Title: Biased credit assignment in motivational learning biases arises through prefrontal influences on striatal learning

Short title: Motivational biases in fronto-striatal circuits

Authors

Johannes Algermissen[1]*, Jennifer C. Swart[1], René Scheeringa[1,2], Roshan Cools[1,3], Hanneke E.M. den Ouden[1]*

Affiliations

[1] Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands
[2] Erwin L. Hahn Institute for Magnetic Resonance Imaging, University of Duisburg-Essen, Essen, Germany
[3] Department of Psychiatry, Radboud University Medical Centre, Nijmegen, The Netherlands
* j.algermissen@donders.ru.nl; h.denouden@donders.ru.nl

Abstract

Actions are biased by the outcomes they can produce: Humans are more likely to show action under reward prospect, but hold back under punishment prospect. Such motivational biases derive not only from biased response selection, but also from biased learning: humans tend to attribute rewards to their own actions, but are reluctant to attribute punishments to having held back. The neural origin of these biases is unclear; in particular, it remains open whether motivational biases arise solely from an evolutionarily old, subcortical architecture or also due to younger, cortical influences. Simultaneous EEG-fMRI allowed us to track which regions encoded biased prediction errors in which order. Biased prediction errors occurred in cortical regions (ACC, vmPFC, PCC) before subcortical regions (striatum). These results highlight that biased learning is not a mere feature of the basal ganglia, but arises through prefrontal cortical contributions, revealing motivational biases to be a potentially flexible, sophisticated mechanism.

Teaser

Cortical influences on subcortical learning explain why we attribute rewards to actions, but not punishments to inactions.

Introduction

Human action selection is biased by potential action outcomes: reward prospect drives us to invigorate action, while threat of punishment holds us back (*1*–*3*). These motivational biases have been evoked to explain why humans are tempted by reward-related cues signaling the chance to gain food, drugs, or money, as they elicit automatic approach behavior. Conversely, punishment-related cues suppress action and lead to paralysis, which may even lie at the core of mental health problems such as phobias and mood disorders (*4*, *5*). While such examples highlight the potential maladaptiveness of biases in some situations, they confer benefits in other situations: Biases could provide sensible "default" actions before context-specific knowledge is acquired (*1*, *6*). They may also provide ready-made alternatives to more demanding action selection mechanisms, especially when speed has to be prioritized (*7*).

Previous research has assumed that motivational biases arise because the valence of prospective outcomes influences action selection (*8*). However, we have recently shown that not only action selection, but also the updating of action values based on obtained outcomes is subject to valence-dependent biases (*3*, *9*, *10*): humans are more inclined to ascribe rewards to active responses, but have problems with attributing punishments to having held back. One the one hand, such biased learning might be adaptive in combining the flexibility of instrumental learning with somewhat rigid "priors" about typical action-outcome relationships. Exploiting lifetime (or evolutionary) experience might lead to learning that is faster and more robust to environmental "noise". On the other hand, biases might be responsible for phenomena of "animal superstition" like auto-shaping or negative maintenance, where rats and pigeons repeat behavioral patterns that co-occurred with the attainment of (factually random) rewards and keep showing such behavior even if it delays or decreases rewards (*1*, *11*, *12*). While reward attainment can lead to an illusory sense of control over outcomes, control is underestimated under threat of punishment: Humans find it hard to comprehend how inactions can cause negative outcomes, which makes them more lenient in judging harms caused by others' inactions (*13*, *14*). Taken together, also credit assignment is subject to motivational biases, with enhanced credit for rewards given to actions, but diminished credit for punishments given to inactions.

While evident in behavior, the neural mechanisms subserving such biased credit assignment are unclear. One strong candidate region is the striatum, part of the evolutionarily old basal ganglia system. Influential computational models of basal ganglia function (*15*, *16*) (henceforth called "asymmetric pathways model") predict such motivational learning biases: Positive prediction errors, elicited by rewards, lead to long-term potentiation in the striatal direct "Go" pathway (and long term depression in the indirect pathway), allowing for a particularly effective acquisition of Go responses after rewards. Conversely, negative prediction errors, elicited by punishments, lead to long term potentiation in the "NoGo" pathway, impairing the unlearning of NoGo responses after punishments. This account suggests that motivational biases arise within the same pathways involved in standard reinforcement learning (RL). An alternative candidate model is that biases arise through the modulation of these evolutionarily old RL systems by external, evolutionarily younger areas that also track past actions, putatively the prefrontal cortex (PFC). Past research has suggested that standard RL can be biased by information stored in PFC, such as explicit instructions (*17*, *18*) or cognitive-map like models of the environment (*19*–*21*). Most notably, the anterior cingulate cortex (ACC) has been found to reflect the impact of explicit instructions (*18*) and of environmental changes on prediction errors (*22*, *23*).

Both candidate models predict that BOLD signal in striatum should be better described by biased compared with "standard" prediction errors. In addition, the model proposing a prefrontal influence on striatal processing makes a notable prediction about the timing of signals: information about the selected action and the obtained outcome should be present first in prefrontal circuits to then later affect processes in the striatum. While fMRI BOLD recordings allow for unequivocal access to striatal activity, the sluggish nature of the BOLD signal prevents clear inferences about temporal precedence of signals from different regions. We thus combined BOLD with simultaneous EEG recordings which allowed us to precisely characterize learning signals in both space and time.

The key question is whether biased credit assignment arises directly from biased RL through the asymmetric pathways in the striatum, or whether striatal RL mechanisms are biased by external prefrontal sources, with the ACC as likely candidate. To this end, participants performed a motivational Go/ NoGo learning task that is well-established to evoke motivational biases of action (*3*, *9*, *24*). We expected to observe biased PEs in striatum and frontal cortical areas. By simultaneously recording fMRI and EEG and correlating trial-by-trial BOLD signal with EEG time-frequency power, we were able to time-lock the peaks of EEG-BOLD correlations for regions reflecting biased PEs and infer their relative temporal precedence. We focused on two well-established electrophysiological signatures of RL, namely theta and delta power (*25*–*30*) as well as beta power (*25*, *31*) over midfrontal electrodes.

Results

Thirty-six participants performed a motivational Go/ NoGo learning task (*3*, *9*) in which required action (Go/ NoGo) and potential outcome (reward/ punishment) were orthogonalized (Fig. 1A-D). They learned by trial-and-error for each of eight cues whether to perform a left button press (Go_{LEFT}), right button press (Go_{RIGHT}), or no button press (NoGo), and whether a correct action increased the chance to win a reward (Win cues) or to avoid a punishment (Avoid cues). Correct actions lead to 80% favorable outcomes (reward, no punishment), with only 20% favorable outcomes for incorrect actions. Participants performed two sessions of 320 trials, with separate cue sets, which were counterbalanced across participants.

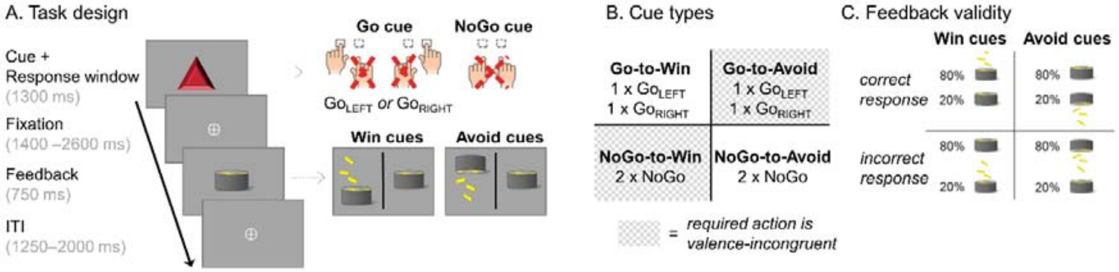

Figure 1. Motivational Go/ NoGo learning task design. (A) On each trial, a Win or Avoid cue appears; valence of the cue is not signaled but should be learned. Cue offset is also the response deadline. Response-dependent feedback follows after a jittered interval. Each cue has only one correct action (Go_{LEFT}, Go_{Right}, or NoGo), which is followed by the favorable outcome 80% of the time. For Win cues, actions can lead to rewards or neutral outcomes; for Avoid cues, actions can lead to neutral outcomes or punishment. Rewards and punishments are depicted by money falling into/ out of a can. (B) There are eight different cues, orthogonalizing cue valence (Win versus Avoid) and required action (Go versus NoGo). The motivationally incongruent cues, for which the motivational action tendencies are incongruent with the instrumental requirements, are highlighted in gray. (C) Feedback is probabilistic: Correct actions to Win cues lead to rewards in 80% of cases, but neutral outcomes in 20% of cases. For Avoid cues, correct actions lead to neutral outcomes in 80% of cases, but punishments in

20% of cases. For incorrect actions, these probabilities are reversed.

Regression analyses of behavior

We performed regression analyze to test whether a) responses were biased by the valence of prospective outcomes (Win/ Avoid), reflecting biased responding and/ or learning, and b) whether response repetition after favorable vs. non-favorable outcomes was biased by whether a Go vs. NoGo response was performed, selectively reflecting biased learning.

For the first purpose, we analyzed choice data (Go/ NoGo) using mixed-effects logistic regression that included factors required action (Go/ NoGo; note that this approach collapses across Go_{LEFT} and Go_{RIGHT} responses), cue valence (Win/ Avoid), and their interaction (also reported in)(*32*). Participants learned the task, i.e., they performed more Go responses towards Go than NoGo cues (main effect of required action: $b = 0.815$, $SE = 0.113$, $\chi^2(1) = 32.008$, $p < .001$). In contrast to previous studies (*3*, *9*), learning did not asymptote (Fig. 2A), which provided greater dynamic range for the biased learning effects to surface. Furthermore, participants showed a motivational bias, i.e., they performed more Go responses to Win than Avoid cues (main effect of cue valence, $b = 0.423$, $SE = 0.073$, $\chi^2(1) = 23.695$, $p < .001$). Replicating other studies with this task, there was no significant interaction between required action and cue valence ($b = 0.030$, $SE = 0.068$, $\chi^2(1) = 0.196$, $p = .658$, Fig. 2A-B), i.e., there was no evidence for the effect of cue valence (motivational bias) differing in size between Go or NoGo cues.

Secondly, as a proxy of (biased) learning, we analyzed cue-based response repetition (probability of repeating a response on the next encounter of the same cue) as a function of outcome valence (favorable vs non-favorable outcome), performed action (Go vs. NoGo), and outcome salience (salient: reward or punishment vs. neutral: no reward or no punishment). As expected, people were more likely to repeat the same response following a favorable outcome (main effect of outcome valence: $b = 0.504$, $SE = 0.053$, $\chi^2(1) = 45.595$, $p < .001$). Most importantly, after salient outcomes, participants adjusted their responses to a larger degree following Go responses than NoGo responses, revealing the presence of a learning bias (Fig. 2C; interaction of valence x action x salience: $b = 0.248$, $SE = 0.048$, $\chi^2(1) = 19.732$, $p < .001$). When selectively analyzing trials with salient outcomes only, rewards (compared to punishments) led to a higher proportion of choice repetitions following Go relative to NoGo responses (valence x response: $b = 0.308$, $SE = 0.064$, $\chi^2(1) = 17.798$, $p < .001$; valence effect for Go only: $b = 1.276$, $SE = 0.115$, $\chi^2(1) = 53.932$, $p < .001$; valence effect for NoGo only: $b = 0.637$, $SE = 0.127$, $\chi^2(1) = 18.228$, $p < .001$; see full results in S02).

Taken together, these results suggest that behavioral adaptation following rewards and punishments is biased by the type of action that led to this outcome (Go or NoGo). However, these analyses only consider behavioral adaptation on the next trial, and cannot pinpoint the precise algorithmic nature of this learning bias. More importantly, it does not provide trial-by-trial estimates of action values as required for model-based fMRI and EEG analyses to test for regions or time points that reflect biased learning. We thus analyzed the impact of past outcomes on participants' choices using computational RL models.

