CrossRef Medline

Daunizeau J, Adam V, Rigoux L (2014) VBA: a probabilistic treatment of nonlinear models for neurobiological and behavioural data. PLoS Comput Biol 10:e1003441. CrossRef Medline

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ (2011) Model-based influences on humans' choices and striatal prediction errors. Neuron 69:1204–1215. CrossRef Medline

Dayan P, Balleine BW (2002) Reward, motivation, and reinforcement learning. Neuron 36:285–298. CrossRef Medline

Debener S, Ullsperger M, Siegel M, Fiehler K, von Cramon DY, Engel AK (2005) Trial-by-trial coupling of concurrent electroencephalogram and functional magnetic resonance imaging identifies the dynamics of performance monitoring. J Neurosci 25:11730–11737. CrossRef Medline

Debener S, Ullsperger M, Siegel M, Engel AK (2006) Single-trial EEG-fMRI reveals the dynamics of cognitive function. Trends Cogn Sci 10:558–563. CrossRef Medline

Dehaene S, Changeux JP (2011) Experimental and theoretical approaches to conscious processing. Neuron 70:200–227. CrossRef Medline

Dehaene S, Charles L, King JR, Marti S (2014) Toward a computational theory of conscious processing. Curr Opin Neurobiol 25:76–84. CrossRef Medline

den Ouden HE, Daw ND, Fernandez G, Elshout JA, Rijpkema M, Hoogman M, Franke B, Cools R (2013) Dissociable effects of dopamine and serotonin on reversal learning. Neuron 80:1090–1100. CrossRef Medline

Evans JS (2008) Dual-processing accounts of reasoning, judgment, and social cognition. Annu Rev Psychol 59:255–278. CrossRef Medline

Evans JS, Stanovich KE (2013) Dual-process theories of higher cognition: advancing the debate. Perspect Psychol Sci 8:223–241. CrossRef Medline

Fischer AG, Ullsperger M (2013) Real and fictive outcomes are processed differently but converge on a common adaptive mechanism. Neuron 79:1243–1255. CrossRef Medline

Fouragnan E, Retzler C, Mullinger K, Philiastides MG (2015) Two spatiotemporally distinct value systems shape reward-based learning in the human brain. Nat Commun 6:8107. CrossRef Medline

Fouragnan E, Queirazza F, Retzler C, Mullinger KJ, Philiastides MG (2017) Spatiotemporal characterization of the neural correlates of outcome valence and surprise during reward learning in humans. Sci Rep 7:4762. CrossRef Medline

Hajcak G, Moser JS, Holroyd CB, Simons RF (2006) The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. Biol Psychol 71:148–154. CrossRef Medline

Hauser TU, Iannaccone R, Stämpfli P, Drechsler R, Brandeis D, Walitza S, Brem S (2014) The feedback-related negativity (FRN) revisited: new insights into the localization, meaning and network organization. Neuroimage 84:159–168. CrossRef Medline

Holroyd CB, Coles MGH (2002) The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol Rev 109:679–709. CrossRef Medline

Holroyd CB, Nieuwenhuis S, Yeung N, Cohen JD (2003) Errors in reward prediction are reflected in the event-related brain potential. Neuroreport 14:2481–2484. CrossRef Medline

Jocham G, Klein TA, Ullsperger M (2011) Dopamine-mediated reinforcement learning signals in the striatum and ventromedial prefrontal cortex underlie value-based choices. J Neurosci 31:1606–1613. CrossRef Medline

Kahneman D (2003) A perspective on judgment and choice: mapping bounded rationality. Am Psychol 58:697–720. CrossRef Medline

Kass RE, Raftery AE (1995) Bayes factor. J Am Stat Assoc 90:773–795. CrossRef

Lamme VA (2006) Towards a true neural stance on consciousness. Trends Cogn Sci 10:494–501. CrossRef Medline

Lau B, Glimcher PW (2005) Dynamic response-by-response models of matching behavior in rhesus monkeys. J Exp Anal Behav 84:555–579. CrossRef Medline

Mars RB, Debener S, Gladwin TE, Harrison LM, Haggard P, Rothwell JC, Bestmann S (2008) Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. J Neurosci 28:12539–12545. CrossRef Medline

Newell BR, Shanks DR (2014) Unconscious influences on decision making: a critical review. Behav Brain Sci 37:1–19. CrossRef Medline

O'Connell RG, Dockree PM, Kelly SP (2012) A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. Nat Neurosci 15:1729–1735. CrossRef Medline

O'Doherty JP, Hampton A, Kim H (2007) Model-based fMRI and its application to reward learning and decision making. Ann N Y Acad Sci 1104:35–53. CrossRef Medline