Computational modeling of behavior

In line with previous work (*3*, *9*), we fitted a series of increasingly complex RL models. We started with a simple Rescorla Wagner model featuring learning rate and feedback sensitivity parameters (M1). We next added a Go bias, capturing participants' overall propensity to make Go responses (M2), and a Pavlovian response bias (M3), reflecting participants' propensity to adjust their likelihood of emitting a Go response in response to Win vs. Avoid cues (*3*). Alternatively, we added an instrumental learning bias (M4), amplifying the learning rate after rewarded Go responses and dampening it after punished NoGo responses (*3*), in line with the asymmetric pathways model. In the final model (M5), we added both a response bias and a learning bias. For the full model space (M1-M5) and model definitions, see the Methods section. For a comparison with an alternative learning bias specification based on the idea that active responses enhance credit assignment (*33*), see S04.

Model comparison showed clear evidence in favor of the full asymmetric pathways model featuring both response and learning biases (M5; model frequency: 86.43%, protected exceedance probability: 100%, see Fig. 2D, H; for model parameters and fit indices, see S03). Posterior predictive checks involving one-step-ahead predictions and model simulations showed that this model captured key behavioral features (Fig. 2E, F), including motivational biases and a greater behavioral adaptation after Go responses followed by salient outcomes than after NoGo responses followed by salient outcomes (Fig. 2 G). This pattern could not be captured by the alternative learning bias model (S04).

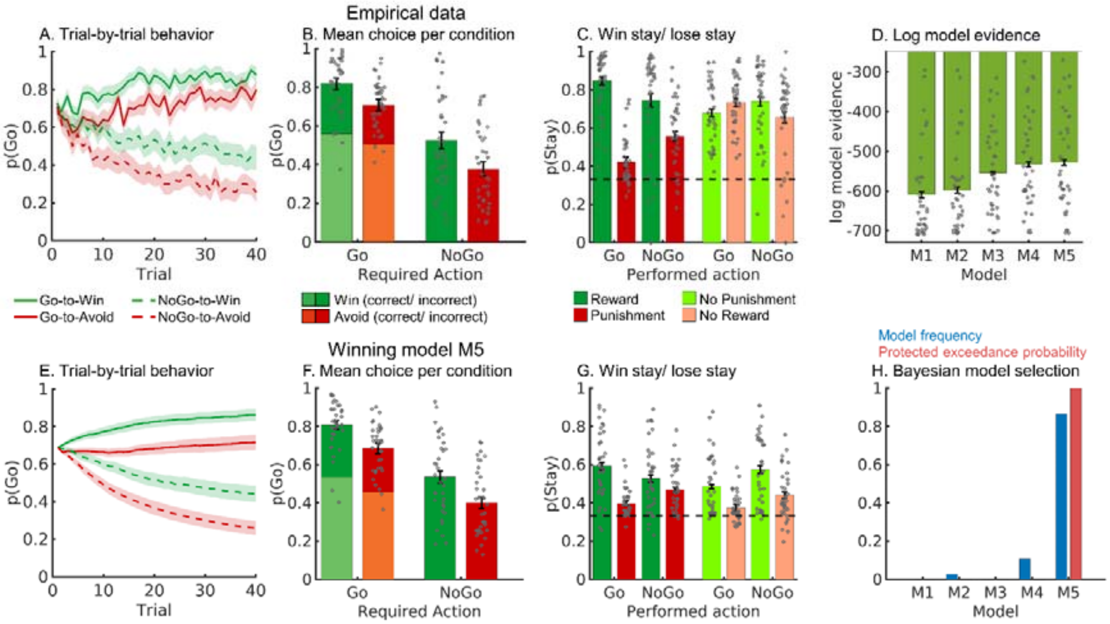

Figure 2. Behavioral performance. (A) Trial-by-trial proportion of Go responses (±SEM across participants) for Go cues (solid lines) and NoGo cues (dashed lines). The motivational bias is already present from very early trials onwards, as participants made more Go responses to Win than Avoid cues (i.e., green lines are above red lines). Additionally, participants clearly learn whether to make a Go response or not (proportion of Go responses increases for Go cues and decreases for NoGo cues). (B) Mean (±SEM across participants) proportion Go responses per cue condition (points are individual participants' means). (C) Probability to repeat a response ("stay") on the next encounter of the same cue as a function of action and outcome. Learning is reflected in higher probability of staying after positive outcomes than after negative outcomes (main effect of outcome valence). Biased learning is evident in learning from salient outcomes, where this valence effect was stronger after Go responses than NoGo responses. Dashed line indicates chance level choice (p_{Stay} = 0.33). (D) Log-model evidence favors the asymmetric pathways model (M5) over simpler models (M1-M4). (E-G) Trial-by-trial proportion of Go responses, mean proportion Go responses, and probability of staying based on one-step-ahead predictions using parameters (hierarchical Bayesian inference) of the winning model (asymmetric pathways model, M5). (H) Model frequency and protected exceedance probability indicate best fit for model M5 (asymmetric pathways model), in line with log model evidence.

fMRI: Basic quality control analyses

First, we performed a GLM as a quality-check to test which regions encoded favorable (rewards, no punishments) vs. unfavorable (no reward/ punishment) outcomes in a “model-free” way, independent of any model-based measure derived from a RL model (for full description of the GLM regressors and contrasts, see S06). Favorable outcomes elicited a higher BOLD response in regions including ventromedial PFC (vmPFC), ventral striatum, and right hippocampus, while unfavorable outcomes elicited higher BOLD in bilateral dorsolateral PFC (dlPFC), left ventrolateral PFC, and precuneous (Fig. 3A, see full report of significant clusters in S07).

We also assessed which regions encoded Go vs. NoGo as well Go_{LEFT} vs. Go_{RIGHT} responses. There was higher BOLD for Go than NoGo responses at the time of response in PFC, ACC, striatum, thalamus, motor cortices, and cerebellum, while BOLD was higher for NoGo than Go responses in right IFG (Fig. 6C left panel; see S04)(*32*). For lateralized Go responses, there was higher BOLD signal in contralateral motor cortex and operculum as well as ipsilateral cerebellum when contrasting hand responses against each other (Fig. 6C, right panel). These results are in line with previous results on outcome processing and response selection and thus assure the general data quality.

fMRI: Biased learning in prefrontal cortex and striatum

To test which brain regions were involved in biased learning, we performed a model-based GLM featuring the trial-by-trial PE update as a parametric regressor (see GLM notation in S06). We used the group-level parameters of the best fitting computational model (M5) to compute trial-by-trial belief updates (i.e., prediction error * learning rate) for every participant. In assessing neural signatures of biased learning, we faced the complication that standard (Rescorla-Wagner learning in M1) and biased PEs (winning model M5) are highly correlated. A mean correlation of 0.92 across participants (range 0.88–0.95) made it difficult to neurally distinguish biased from standard learning. To circumvent this collinearity problem, we decomposed the biased PE (computed using model M5) into the standard PE (computed using model M1) plus a difference term (*19*, *34*):

$$PE_{BIAS} = PE_{STD} + PE_{DIF}$$

A neural signature of biased learning should, significantly and with the same sign, encode both components of this biased PE term. Standard PEs and difference term were uncorrelated (mean correlation of -0.02 across participants; range -0.33–0.24). We tested for biased PEs PE_{BIAS} by computing which regions significantly encoded the conjunction of both its components, i.e., standard prediction errors PE_{STD} and the difference to biased PEs PE_{DIF}. While PE_{STD} was encoded in a range of cortical and subcortical regions (Fig. 3B, S07) previously reported in the literature (*35*), significant encoding of both PE_{STD} and PE_{DIF} (conjunction) occurred in striatum (caudate, nucleus accumbens), vmPFC/ perigenual ACC (area 32d), ventral ACC (area 23/24), posterior cingulate cortex (PCC), left motor cortex, left inferior temporal gyrus, and early visual regions (Fig. 3C; see full report of significant clusters in S07). Thus, BOLD signal in these regions was better described (i.e., more variance explained) by biased learning than by standard prediction error learning.

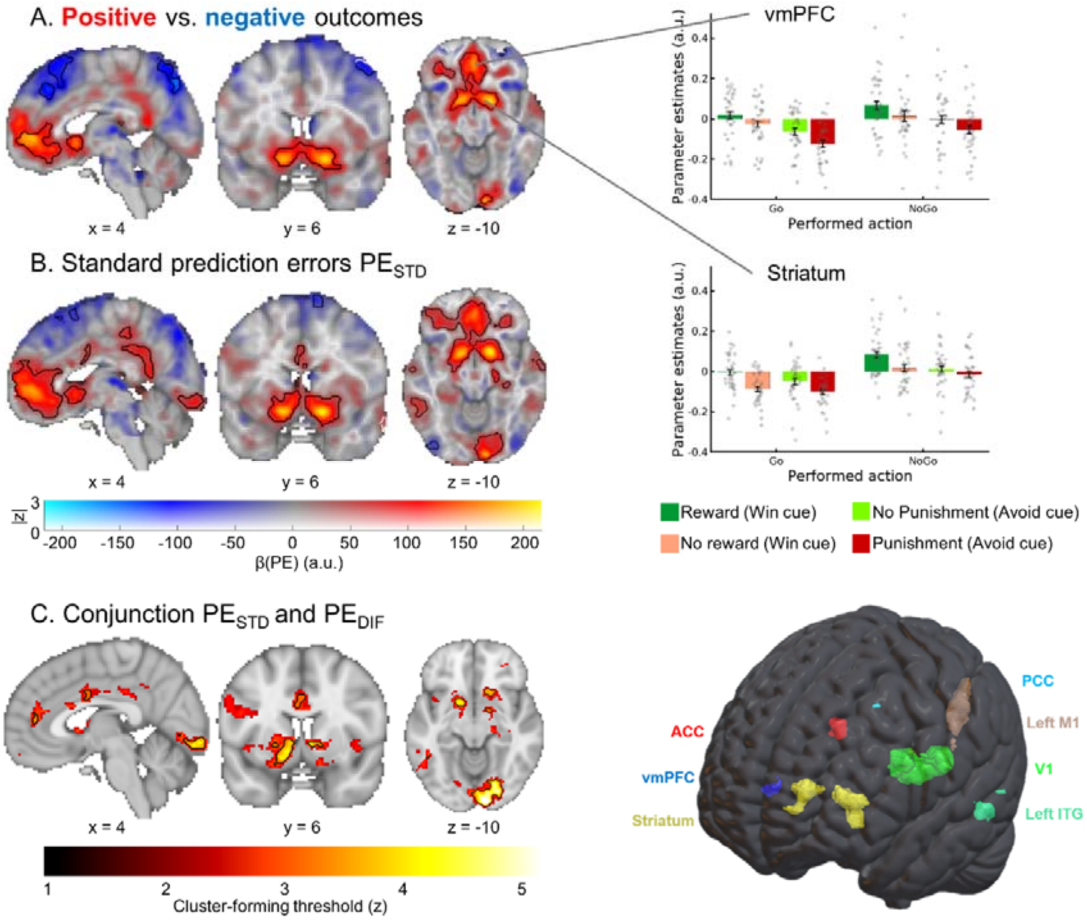

Figure 3. BOLD signal reflecting outcome processing. BOLD effects displayed using a dual-coding visualization: color indicates the parameter estimates and opacity the associated z-statistics. Significant clusters are surrounded by black edges. (A) significantly higher BOLD signal for favorable outcomes (rewards, no punishments) compared with unfavorable outcomes (no rewards, punishments) was present in a range of regions including bilateral ventral striatum and vmPFC. Bar plots show mean parameter estimates per condition (±SEM across participants; dots indicating individual participants) (B) BOLD signals correlated positively to "standard" RL prediction errors in several regions, including the ventral striatum, vmPFC, PCC and ACC. (C) Left panel: Regions encoding both the standard PE term and the difference term to biased PEs (conjunction) at different cluster-forming thresholds ($1 < z < 5$, color coding; opacity constant). Clusters significant at a threshold of $z > 3.1$ are surrounded by black edges. In bilateral striatum, ACC, vmPFC, PCC, left motor cortex, left inferior temporal gyrus, and primary visual cortex, BOLD is significantly better explained by biased learning than by standard learning. Right panel: 3D representation with all seven regions encoding biased learning (used in fMRI-informed EEG analyses).

EEG: Biased learning in midfrontal delta, theta, and beta power

Similar to the fMRI analyses, we next tested whether midfrontal power encoded biased PEs rather than standard PEs. While fMRI provides spatial specificity of where PEs are encoded, EEG power provides temporal specificity of when signals encoding prediction errors occur (*26*, *31*). In line with our fMRI analysis, we used the standard PE term PE_{STD} and the difference to the biased PE term PE_{DIF} as trial-by-trial regressors for EEG power at each channel-time-frequency bin for each participant and then performed cluster-based permutation tests across the *b*-maps of all participants. Note that differently from BOLD signal, EEG signatures of learning typically do not encode the full prediction error. Instead, PE sign (favorable vs. unfavorable outcomes) and PE magnitude (saliency, surprise) have been found encoded separately in the theta and delta band, respectively (*28–30*). We thus added PE sign as an additional regressor to test for separate correlates of PE sign and PE magnitude. Note that PE sign is identical for standard and biased PEs; only PE magnitude distinguishes both learning models.

Both midfrontal theta and beta power reflected PE sign: Theta power was higher for unfavorable than favorable outcomes (225–475 ms, *p* = .006; Fig. 4A-B), while beta power was higher for favorable than unfavorable outcomes (300–1,250 ms, *p* = .002; Fig. 4A, C). Differences in theta power were clearly strongest over frontal channels, while the effect in the beta range was more diffuse, spreading over frontal and parietal channels (Fig. 4B-C). All results held when the condition-wise ERP was removed from the data (see S08), suggesting that differences between conditions were due to induced (rather than evoked) activity (for results in the time domain, see S09).

Delta power was indeed positively, though not significantly correlated with both PE_{STD} (*p* = 0.074, Fig. 4E) and PE_{DIF} (*p* = 0.185; Fig. 4F). Only the sum of both terms, i.e., the PE_{BIAS} term, was significantly encoded by delta power (225–475 ms; *p* = .017; Fig. 4D). For a similar observation in the time-domain EEG signal, see S10. Beyond delta power, beta power correlated positively, though not significantly with PE_{STD} (*p* = 0.110, Fig. 4E) and significantly negatively with PE_{DIF} (*p* = .001, 425 – 850 ms). Encoding of PE_{BIAS} was not significant either (*p* = 0.550, Fig 4D).

In sum, both midfrontal theta power (negatively) and beta power (positively) encoded PE sign. In addition, delta power encoded PE magnitude (positively). This encoding was only significant for biased PEs, but not standard PEs. Taken together, as was the case for BOLD signal, midfrontal EEG power also reflected biased learning. As a next step, we tested whether the identified EEG phenomena were correlated with trial-by-trial BOLD signal in identified regions. Crucially, this allowed us to test whether EEG correlates of cortical learning precede EEG correlates of subcortical learning.

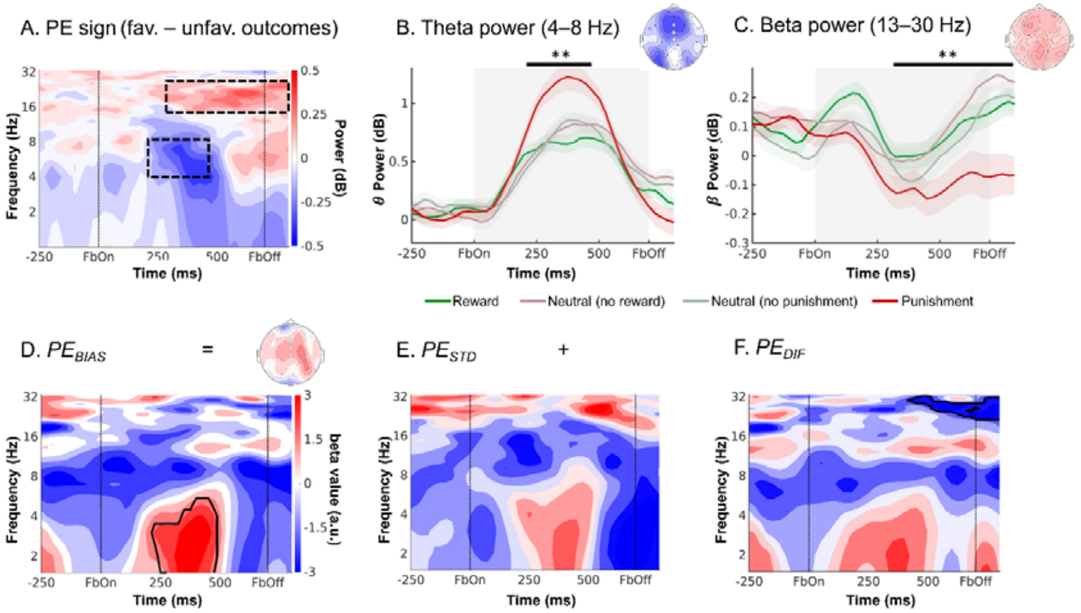

Figure 4. EEG time-frequency power over midfrontal electrodes (Fz/ FCz/ Cz). reflecting outcome processing. (A) Time-frequency plot (logarithmic y-axis) displaying higher theta (4–8 Hz) power for unfavorable outcomes and higher beta power (16–32 Hz) for favorable outcomes. Black square dot boxes indicate clusters above threshold that drive significance in a-priori defined frequency ranges. (B). Theta power transiently increases for any outcome, but more so for unfavorable outcomes (especially punishments) around 225–475 ms after feedback onset. Black horizontal lines

indicate the time range for which the cluster driving significance is above threshold. (C) Beta power is higher for favorable than unfavorable outcomes over a long time period around 300–1,250 ms after feedback onset. (D-F). Correlations between midfrontal EEG power and trial-by-trial PEs controlling for PE sign. Solid black lines indicate clusters above threshold. Biased PEs were significantly positively correlated with midfrontal delta power (D). The correlations of delta with the standard PEs (E) and the difference term to biased PEs (F) were positive as well, though not significant. Beta power only encoded the difference term to biased PEs (F). ** $p < 0.01$.

Combined EEG-fMRI: Prefrontal cortex signals precede striatum during biased outcome processing

The observation that also cortical areas (vmPFC, ACC, PCC) show biased PEs is consistent with the "external model" of cortical signals biasing learning processes in the striatum. However, this model makes the crucial prediction that these bias signals should be present first in cortical areas and only later in the striatum. Next, we used trial-by-trial BOLD signal from those regions encoding biased PE to predict midfrontal EEG power. By determining the time points at which different regions correlated with EEG power, we were able to infer the relative order of biased PE processing across cortical and subcortical regions, revealing whether cortical processing preceded striatal processing. We used trial-by-trial BOLD signal from the seven regions encoding biased PEs, i.e., striatum, ACC, left motor cortex, vmPFC, PCC, left ITG, and primary visual cortex (see masks in S05) as regressors on average EEG power over midfrontal electrodes (Fz/ FCz/ Cz). We controlled for biased PEs themselves to capture additional variance in EEG explained by BOLD signal beyond the task regressors. As the timeseries of all seven regions were included in one single regression, their regression weights reflect each region's unique contribution, controlling for any shared variance. In line with the "external model", BOLD signal from prefrontal cortical regions correlated with midfrontal EEG power earlier after outcome onset than did striatal BOLD signal:

First, ACC BOLD was significantly negatively correlated with alpha/ theta power early after outcome onset (100–575 ms, 2 – 17 Hz, $p = .016$; Fig. 5A). This cluster started in the alpha/ theta range and then spread into the theta/delta range (henceforth called "lower alpha band power"). It was not observed in the EEG-only analyses reported above.

Second, while vmPFC/ perigenual ACC BOLD did not correlate significantly with midfrontal EEG power (p = .184), BOLD in PCC was negatively correlated with theta/ delta power (Fig. 5B; 175–500 ms, 1–6 Hz, $p = .014$). This finding bears resemblance in terms of time-frequency space to the cluster of (negative) PE sign encoding in the theta band and (positive) PE magnitude encoding in the delta band identified in the EEG-only analyses (Fig. 4A). As a reverse check of this link, we added the trial-by-trial power in the EEG-only theta/delta band cluster as a regressor to the fMRI GLM featuring prediction errors, which yielded significant clusters of negative EEG-BOLD correlation in vmPFC and PCC (Fig. 5F; S13). We thus discuss vmPFC and PCC together in the following.

Third, there was a significant positive correlation between striatal BOLD and midfrontal beta/ alpha power (driven by a cluster at 100–800 ms, 7–23 Hz, $p = .010$; Fig. 5C). This finding bears resemblance in time-frequency space to the cluster of positive PE sign encoding in beta power identified in the EEG-only analyses (Fig. 4A). Again, to substantiate this link, we performed the reverse approach of using trial-by-trial power in the EEG-only beta band cluster as a regressor added to the fMRI GLM. Clusters of positive EEG-BOLD correlations in right dorsal caudate (and left parahippocampal gyrus) as well as clusters of negative correlations in bilateral dorsolateral PFC (dlPFC) and supramarginal gyrus (SMG; Fig. 5G; see S13) confirmed the positive striatal BOLD-beta power association. Given that the striatum is unlikely to be the source of midfrontal beta power over the scalp, this analysis suggests dlPFC and SMG as likely candidate sources.

Finally, regarding the other three regions that showed a significant BOLD signature of biased PEs: BOLD in left motor cortex was significantly negatively correlated with early midfrontal beta power (*p* = .002; around 0 – 625 ms; see S11). There were no significant correlations between midfrontal EEG power and left inferior temporal gyrus or primary visual cortex BOLD (see S11). All results were robust to different analysis approaches including shorter trial windows, different GLM specifications, inclusion of task-condition and fMRI motion realignment regressors, and individual modelling of each region, and were not reducible to phenomena in the time domain (see S12).

In sum, there were negative correlations between ACC BOLD and midfrontal lower alpha band power early after outcome onset, negative correlations between PCC BOLD and midfrontal theta/ delta power at intermediate time points, and positive correlations between striatal BOLD and midfrontal beta power at late time points (Fig. 5D, H). These results are consistent with an "external model" of motivational biases arising from early cortical processes biasing later learning processes in the striatum.