Overbeek TJM, Nieuwenhuis S, Ridderinkhof KR (2005) Dissociable components of error processing: on the functional significance of the pe vis-à-vis the ERN/Ne. J Psychophysiol 19:319–329. CrossRef

Overgaard M, Sandberg K (2012) Kinds of access: different methods for report reveal different kinds of metacognitive access. Philos Trans R Soc Lond B Biol Sci 367:1287–1296. CrossRef Medline

Palminteri S, Khamassi M, Joffily M, Coricelli G (2015) Contextual modulation of value signals in reward and punishment learning. Nat Commun 6:8096. CrossRef Medline

Palminteri S, Wyart V, Koechlin E (2017) The importance of falsification in computational cognitive modeling. Trends Cogn Sci 21:425–433. CrossRef Medline

Park SQ, Kahnt T, Talmi D, Rieskamp J, Dolan RJ, Heekeren HR (2012) Adaptive coding of reward prediction errors is gated by striatal coupling. Proc Natl Acad Sci U S A 109:4285–4289. CrossRef Medline

Pearce JM, Hall G (1980) A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol Rev 87:532–552. CrossRef Medline

Pernet CR, Sajda P, Rousselet GA (2011) Single-trial analyses: why bother? Front Psychol 2:322. CrossRef Medline

Pessiglione M, Schmidt L, Draganski B, Kalisch R, Lau H, Dolan RJ, Frith CD (2007) How the brain translates money into force: a neuroimaging study of subliminal motivation. Science 316:904–906. CrossRef Medline

Pfabigan DM, Alexopoulos J, Bauer H, Sailer U (2011) Manipulation of feedback expectancy and valence induces negative and positive reward prediction error signals manifest in event-related brain potentials. Psychophysiology 48:656–664. CrossRef Medline

Rangel A, Camerer C, Montague PR (2008) A framework for studying the neurobiology of value-based decision making. Nat Rev Neurosci 9:545–556. CrossRef Medline

Rutledge RB, Lazzaro SC, Lau B, Myers CE, Gluck MA, Glimcher PW (2009) Dopaminergic drugs modulate learning rates and perseveration in Parkinson's patients in a dynamic foraging task. J Neurosci 29:15104–15114. CrossRef Medline

Sandberg K, Timmermans B, Overgaard M, Cleeremans A (2010) Measuring consciousness: is one measure better than the other? Conscious Cogn 19:1069–1078. CrossRef Medline

Schmidt L, Lebreton M, Cléry-Melin ML, Daunizeau J, Pessiglione M (2012) Neural mechanisms underlying motivation of mental versus physical effort. PLoS Biol 10:e1001266. CrossRef Medline

Schönberg T, Daw ND, Joel D, O'Doherty JP (2007) Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. J Neurosci 27:12860–12867. CrossRef Medline

Schultz W (2007) Multiple dopamine functions at different time courses. Annu Rev Neurosci 30:259–288. CrossRef Medline

Seymour B, Daw ND, Roiser JP, Dayan P, Dolan R (2012) Serotonin selectively modulates reward value in human decision-making. J Neurosci 32:5833–5842. CrossRef Medline

Sutton RS, Barto AG (1998) Introduction to reinforcement learning. Cambridge, MA: MIT.

Ullsperger M, Fischer AG, Nigbur R, Endrass T (2014) Neural mechanisms and temporal dynamics of performance monitoring. Trends Cogn Sci 18:259–267. CrossRef Medline

van Gaal S, Lamme VA (2012) Unconscious high-level information processing: implication for neurobiological theories of consciousness. Neuroscientist 18:287–301. CrossRef Medline

Voon V, Derbyshire K, Rück C, Irvine MA, Worbe Y, Enander J, Schreiber LR, Gillan C, Fineberg NA, Sahakian BJ, Robbins TW, Harrison NA, Wood J, Daw ND, Dayan P, Grant JE, Bullmore ET (2015) Disorders of compulsivity: a common bias towards learning habits. Molecular Psychiatry 20: 345–352. CrossRef Medline

Walsh MM, Anderson JR (2012) Learning from experience: event-related potential correlates of reward processing, neural adaptation, and behavioral choice. Neurosci Biobehav Rev 36:1870–1884. CrossRef Medline

Weber EU, Johnson EJ (2009) Mindful judgment and decision making. Annu Rev Psychol 60:53–85. CrossRef Medline

Zedelius CM, Veling H, Aarts H (2012) When unconscious rewards boost cognitive task performance inefficiently: the role of consciousness in integrating value and attainability information. Front Hum Neurosci 6:219. CrossRef Medline

Zylberberg A, Dehaene S, Roelfsema PR, Sigman M (2011) The human turing machine: a neural framework for mental programs. Trends Cogn Sci 15:293–300. CrossRef Medline