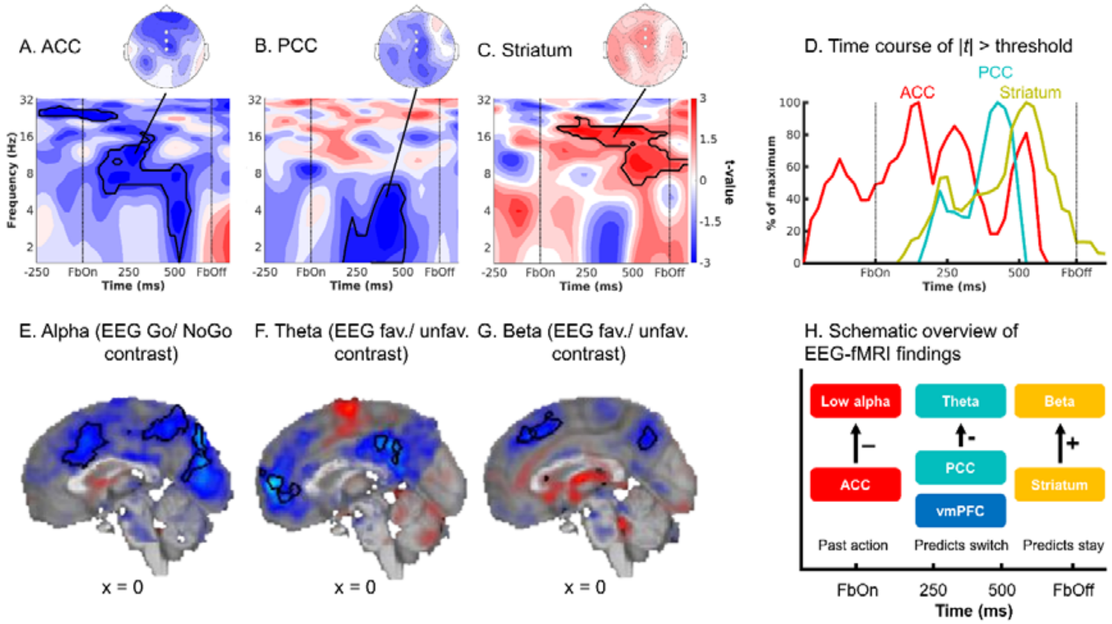

Figure 5. fMRI-informed EEG analyses. Unique temporal contributions of BOLD signal in (A) ACC, (B) PCC, and (C) striatum to average EEG power over midfrontal electrodes (Fz/ FCz/ Cz). Group-level *t*-maps display the modulation of the EEG power by trial-by-trial BOLD signal in the selected ROIs. ACC BOLD correlates negatively with early alpha/ theta power, PCC BOLD negatively with theta/ delta power, striatal BOLD positively with beta/ alpha power. Areas surrounded by a black edge indicate clusters of |*t*| > 2 with *p* < .05 (cluster-corrected). Topoplots indicate the topography of the respective cluster. (D) Time course of ACC, PCC, and striatal BOLD correlations, normalized to the peak of the time course of each region. ACC-lower alpha band correlations emerge first, followed by (negative) PCC-theta correlations and finally positive striatum-beta correlations. Reverse approach using lower alpha (E), theta (F) and beta (G) power as trial-by-trial regressors in fMRI GLMs. These EEG-informed fMRI analyses corroborate the fMRI-informed EEG analyses: Lower alpha band power correlated negatively with the ACC BOLD, theta power negatively with vmPFC and PCC BOLD, and beta power positively with striatal BOLD. (H) Schematic overview of the main EEG-fMRI results: ACC encodes the previously performed response and correlates with early midfrontal lower alpha band power. vmPFC/ PCC (correlated with theta power) and striatum (correlated with beta power) both encode outcome valence, but have opposite effects on subsequent behavior. Note that activity in these regions temporally overlaps; boxes are ordered in temporal precedence of peak activity.

ACC BOLD and midfrontal lower alpha band power encode the previously performed action during outcome presentation

While the clusters of EEG-fMRI correlation in the theta/ delta and beta range matched the clusters identified in EEG-only analyses, the cluster of negative correlations between ACC BOLD and early midfrontal lower alpha band power was novel and did not match our expectations. Given that these correlations arose very soon after outcome onset, we hypothesized that ACC BOLD and midfrontal lower alpha band power might reflect a process occurring even before outcome onset, such as the maintenance (“eligibility trace”) of the previously performed response to which credit may later be assigned. We therefore assessed whether information of the previous response was present in ACC BOLD and in the lower alpha band around the time of outcome onset.

First, we tested for BOLD correlates of the previous response at the time of *outcomes* (eight outcome-locked regressors for every Go/ NoGo x reward/ no reward/ no punishment/ punishment combination) while controlling for motor-related signals at the time of the *response* (response-locked regressors for left-hand and right-hand button presses). At the time of outcomes, there was higher BOLD signal for NoGo than Go responses across several cortical and subcortical regions, peaking in both the ACC and striatum (Fig. 6E). This inversion of effects—higher BOLD for Go than NoGo responses at the time of response (see quality checks), but the reverse at the time of outcome—was also observed in the upsampled raw BOLD and was independent of the response of the next trial (S14). In sum, large parts of cortex, including the ACC, indeed encoded the previously performed response at the moment outcomes were presented, in line with the idea that the ACC maintains an “eligibility trace” of the previously performed response.

Second, we tested for differences between Go and NoGo responses at the time of outcomes in midfrontal broadband EEG power. Power was significantly higher on trials with Go than on trials with NoGo responses, driven by clusters in the lower alpha band (spreading into the theta band; around 0.000–0.425 sec., 1–11 Hz, p = .012) and in the beta band (around 0.200–0.450 sec., 18–27 Hz, p = .022; Fig. 6A, B). The first cluster matched the time-frequency pattern of ACC BOLD-alpha power correlations (Fig. 5A).

If this activity cluster contained a signature of the previously performed response, it might have been present throughout the delay between cue offset and outcome onset. When repeating the above permutation test including the last second before outcome onset, there were significant differences again, driven by a sustained cluster in the beta band (-1–0 sec., 13–33 Hz, p = .002) and two clusters in the alpha/ theta band (Cluster 1: -1.000– -0.275 sec., 1–10 Hz, p = 0.014; Cluster 2: -0.225–0.425 sec., 1–11 Hz, p = .022; Fig. 6B). These findings suggest that lower alpha band power might reflect a sustained memory of the previously performed response. Supplemental analyses (S14) yielded that this Go-NoGo trace during outcome processing did not change over the time course of the experiment, suggesting that it did not reflect typical fatigue/ time-on task effects often observed in the alpha band.

Again, we performed the reverse EEG-fMRI analysis using trial-by-trial power in the identified lower alpha band cluster (Fig. 6B) as an additional regressor in the quality-check fMRI GLM. Clusters of negative EEG-BOLD occurred correlation in a range of cortical regions, including ACC and precuneous (Fig. 5E; see S13). In sum, both ACC BOLD signal and midfrontal lower alpha band power contained information about the previously performed response, consistent with the idea that both signals reflect an “eligibility trace” of the response to which credit is assigned once an outcome is obtained.

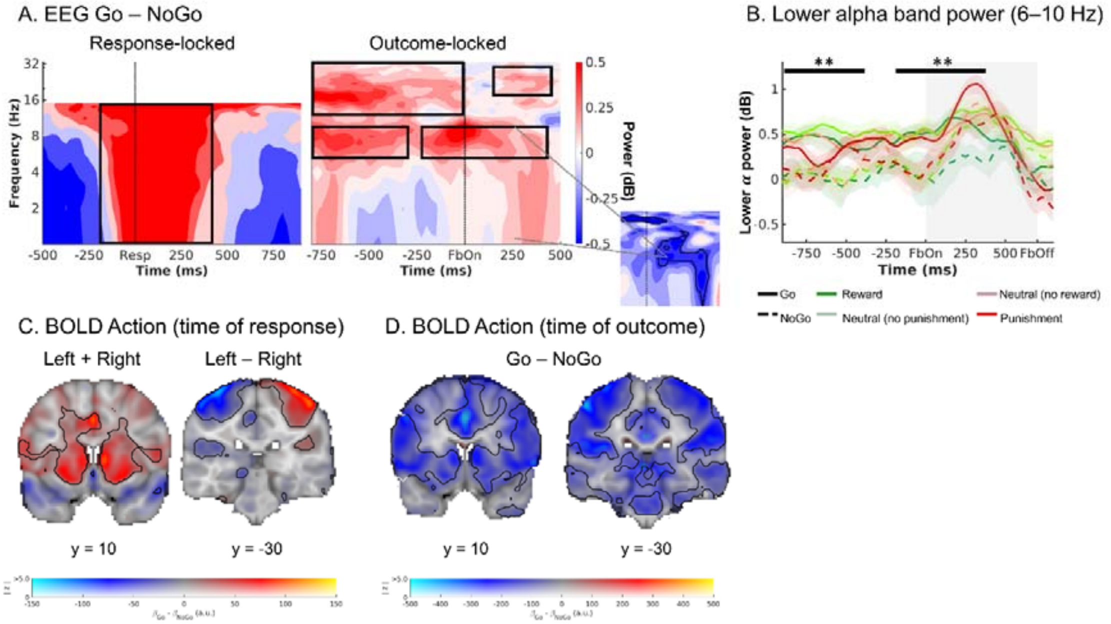

Figure 6. Exploratory follow-up analyses on ACC BOLD signal and midfrontal lower alpha band power. (A) Midfrontal time-frequency response-locked (left panel) and outcome-locked (right panel). Before and shortly after outcome onset, power in the lower alpha band is higher on trials with Go actions than on trials with NoGo actions. The shape of this difference resembles the shape of ACC BOLD-EEG TF correlations (small plot; note that this plot depicts BOLD-EEG correlations, which are negative). Note that differences between Go and NoGo trials occurred already before outcome onset in the alpha and beta range, reminiscent of delay activity, but were not fully sustained throughout the delay between response and outcome. (B) Midfrontal power in the lower alpha band per action x outcome condition. Lower alpha band power is consistently higher on trials with Go actions than on trials with NoGo actions, starting already before outcome onset. (C) BOLD signal differences between Go and NoGo actions (left panel) and left vs. right hand responses (right panel) at the time or responses. Response-locked ACC BOLD is significantly higher for Go than NoGo actions. (D) BOLD signal differences between Go and NoGo actions at the time of outcomes. Outcome-locked ACC BOLD (and BOLD in other parts of cortex) is significantly lower on trials with Go than on trials with NoGo actions.

Striatal and vmPFC/ PCC BOLD differentially relate to action policy updating

EEG correlates of PCC BOLD and striatal BOLD occurred later than for the ACC BOLD, and overlapped with classical feedback-related midfrontal theta and beta power responses. We hypothesized that those neural signals might be more closely related to updating of action policies (i.e., which action to perform for each cue) and might thus predict the next response to the same cue (*27*, *36*). We thus used the trial-by-trial BOLD responses in ACC, PCC, vmPFC and striatum to predict whether participants would repeat the same response on the next trial with the same cue ("stay") or switch to another response ("shift"). Mixed-effects logistic regression yielded that ACC BOLD did not significantly predict response repetition (b = -0.019, SE = 0.016, $\chi^2(1)$ = 1.294, p = .255). In contrast, BOLD in PCC/ vmPFC and striatum did predict response repetition, though in opposite directions: Participants were significantly more likely to repeat the same response when striatal BOLD was high (b = 0.067, SE = 0.024, $\chi^2(1)$ = 9.051, p = .003), but more likely to switch to another response when vmPFC BOLD (b = -0.076, SE = 0.017, $\chi^2(1)$ = 15.559, p < .001) or PCC BOLD (b = -0.036, SE = 0.016, $\chi^2(1)$ = 3.691, p = .030; Fig. 5H) was high (for plots, see S15). We also inspected the raw upsampled HRF shapes per region per condition, confirming that differential relationships were not driven by differences in HRF shapes across regions.

We also tested whether trial-by-trial midfrontal lower alpha band, theta, or beta power (within the clusters identified in the EEG-only analyses) predicted action policy updating. Participants were significantly more likely to repeat the same response when beta power was high (b = 0.145, SE = 0.041, $\chi^2(1)$ = 11.886, p < .001), but more likely to switch when theta power was high (b = -0.099, SE = 0.047, $\chi^2(1)$ = 4.179, p = .041). Notably, unlike its BOLD correlate in ACC, lower alpha band power did predict response repetition, with more repetition when alpha power was high (b = .0.179, SE = 0.052, $\chi^2(1)$ = 10.711, p = .001; for plots, see S15).

In sum, high striatal BOLD and midfrontal beta power predicted that the same response would be repeated on the next encounter of a cue, while high vmPFC and PCC BOLD and high theta power predicted that participants would switch to another response. Thus, although both striatal and vmPFC/PCC BOLD positively encoded biased prediction errors, these two sets of regions had opposite roles in learning: while the striatum reinforces previous responses, vmPFC/PCC trigger the shift to another response strategy (Fig. 5H).

Discussion

We investigated neural correlates of biased learning for Go and NoGo responses. In line with previous research (*3*, *9*), participants' behavior was best described by a computational model featuring faster learning from rewarded Go responses and slower learning from punished NoGo responses. Neural correlates of biased PEs were present in BOLD signals in several regions, including ACC, PCC, vmPFC, and striatum. These regions exhibited distinct midfrontal EEG power correlates. Most importantly, correlates of prefrontal cortical BOLD preceded correlates of striatal BOLD: Trial-by-trial ACC BOLD correlated with lower alpha band power immediately after outcome onset, followed by PCC (and vmPFC) BOLD correlated with theta power, and finally striatal BOLD correlated with beta power. These results are in line with a model of PFC biasing striatal outcome processing, giving rise to motivational learning biases in behavior.

Biased learning in PFC precedes the striatum

The dominant idea about the origin of motivational biases has been that these biases are an emergent feature of the asymmetric direct/ indirect pathway architecture in the basal ganglia (*2*, *16*). We find evidence that these biases are present first in prefrontal cortical areas, notably ACC and vmPFC. This argues against biases purely being a "fixed" leftover of evolutionary ancient, subcortical circuits. Rather, motivational learning biases might be an instance of sophisticated, even "model-based" learning processes in the striatum instructed by the prefrontal cortex (*37*, *38*). An influence of PFC on striatal RL has prominently been observed in the case of model-based vs. model-free learning (*20*, *21*) and has been stipulated as a mechanism of how instructions can impact RL learning (*17*, *18*). Although there are reports of striatal processes preceding prefrontal processes within learning tasks (*39*, *40*), the opposite pattern of PFC preceding striatum has been observed as well (*41*) and a causal impact of PFC on striatal learning is well established (*42*, *43*).

The particular subregion of PFC showing the earliest EEG correlates was the ACC. This observation is in line with an earlier EEG-fMRI study reporting ACC to be part of an early valuation system preceding a later system comprising vmPFC and striatum (*44*). The ACC has been suggested to encode models of agents' environment (*45*, *46*) that are relevant for interpreting outcomes. ACC BOLD has been found to scale with the size of PEs (*22*, *23*), indexing how much should be learned from new outcomes. We hypothesize that, at the moment of outcome, ACC maintains an "eligibility trace" of the previously performed response (*47*), which might modulate the processing of outcomes as soon as they become available (*48*, *49*). Notably, ACC exhibited stronger BOLD signal for Go than NoGo responses at the time of participants' response, but this pattern reversed at the time of outcomes. This reversal rules out the possibility that response-locked BOLD signal simply spilled over into the time of outcomes. Future research will be necessary to corroborate such a motor "eligibility trace" in ACC.

In sum, the ACC might be in a designated position to inform subsequent outcome processing in downstream regions by modulating the learning rate as a function of previously performed response and the obtained outcome. Rather than striatal circuits being sufficient for the emergence of motivational biases, the more "flexible" PFC seems to play role in instructing downstream striatal learning processes.

Striatum and midfrontal beta power signal maintenance of action policies

Striatal, vmPFC and PCC BOLD encoded biased PEs. In line with previous research, striatal BOLD positively linked to midfrontal beta power (*50*, *51*), which positively encoded PE sign (*25*, *31*, *52*). PCC and vmPFC BOLD negatively linked to midfrontal theta/ delta power (*32*, *53*, *54*), which encoded PE sign negatively, but PE magnitude positively. Notably, theta/ delta power correlates of vmPFC/ PCC BOLD preceded beta power correlates of striatal BOLD in time, which aligns with previous findings of motivational response biases being first visible in the vmPFC BOLD before they impact striatal action selection (*32*).

Positive encoding of prediction errors in striatal BOLD signal is a well-established phenomenon (*35*, *55*). Striatal BOLD was better described by biased PEs than by standard PEs, corroborating the presence of motivational learning biases also in striatal learning processes. Notably, EEG correlates of striatal BOLD peaked rather late, suggesting that these processes are informed by early sources in PFC which are connected to the striatum via recurrent feedback loops (*15*, *56*). Positive prediction errors increase the value of a performed action and thus strengthen action policies. Hence, it is not surprising that high striatal BOLD signal and midfrontal beta power predicted action repetition (*57*, *58*).

vmPFC and midfrontal theta/ delta power signal updating of action policies

In contrast to striatal learning signals, the PCC and vmPFC BOLD as well as midfrontal theta and delta power signals were more complicated: Theta encoded PE sign, delta encoded PE magnitude. Both correlates showed opposite polarities. This observation is in line with previous literature suggesting that midfrontal theta and delta power (resp. the feedback-related negativity and reward positivity components in the time domain EEG signal) might reflect the "saliency" or "surprise" aspect of PEs (*28*, *29*, *59*). Surprises have the potential to disrupt an ongoing action policy (*60*) and motivate a shift to another policy, which might explain why these signals predicted switching to another response (*61*, *62*). Notably, this EEG surprise signal was only significantly correlated with the biased (but not the standard) PE term, corroborating that the surprise attributed to outcomes depends on previously performed response, reflecting motivational learning biases. In sum, both vmPFC and striatum encode biased PEs, though with different consequences for future action policies.

Limitations

Taken together, distinct brain regions processed outcomes in a biased fashion at distinct time points with distinct EEG power correlates. Simultaneous EEG-fMRI recordings allowed us to infer when those regions reached their peak activity (*63*). However, the correlational nature of BOLD-EEG links precludes strong statements about these regions actually generating the respective power phenomena. Alternatively, activity in those regions might merely modulate the amplitude of time-frequency responses originating from other sources. Furthermore, while the observed associations align with previous literature (*32*, *50*, *51*, *53*, *54*), the considerable distance of the striatum to the scalp raises the question whether scalp EEG could in principle reflect striatal activity, at all (*64*, *65*). Intracranial recordings have observed beta oscillations during outcome processing in the striatum before (*58*, *66*, *67*). Also, our analysis controlled for BOLD signal in motor cortex, an alternative candidate source for beta power, suggesting that late midfrontal beta power does not merely reflect motor cortex beta. Even if the striatum is not the generator of the beta oscillations over the scalp, their true (cortical) generator might be tightly coupled to the striatum and thus act as a "transmitter" of striatal beta oscillations. In fact, the analyses using trial-by-trial beta power to predict BOLD yielded significant clusters in dlPFC and SMG, two candidate regions for such a "transmitter".

Finally, the correlational nature of the study prevents strong statements over any causal interactions between the observed regions. We assume here that a region showing an earlier midfrontal EEG correlate influences other regions showing later midfrontal EEG correlates, and such an influence is plausible given findings of feedback loops between prefrontal regions and the striatum (*56*). Future studies targeting those regions via selective causal manipulations will be necessary to test for the causal role of PFC in informing striatal learning.

The role of motivational biases in credit assignment and learning

In conclusion, biased learning—increased credit assignment to rewarded action, decreased credit assignment to punished inaction—was visible both in behavior and in BOLD signal in a range of regions. EEG correlates of prefrontal cortical regions, notably ACC and vmPFC, *preceded* correlates of the striatum, consistent with a model of the PFC biasing RL in the striatum. The ACC appeared to hold a "motor eligibility trace" of the past response, biasing early outcome processing. Subsequently, biased learning was also present in vmPFC/ PCC and striatum, with opposite roles in adjusting vs. maintaining action policies. These results refine previous views on the neural origin of these learning biases, which might not purely be "naïve" remnants of evolutionary ancient, "primitive" parts of the brain, but rather incorporate sophisticated, even "model-based" processes relying on frontal inputs. The PFC is typically believed to facilitate goal-directed over instinctive processes. Hence, PFC involvement into biased learning suggests that these biases are not necessarily agents' inescapable "fate", but rather likely act as global "priors" that facilitate learning of more local relationships. They allow for combining "the best of both worlds"—long-term experience with consequences of actions and inactions together with flexible learning from rewards and punishments.

Materials and methods

Participants

Thirty-six participants (M_{age} = 23.6, SD_{age} = 3.4, range 19–32; 25 women; all right-handed; all normal or corrected-to-normal vision) took part in a single 3-h data collection session, for which they received €30 flat fee plus a performance-dependent bonus (range €0–5, M_{bonus} = €1.28, SD_{bonus} = 1.54). The study was approved by the local ethics committee (CMO2014/288; Commissie Mensengeboden Onderzoek Arnhem-Nijmegen) and all participants provided written informed consent. Exclusion criteria comprised claustrophobia, allergy to gels used for EEG electrode application, hearing aids, impaired vision, colorblindness, history of neurological or psychiatric diseases (including heavy concussions and brain surgery), epilepsy and metal parts in the body, or heart problems. Sample size was based on previous EEG studies with a comparable paradigm (*9*, *68*).

Behavioral and modeling results include all 36 participants. The following participants were excluded from analyses of neural data: For two participants, fMRI functional-to-standard image registration failed; hence, all fMRI-only results are based on 34 participants (M_{age} = 23.47, 25 women). Four participants exhibited excessive residual noise in their EEG data (> 33% rejected trials) and were thus excluded from all EEG analyses; hence, all EEG-only analyses are based on 32 participants (M_{age} = 23.09, 23 women). For combined EEG-fMRI analyses, we excluded the above-mentioned six participants plus one more participant whose regression weights for every regressor were about ten times larger than for other participants, leaving 29 participants (M_{age} = 23.00, 22 women). Exclusions were in line with a previous analysis of this data set (*32*). fMRI- and EEG-only results held when analyzing only those 29 participants (see S01).

Task

Participants performed a motivational Go/ NoGo learning task (*3*, *9*) administered via MATLAB R2014b (MathWorks, Natick, MA, United States) and Psychtoolbox-3.0.13. On each trial, participants saw a gem-shaped cue for 1300 ms, which signaled whether they could potentially win a reward (Win cues) or avoid a punishment (Avoid cues), and whether they had to perform a Go (Go cue) or NoGo response (NoGo cue). They could press a left (Go_{LEFT}), right (Go_{RIGHT}), or no (NoGo) button while the cue was presented. Only one response option was correct per cue. Participants had to learn both cue valence and required action from trial-and-error. After a variable inter-stimulus-interval of 1,400–1,600 ms, the outcome was presented for 750 ms. Potential outcomes were a reward (symbolized by coins falling into a can) or neutral outcome (can without money) for Win cues, and a neutral outcome or punishment (symbolized by money falling out of a can) for Avoid cues. Feedback validity was 80%, i.e., correct responses were followed by favorable outcomes (rewards/ no punishments) on only 80% of trials, while incorrect responses were still followed by favorable outcomes on 20% of trials. Trials ended with a jittered inter-trial interval of 1250–2000 ms, yielding total trial lengths of 4700–6650 ms.

Participants gave left and right Go responses via two button boxes positioned lateral to their body. Each box featured four buttons, but only one button per box was required in this task. When participants accidentally pressed a non-instructed button, they received the message “Please press one of the correct keys” instead of an outcome. In the analyses, these responses were recoded into the instructed button on the respective button box. In the fMRI GLMs, such trials were modeled with a separate regressor.

Before the task, participants were instructed that each cue could be followed by either reward or punishment, each cue had one optimal response, feedback was probabilistic, and that the rewards and punishments were converted into a monetary bonus upon completion of the study. They performed an elaborate practice session in which they got familiarized first with each condition separately (using practice stimuli) and finally practiced all conditions together. They then performed 640 trials of the main task, separated into two sessions of 320 trials with separate cue sets. Introducing a new set of cues allowed us to prevent ceiling effects in performance and investigate continuous learning throughout the task. Each session featured eight cues that were presented 40 times. After every 100–110 trials (~ 6 min.), participants could take a self-paced break. The assignment of the gems to cue conditions was counterbalanced across participants, and trial order was pseudo-random (preventing that the same cue occurred on more than two consecutive trials).

Behavior analyses

We used mixed-effects logistic regression (as implemented in the R package *lme4*) to analyze behavioral responses (Go vs. NoGo) as a function of required action (Go/ NoGo), cue valence (Win/ Avoid), and their interaction. We included a random intercept and all possible random slopes and correlations per participant to achieve a maximal random-effects structure (*69*). Sum-to-zero coding was employed for the factors. Type 3 *p*-values were based on likelihood ratio tests (implemented in the R package *afex*). We used a significance criterion of α = .05 for all the analyses.

Furthermore, we used mixed-effects logistic regression to analyze "stay behavior", i.e., whether participants repeated an action on the next encounter of the same cue, as a function of outcome valence (positive: reward or no punishment/ negative: no reward or punishment), outcome salience (salient: reward or punishment/ neutral: no reward or no punishment), and performed action (Go/ NoGo). We again included all possible random intercepts, slopes, and interactions.

Computational modeling

We fit a series of increasingly complex RL models to participants' choices to decide between different algorithmic explanations for the emergence of motivational biases in behavior. We employed the same set of nested models as in previous studies using this task (*3*, *9*). For tests of alternative biases specifications, see S03.

Model space

To determine whether a Pavlovian response bias, an instrumental learning bias, or both biases jointly predicted behavior best, we fitted a series of increasing complex computational models. In each trial (t), choice probabilities for all three response options (a) given the displayed cue (s) were computed from their action weights (modified Q-values) using a softmax function:

$$p(a_t|s_t) = \frac{\exp\,(w(a_t,s_t))}{\sum_a \exp\,(w(a\prime,s_t))} \qquad (1)$$

After each response, action values were updated with the prediction error based on the obtained outcome $r \in \{-1; 0; 1\}$. As the starting model (M1), we fitted an standard delta-learning model (*70*) in which action values were updated with prediction errors, i.e.,the deviation between the experienced outcome and expected outcome. This model contained two free parameters: the learning rate (ε) scaling the updating term and the feedback sensitivity (ρ) scaling the received outcome:

$$Q_t(a_t, s_t) = Q_{t-1}(a_t, s_t) + \varepsilon(\rho r - Q_{t-1}(a_t, s_t)) \qquad (2)$$

In this model, choice probabilities were fully determined by action values, without any bias. We assigned cue valence V_s to 0.5 for Win cues and -0.5 for Avoid cues and used cue valence scaled by participants' individual feedback sensitivity as initial action values Q_0. Unlike previous versions of the task (*3*), cue valences were not instructed, but had to be learned from outcomes, as well (*9*). Thus, until experiencing the first reward/ punishment for a cue, participants could not know its valence (and not learn from neutral feedback). Hence, for these trials, action values were multiplied with zero when computing choice probabilities. After the first encounter of a valenced outcome, action values were "unmuted" and started to influence choices probabilities, retrospectively considering all previous outcomes.

In M2, we added the Go bias parameter *b*, which accounted for individual differences in participants' overall propensity to make Go responses, to the action values Q, resulting in action weights w:

$$w(a_t, s_t) = \begin{cases} Q_t(a_t, s_t) + b & if\ a = Go \\ Q_t(a_t, s_t) & else \end{cases} \qquad (3)$$

In M3, we added a Pavlovian response bias π, scaling how positive/ negative cue valence (Pavlovian values) increased/ decreased the weights of Go responses:

$$w(a_t, s_t) = \begin{cases} Q_t(a_t, s_t) + b + \pi V(s) & if\ a = Go \\ Q_t(a_t, s_t) & else \end{cases} \tag{4}$$

We assigned cue valence to 0.5 for Win cues and -0.5 for Avoid cues. Cue valence became effective only once the participant had experienced the first reward/ punishment for that cue; beforehand, it was treated as zero. The Pavlovian response bias affected left-hand and right-hand Go responses similarly and thus reflected generalized activation/ inactivation by the cue valence.

In M4, we added an instrumental learning bias κ, increasing the learning rate for rewards after Go responses and decreasing it for punishments after NoGo responses:

$$\varepsilon = \begin{cases} \varepsilon_0 + \kappa & if\ r_t = 1\ and\ a = go \\ \varepsilon_0 - \kappa & if r_t = -1\ and\ a = nogo \\ \varepsilon_0 & else \end{cases} \tag{5}$$

The instrumental learning bias was specific to the response shown, thus reflecting a specific enhancement in action learning/ impairment in unlearning for that particular response.

In the model M5, we included both the Pavlovian response bias and the instrumental learning bias.

The hyperpriors were $X_\rho \sim \mathcal{N}(2{,}3)$, $X_\varepsilon \sim \mathcal{N}(0{,}2)$, $X_{b,\pi,\kappa} \sim \mathcal{N}(0{,}3)$. For computing the participant-level parameters, ρ was exponentiated to constrain it to positive values, and the inverse-logit transformation was applied to ε to constraint it to the range [0 1]. We made sure that the effect of κ on ε was symmetrical by computing it as:

$$\varepsilon = \begin{cases} \varepsilon_0 = inv.logit(\varepsilon) & \\ \varepsilon_{punished\ NoGo} = inv.logit(\varepsilon - \kappa) & if\ \varepsilon_0 < .5 \\ \varepsilon_{rewarded\ Go} = \varepsilon_0 + \left(\varepsilon_0 - \varepsilon_{punished\ NoGo}\right) & if\ \varepsilon_0 < .5 \end{cases} \tag{6}$$

$$\varepsilon = \begin{cases} \varepsilon_{rewarded\ Go} = inv.logit(\varepsilon + \kappa) & if\ \varepsilon_0 > .5 \\ \varepsilon_{punished\ NoGo} = \varepsilon_0 + (\varepsilon_0 - \varepsilon_{rewarded\ Go}) & if\ \varepsilon_0 > .5 \end{cases}$$

Model fitting and comparison

For model fitting and comparison, we used Hierarchical Bayesian inference as implemented in the CBM toolbox in Matlab (*71*). This approach combines hierarchical Bayesian parameter estimation with random-effects model comparison (*72*). The fitting procedure involves two steps, starting with the Laplace approximation of the model evidence to compute the group evidence, which quantifies how well each model fits the data while penalizing for model complexity. Both group-level and individual-level parameters are estimated using an iterative algorithm. We used wide Gaussian priors (see hyperpriors above) and exponential and sigmoid transforms to constrain parameter spaces. Subsequent random-effects model selection allows for the possibility that different models generated the data for different participants. Participants contribute to the group-level parameter estimation in proportion to how well a given model fits their data, quantified via a responsibility measure (i.e., the probability that the model at hand is responsible for generating data of the respective participant). This model-comparison approach has been shown to be less susceptible to the influence of outliers (*71*). We selected the "winning" model based on the protected exceedance probability.

Model validation

We assured that the winning model was able to reproduce the data, using the sampled combinations of participant-level parameter estimates to create 3,600 agents that "played" the task.

We employed two approaches to simulate the task: *posterior predictive model simulations* and *one-step-ahead model predictions*. In the posterior predictive model simulations, agents' choices were sampled probabilistically based on their action values, and outcomes probabilistically sampled based on their choices. This method ignores participant-specific choice histories and can thus yield choice/ outcome sequences that diverge considerably from participants' actual experiences. In contrast, one-step-ahead predictions use participants' actual choices and experienced outcomes in each trial to update action values. We simulated choices for each participant using both methods, which confirmed that the winning model M5 ("asymmetric pathways model") was able to qualitatively reproduce the data, while an alternative implementation of biased learning ("action priming model") failed to do so (see S03).

fMRI data acquisition

fMRI data were collected on a 3T Siemens Magnetom Prisma fit MRI scanner with a 64-channel head coil. During scanning, participants' heads were restricted using foam pillows and strips of adhesive tape were applied to participants' forehead to provide active motion feedback and minimize head movement (*73*). After two localizer scans to position slices, we collected functional scans with a whole-brain T2*-weighted sequence (68 axial-oblique slices, TR = 1400 ms, TE = 32 ms, voxel size 2.0 mm isotropic, interslice gap 0 mm, interleaved multiband slice acquisition with acceleration factor 4, FOV 210 mm, flip angle 75°, A/ P phase encoding direction). The first seven volumes of each run were automatically discarded. This sequence was chosen because of its balance between a short TR and relatively high spatial resolution, which was required to disentangle cue and outcome-related neural activity. Pilots using different sequences yielded that this sequence performed best in reducing signal loss in striatum.

Furthermore, after task completion, we removed the EEG cap and collected a high-resolution anatomical image using a T1-weighted MP-RAGE sequence (192 sagittal slices per slab, GRAPPA acceleration factor = 2, TI = 1100 ms, TR = 2300 ms, TE = 3.03 ms, FOV 256 mm, voxel size 1.0 mm isotropic, flip angle 8°) which was used to aid image registration, and a gradient fieldmap (GRE; TR = 614 ms, TE1 = 4.92 ms, voxel size 2.4 mm isotropic, flip angle 60°) for distortion correction. For one participant, no fieldmap was collected due to time constraints. At the end of each session, an additional DTI data collection took place; results will be reported elsewhere.

fMRI preprocessing

All fMRI pre-processing was performed in FSL 6.0.0. After cleaning images from non-brain tissue (brain-extraction with BET), we performed motion correction (MC-FLIRT), spatial smoothing (FWHM 3 mm), and used fieldmaps for B0 unwarping and distortion correction in orbitofrontal areas. We used ICA-AROMA (*74*) to automatically detect and reject independent components associated with head motion. Finally, images were high-pass filtered at 100 s and pre-whitened. After the first-level GLM analyses, we computed and applied co-registration of EPI images to high-resolution images (linearly with FLIRT using boundary-based registration) and to MNI152 2mm isotropic standard space (non-linearly with FNIRT using 12 DOF and 10 mm warp resolution).

ROI selection

For fMRI-informed EEG analyses, we first created a functional mask as the conjunction of the PE_{STD} and PE_{DIF} contrasts by thresholding both z-maps at $z > 3.1$, binarizing, and multiplying them (see S05). After visual inspection of the respective clusters, we created seven anatomical masks based on the probabilistic Harvard-Oxford Atlas (thresholded at 10%): striatum and ACC (see above), vmPFC (combined frontal pole, frontal medial cortex, and paracingulate gyrus), motor cortex (combined precentral and postcentral gyrus), PCC (Cingulate Gyrus, posterior division), ITG (Inferior Temporal Gyrus, posterior division, and Inferior Temporal Gyrus, temporooccipital part) and primary visual cortex (Lingual Gyrus, Occipital Fusiform Gyrus, Occipital Pole). We then multiplied this functional mask with each of the seven anatomical masks, returning seven masks focused on the respective significant clusters, which were then used for signal extraction. For the ACC mask, we manually excluded voxels in subgenual ACC belonging to a distinct cluster. Masks were back-transformed to each participant's native space.

For bar plots in Fig. 3A, we multiplied the anatomical masks of vmPFC and striatum specified above with the binarized outcome valence contrast.

fMRI analyses

For each participant, data were modelled using two event-related GLMs. First, we performed a model-based GLM in which used trial-by-trial estimates of biased PEs as regressors. Second, we used another model-free GLM in which we model all possible action x outcome combinations via outcome-locked categorical regressors while at the same time modeling response-locked left- and right-hand response regressors. This model free GLM also contained the valence contrast reported as an initial manipulation check.

In the model-based GLM, we used two model-based regressors that reflected the trial-by-trial prediction error (PE) update term. For this purpose, we extracted the group-level parameters of the best fitting computational model M5 (asymmetric pathways model) and used those parameters to compute the prediction error on every trial for every participant. Using the same parameter for each participant is warranted when testing for the same qualitative learning pattern across participants (*75*). Given that both standard (base model M1) and biased (winning model M5) PEs were highly correlated (mean correlation of 0.921 across participants, range 0.884–0.952), it appeared difficult to distinguish standard learning from biased learning. As a remedy, we decomposed the biased PE into the standard PE plus a difference term as $PE_{BIAS} = PE_{STD} + PE_{DIF}$ (*19*, *34*). Any region displaying truly biased learning should significantly encode *both* the standard PE term and the difference term. The standard PE and difference term were much less correlated (mean correlation of -0.020, range -0.326–0.237). To control for cue-related activation, we furthermore added four regressors spanned by crossing cue valence and performed action (Go response to Win cue, Go response to Avoid cue, NoGo response to Win cue, NoGo response to Avoid cue).

The model-free GLM included a separate regressor for each of the eight conditions obtained when crossing performed action (Go/ NoGo) and obtained outcome (reward/ no reward/ no punishment/ punishment). We fitted four contrasts: 1) one contrast comparing conditions with favorable (reward/ no punishment) and non-favorable (no reward/ punishment) outcomes, used as a quality check to identify regions that encoded outcome valence; 2) one contrast comparing Go vs. NoGo responses at the time of the outcome; 3) one contrast summing of left- and right-hand responses, reflecting Go vs. NoGo responses at the time of the response; and 4) one contrast subtracting right- from left-handed responses, reflecting lateralized motor activation. As this GLM resulted in empty regressors for several participants when fitted on a block level, making it impossible to use the data of the respective blocks on a higher level, we instead concatenated blocks and performed a single GLM per participant. We therefore registered the data from all blocks to the middle image of the first block (default reference volume in FSL) using MCFLIRT. The first and last 20 seconds of each block did not feature any task-related events, such that carry-over effects of task events in the design matrix from one block to another were not possible.

In both GLMs, we added four regressors of no interest: one for the motor response (left = +1, right = -1, NoGo = 0), one for error trials, one for outcome onset, and one for trials with invalid motor response (and no outcome respectively). We also added nine or more nuisance regressors: the six realignment parameters from motion correction, mean cerebrospinal fluid (CSF) signal, mean out-of-brain (OBO) signal, and a separate spike regressor for each volume with a relative displacement of more than 2 mm (occurred in 10 participants; in those participants: M = 7.40, range 1–29). For the model-free GLM, nuisance regressors were added separately for each block as well as an overall intercept per block. We convolved task regressors with double-gamma haemodynamic response function (HRF) and high-pass filtered the design matrix at 100 s.

First-level contrasts were fit in native space. Afterwards, co-registration and reslicing was applied to participants' contrast maps, which were then combined on a (participant and) group level using FSL's mixed effects models tool FLAME with a cluster-forming threshold of z > 3.1 and cluster-level error control at α < .05 (i.e., two one-sided tests with α < .025).

EEG data acquisition

We recorded EEG data with 64 channels (BrainCap-MR-3-0 64Ch-Standard; Easycap GmbH; Herrsching, Germany; international 10-20 layout, reference electrode at FCz) plus channels for electrocardiogram, heart rate, and respiration (used for MR artifact correction) at a sampling rate of 1000 Hz. We placed MRI-compatible EEG amplifiers (BrainAmp MR plus; Brain Products GmbH, Gilching, Germany) behind the MR scanner and attached cables to the participants once they were located in final position in the scanner. Furthermore, we fixated cables using sand-filled pillows to reduce artifacts induced through cable movement in the magnetic field. During functional scans, the MR helium pump was switched off to reduce EEG artifacts. After the scanning, we recorded the exact EEG electrode locations on participants' heads relative to three fiducial points using a Polhemus FASTRAK device. For four participants, no such data were available due to time constraints/ technical errors, in which case we used the average electrode locations of the remaining 32 participants.

EEG pre-processing

First, raw EEG data were cleaned from MR scanner and cardioballistic artifacts using BrainVisionAnalyzer (*76*). The rest of the pre-processing was performed in Fieldtrip (*77*). After rejecting channels with high residual MR noise (mean 4.8 channels per participant, range 1–13), we epoched trials into time windows of -1,400–2,000 ms relative to the onset of outcomes. Timing of this epochs was determined by the minimal inter-stimulus interval beforehand until the minimal inter-trial interval afterwards. Data was re-referenced to the grand average, which allowed us to recover the reference as channel FCz, and then band-pass filtered using a two-pass 4th order Butterworth IIR filter (Fieldtrip default) in the range of 0.5–35 Hz. These filter settings allowed us to distinguish the delta, theta, alpha, and beta band, while filtering out residual high-frequency MR noise. This low-pass filter cut-off was different from a previous analysis of this data in which we set it at 15 Hz (*32*) because in this analysis, we had a hypothesis on outcome valence encoding in the beta range. We then applied linear baseline correction based on the 200 ms prior to cue onset and used ICA to detect and reject independent components related to eye-blinks, saccades, head motion, and residual MR artifacts (mean number of rejected components per participant: 32.694, range 24–45). Afterwards, we manually rejected trials with residual motion (for all 36 participants: *M* = 117.722, range 11–499). Based on trial rejection, four participants for which more than 211 (33%) of trials were rejected were excluded from any further analyses (rejected trials after excluding those participants: *M* = 81.875, range 11–194). Finally, we computed a Laplacian filter with the spherical spline method to remove global noise (using the exact electrode positions recorded with Polhemus FASTRAK), which we also used to interpolate previously rejected channels. This filter attenuates more global signals (e.g., signal from deep sources or global noise) and noise (heart-beat and muscle artifacts) while accentuating more local effects (e.g., superficial sources).

EEG TF decomposition

We decomposed the trial-by-trial EEG time series into their time-frequency representations using 33 Hanning tapers between 1 and 33 Hz in steps of 1 Hz, every 25 ms from -1000 until 1,300 ms relative to outcome onset. We first zero-padded trials to a length of 8 sec. and then performed time-frequency decomposition in steps of 1 Hz by multiplying the Fourier transform of the trail with the Fourier transform of a Hanning taper of 400 ms width, centered around the time point of interest. This procedure results in an effective resolution of 2.5 Hz (Rayleigh frequency), interpolated in 1 Hz steps, which is more robust to the choice of exact frequency bins. To exclude the possibility of slow drifts in power over the time course of the experiment, we performed baseline correction across participants and trials by fitting a linear model for each channel/ frequency combination with trial number as predictor and the average power 250–50 ms before outcome onset as outcome, and subtracting the power predicted by this model from the data. This procedure is able to remove slow linear drifts in power over time from the data. In absence of such drifts, it is equivalent to correcting all trials by the grand mean across trials per frequency in the selected baseline time window. Afterwards, we averaged power over trials within each condition spanned by performed action (Go/ NoGo) and outcome (reward/ no reward/ no punishment/ punishment). We finally converted the average time-frequency data per condition to decibel to ensure that data across frequencies, time points, electrodes, and participants were on same scale.

EEG analyses

All analyses were performed on the average signal of a-priori selected channels Fz, FCz, and Cz based on (*9*, *32*). We again performed model-free and model-based analyses. For the model-free analyses, we sorted trials based on the performed action (Go/ NoGo) and obtained outcome (reward/ no reward/ no punishment/ punishment) and computed the mean TF power across trials for each of the resultant eight conditions for each participant. We tested whether theta power (average power 4–8 Hz) and beta power (average power 13–30 Hz) encoded outcome valence by contrasting favorable (reward/ no punishment) and unfavorable (no reward/ punishment) conditions (irrespective of the performed action). We also tested for differences between Go and NoGo responses in the lower alpha band (6–10 Hz). For all contrasts, we employed two-sided cluster-based permutation tests in a window from 0–1,000 ms relative to outcome onset. For beta power, results were driven by a cluster that was at the edge of 1,000 ms; to more accurately report the time span during which this cluster exceeded the threshold, we extended the time window to 1,300 ms in this particular analysis. Such tests are able to reject the null hypothesis of exchangeability of two experimental conditions, but they are not suited to precisely locate clusters in time-frequency space. Hence, interpretations are mostly based on the visual inspection of plots of the signal time courses.

For model-based analyses, similar to fMRI analyses, we used the group-level parameters from the best fitting computational model M5 to compute the trial-by-trial biased PE term and decomposed it into the standard PE term and the difference to the biased PE term. We used both terms as predictors in a multiple linear regression for each channel-time-frequency bin for each participant, and then performed one-sample cluster-based permutation-tests across the resultant *b*-maps of all participants (*78*). For further details on this procedure, see fMRI-inspired EEG analyses.

fMRI-informed EEG analyses

The BOLD signal is sluggish. It is thus hard to determine when different brain regions become active. In contrast, EEG provides much higher temporal resolution. A fruitful approach can be to identify distinct EEG correlates of the BOLD signal in different regions, allowing to test hypotheses about the temporal order in which regions might become active and modulated EEG power (*32*, *63*). Furthermore, by using the BOLD signal from different regions in a multiple linear regression, one can control for variance shared among regions (e.g., changes in global signal; variance due to task regressors) and test which region is the best unique predictor of a certain EEG signal. In such an analysis, any correlation between EEG and BOLD signal from a certain region reflects an association above and beyond those induced by task conditions.

We used the trial-by-trial BOLD signal in selected regions in a multiple linear regression to predict EEG signal over the scalp (*32*, *63*) (building on existing code from https://github.com/tuhauser/TAfT). As a first step, we extracted the volume-by-volume signal (first eigenvariate) from each of the seven regions identified to encode biased PEs (conjunction of PE_{STD} and PE_{DIF}: striatum, ACC, vmPFC, left motor cortex, PCC, left ITG, and primary visual cortex). We applied a highpass-filter at 128 s and regressed out nuisance regressors (6 realignment parameters, CSF, OOB, single volumes with strong motion, same as in the fMRI GLM). We then upsampled the signal by a factor 10, epoched it into trials of 8 s duration, and fitted a separate HRF (based on the SPM template) to each trial (58 upsampled data points), resulting in trial-by-trial regression weights reflecting the respective BOLD response. We then combined the regression weights of all trials and regions of a certain participant into a design matrix with trials as rows and the seven ROIs as columns, which we used to predict power at each time-frequency-channel bin. As further control variables, we added the behavioral PE_{STD} and PE_{DIF} regressors to the design matrix. All predictors and outcomes were demeaned such that the intercept became zero. Such a multiple linear regression was performed for each participant, resulting in a time-frequency-channel-ROI *b*-map reflecting the association between trial-by-trial BOLD signal and TF power at each time-frequency-channel bin. *B*-maps were Fisher-*z* transformed, which makes the sampling distribution of correlation coefficients approximately normal and allows for combining them across participants, and analyzed with a cluster-based one-sample permutation *t*-test (*78*) on the mean regression weights over channels Fz, FCz, and Cz across participants in the range of 0–1000 ms, 1–33 Hz. We first obtained a null distribution of maximal cluster mass statistics from 10000 permutations. For each permutation, we flipped the sign of the *b*-map of a random subset of participants, computed a separate *t*-test at each time-frequency bin (bins of 25 ms, 1 Hz) across participants (results in *t*-map), thresholded these maps at $|t| > 2$, and finally computed the maximal cluster mask statistic (sum of all *t*-values) for any cluster (adjacent voxels above threshold). Afterwards, we computed the same *t*-map for the real data, identified the cluster with the biggest cluster-mass statistic, and computed the corresponding *p*-value as number of permutations in the null distribution that were larger than the maximal cluster mass statistic in the real data.

EEG-informed fMRI analyses

For the EEG-informed fMRI analyses, we fit three additional GLMs for which we entered the trial-by-trial theta/ delta power (1–8 Hz), beta power (13–30 Hz), and lower alpha band power (6–10 Hz) as parametric regressors on top of the task regressors of the model-free GLM. These measures were created by using the 3-D (time-frequency-channel) *t*-map obtained when contrasting positive vs. negative outcomes (theta/ delta and beta) and Go vs. NoGo conditions (lower alpha band) as a linear filter. We enforced strict frequency cut-offs. For lower alpha band and beta, we used midfrontal channels (Fz/ FCz/ Cz). For theta/ delta power, given the topography that reached far beyond midfrontal channels and over the entire frontal scalp, we used a much wider ROI (AF3/ AF4/ AF7/ AF8/ F1/ F2/ F3/ F4/ F5/ F6/ F7/ F8/ FC1/ FC2/ FC3/ FC4/ FC5/ FC6/ FCz/ Fp1/ Fp2/ Fpz/ Fz). We extracted those maps and retained all voxels with t > 2. These masks were applied to the trial-by-trial time-frequency data to create weighted summary measures of the average power in the identified clusters in each trial. For trials for which EEG data was rejected, we imputed the participant mean value of the respective action (Go/ NoGo) x outcome (reward/ no reward/ no punishment/ punishment) condition. Note that this approach accentuates differences between conditions, which are already captured by the task regressors in the GLM, but decreases trial-by-trial variability within each condition, which is of interest in this analysis. This imputation approach is thus conservative. While trial-by-trial beta and theta power were largely uncorrelated, mean *r* = 0.104, range -0.118–0.283 across participants, and so were beta and alpha, mean *r* = 0.097, range -0.162–0.284 across participants, theta and alpha power moderately correlate, mean *r* = 0.412, range 0.121–0.836 across participants, warranting the use of a separate channel ROI for theta and using separate GLMs for each frequency band.

Analyses of behavior as a function of BOLD signal and EEG power

We used mixed-effects logistic regression to analyze “stay behavior”, i.e., whether participants repeated an action on the next encounter of the same cue, as a function of BOLD signal and EEG power in selected regions. For analyses featuring BOLD signal, we used the trial-by-trial HRF amplitude also used for fMRI-informed EEG analyses. For analyses featuring EEG, we used the trial-by-trial EEG power also used in the EEG-informed fMRI analyses.

References

1. P. Dayan, Y. Niv, B. Seymour, N. Daw, The misbehavior of value and the discipline of the will. *Neural Networks*. **19**, 1153–1160 (2006).

2. M. Guitart-Masip, E. Duzel, R. Dolan, P. Dayan, Action versus valence in decision making. *Trends Cogn. Sci.* **18**, 194–202 (2014).

3. J. C. Swart, M. I. Froböse, J. L. Cook, D. E. M. Geurts, M. J. Frank, R. Cools, H. E. den Ouden, Catecholaminergic challenge uncovers distinct Pavlovian and instrumental mechanisms of motivated (in)action. *Elife*. **6**, 1–54 (2017).

4. A. Mkrtchian, J. Aylward, P. Dayan, J. P. Roiser, O. J. Robinson, Modeling avoidance in mood and anxiety disorders using reinforcement learning. *Biol. Psychiatry*. **82**, 532–539 (2017).

5. Q. J. M. Huys, M. Gölzer, E. Friedel, A. Heinz, R. Cools, P. Dayan, R. J. Dolan, The specificity of Pavlovian regulation is associated with recovery from depression. *Psychol. Med.* **46**, 1027–1035 (2016).

6. Q. J. M. Huys, R. Cools, M. Gölzer, E. Friedel, A. Heinz, R. J. Dolan, P. Dayan, Disentangling the roles of approach, activation and valence in instrumental and Pavlovian responding. *PLoS Comput. Biol.* **7**, e1002028 (2011).

7. Y.-L. Boureau, P. Sokol-Hessner, N. D. Daw, Deciding how to decide: Self-control and meta-decision making. *Trends Cogn. Sci.* **19**, 700–710 (2015).

8. M. Guitart-Masip, Q. J. M. Huys, L. Fuentemilla, P. Dayan, E. Duzel, R. J. Dolan, Go and no-go

learning in reward and punishment: Interactions between affect and effect. *Neuroimage*. **62**, 154–166 (2012).

9. J. C. Swart, M. J. Frank, J. I. Määttä, O. Jensen, R. Cools, H. E. M. den Ouden, Frontal network dynamics reflect neurocomputational mechanisms for reducing maladaptive biases in motivated action. *PLOS Biol.* **16**, e2005979 (2018).

10. L. de Boer, J. Axelsson, R. Chowdhury, K. Riklund, R. J. Dolan, L. Nyberg, L. Bäckman, M. Guitart-Masip, Dorsal striatal dopamine D1 receptor availability predicts an instrumental bias in action learning. *Proc. Natl. Acad. Sci.* **116**, 261–270 (2019).

11. D. R. Williams, H. Williams, Auto-maintenance in the pigeon: Sustained pecking despite contingent non-reinforcement. *J. Exp. Anal. Behav.* **12**, 511–520 (1969).

12. P. L. Brown, H. M. Jenkins, Autoshaping of pigeon's key-peck. *J. Exp. Anal. Behav.* **11**, 1–8 (1968).

13. I. Ritov, J. Baron, Reluctance to vaccinate: Omission bias and ambiguity. *J. Behav. Decis. Mak.* **3**, 263–277 (1990).

14. M. Zeelenberg, J. van der Pligt, N. K. de Vries, Attributions of responsibility and affective reactions to decision outcomes. *Acta Psychol. (Amst).* **104**, 303–315 (2000).

15. M. J. Frank, Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. *J. Cogn. Neurosci.* **17**, 51–72 (2005).

16. A. G. E. Collins, M. J. Frank, Opponent actor learning (OpAL): Modeling interactive effects of striatal dopamine on reinforcement learning and choice incentive. *Psychol. Rev.* **121**, 337–366 (2014).

17. B. B. Doll, W. J. Jacobs, A. G. Sanfey, M. J. Frank, Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Res.* **1299**, 74–94 (2009).

18. L. Y. Atlas, B. B. Doll, J. Li, N. D. Daw, E. A. Phelps, Instructed knowledge shapes feedback-driven aversive learning in striatum and orbitofrontal cortex, but not the amygdala. *Elife*. **5**, 1–26 (2016).

19. N. D. Daw, S. J. Gershman, B. Seymour, P. Dayan, R. J. Dolan, Model-based influences on humans' choices and striatal prediction errors. *Neuron*. **69**, 1204–1215 (2011).

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

21. P. Piray, I. Toni, R. Cools, Human choice strategy varies with anatomical projections from ventromedial prefrontal cortex to medial striatum. *J. Neurosci.* **36**, 2857–2867 (2016).

22. T. E. J. Behrens, M. W. Woolrich, M. E. Walton, M. F. S. Rushworth, Learning the value of information in an uncertain world. *Nat. Neurosci.* **10**, 1214–1221 (2007).

23. D. Meder, N. Kolling, L. Verhagen, M. K. Wittmann, J. Scholl, K. H. Madsen, O. J. Hulme, T. E. J. Behrens, M. F. S. Rushworth, Simultaneous representation of a spectrum of dynamically changing value estimates during decision making. *Nat. Commun.* **8**, 1942 (2017).

24. A. J. van Nuland, R. C. Helmich, M. F. Dirkx, H. Zach, I. Toni, R. Cools, H. E. M. den Ouden, Effects of dopamine on reinforcement learning in Parkinson's disease depend on motor phenotype. *Brain*, 1–13 (2020).

25. I. van de Vijver, K. R. Ridderinkhof, M. X. Cohen, Frontal oscillatory dynamics predict feedback

learning and action adjustment. *J. Cogn. Neurosci.* **23**, 4106–4121 (2011).

26. M. X. Cohen, K. A. Wilmes, I. van de Vijver, Cortical electrophysiological network dynamics of feedback learning. *Trends Cogn. Sci.* **15**, 558–566 (2011).

27. J. F. Cavanagh, M. J. Frank, T. J. Klein, J. J. B. Allen, Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. *Neuroimage*. **49**, 3198–3209 (2010).

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

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

30. E. M. Bernat, L. D. Nelson, A. R. Baskin-Sommers, Time-frequency theta and delta measures index separable components of feedback processing in a gambling task. *Psychophysiology*. **52**, 626–637 (2015).

31. J. Marco-Pallarés, T. F. Münte, A. Rodríguez-Fornells, The role of high-frequency oscillatory activity in reward processing and learning. *Neurosci. Biobehav. Rev.* **49**, 1–7 (2015).

32. J. Algermissen, J. C. Swart, R. Scheeringa, R. Cools, H. E. M. den Ouden, Striatal BOLD signal and midfrontal theta oscillations express motivation for action. *Cereb. Cortex* (2021).

33. J. Cockburn, A. G. E. Collins, M. J. Frank, A reinforcement learning mechanism responsible for the valuation of free choice. *Neuron*. **83**, 551–557 (2014).

34. B. C. Wittmann, N. D. Daw, B. Seymour, R. J. Dolan, Striatal activity underlies novelty-based choice in humans. *Neuron*. **58**, 967–973 (2008).

35. O. Bartra, J. T. McGuire, J. W. Kable, The valuation system: A coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. *Neuroimage*. **76**, 412–427 (2013).

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

37. M. J. Sharpe, C. Y. Chang, M. A. Liu, H. M. Batchelor, L. E. Mueller, J. L. Jones, Y. Niv, G. Schoenbaum, Dopamine transients are sufficient and necessary for acquisition of model-based associations. *Nat. Neurosci.* **20**, 735–742 (2017).

38. M. J. Sharpe, T. Stalnaker, N. W. Schuck, S. Killcross, G. Schoenbaum, Y. Niv, An integrated model of action selection: Distinct modes of cortical control of striatal decision making. *Annu. Rev. Psychol.* **70**, 53–76 (2019).

39. A. Pasupathy, E. K. Miller, Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*. **433**, 873–876 (2005).

40. E. G. Antzoulatos, E. K. Miller, Increases in functional connectivity between prefrontal cortex and striatum during category learning. *Neuron*. **83**, 216–225 (2014).

41. M. Seo, E. Lee, B. B. Averbeck, Action selection and action value in frontal-striatal circuits. *Neuron*. **74**, 947–960 (2012).

42. J. D. Howard, R. Reynolds, D. E. Smith, J. L. Voss, G. Schoenbaum, T. Kahnt, Targeted stimulation of human orbitofrontal networks disrupts outcome-guided behavior. *Curr. Biol.* **30**, 490-498.e4 (2020).

43. M. R. van Schouwenburg, J. O'Shea, R. B. Mars, M. F. S. Rushworth, R. Cools, Controlling

human striatal cognitive function via the frontal cortex. *J. Neurosci.* **32**, 5631–5637 (2012).

44. E. Fouragnan, C. Retzler, K. Mullinger, M. G. Philiastides, Two spatiotemporally distinct value systems shape reward-based learning in the human brain. *Nat. Commun.* **6**, 8107 (2015).

45. W. H. Alexander, J. W. Brown, Medial prefrontal cortex as an action-outcome predictor. *Nat. Neurosci.* **14**, 1338–1344 (2011).

46. W. H. Alexander, J. W. Brown, Frontal cortex function as derived from hierarchical predictive coding. *Sci. Rep.* **8**, 3843 (2018).

47. P. Enel, J. D. Wallis, E. L. Rich, Stable and dynamic representations of value in the prefrontal cortex. *Elife*. **9**, 1–23 (2020).

48. S. Vyas, D. J. O'Shea, S. I. Ryu, K. V. Shenoy, Causal role of motor preparation during error-driven learning. *Neuron*, 1–11 (2020).

49. R. Shadmehr, M. A. Smith, J. W. Krakauer, Error correction, sensory prediction, and adaptation in motor control. *Annu. Rev. Neurosci.* **33**, 89–108 (2010).

50. S. Sadaghiani, R. Scheeringa, K. Lehongre, B. Morillon, A.-L. Giraud, A. Kleinschmidt, Intrinsic connectivity networks, alpha oscillations, and tonic alertness: A simultaneous electroencephalography/functional magnetic resonance imaging study. *J. Neurosci.* **30**, 10243–10250 (2010).

51. C. Andreou, H. Frielinghaus, J. Rauh, M. Mußmann, S. Vauth, P. Braun, G. Leicht, C. Mulert, Theta and high-beta networks for feedback processing: A simultaneous EEG-fMRI study in healthy male subjects. *Transl. Psychiatry*. **7**, e1016–e1016 (2017).

52. J. Marco-Pallarés, D. Cucurell, T. Cunillera, R. García, A. Andrés-Pueyo, T. F. Münte, A. Rodríguez-Fornells, Human oscillatory activity associated to reward processing in a gambling task. *Neuropsychologia*. **46**, 241–248 (2008).

53. R. Scheeringa, M. C. M. Bastiaansen, K. M. Petersson, R. Oostenveld, D. G. Norris, P. Hagoort, Frontal theta EEG activity correlates negatively with the default mode network in resting state. *Int. J. Psychophysiol.* **67**, 242–251 (2008).

54. R. Scheeringa, K. M. Petersson, R. Oostenveld, D. G. Norris, P. Hagoort, M. C. M. Bastiaansen, Trial-by-trial coupling between EEG and BOLD identifies networks related to alpha and theta EEG power increases during working memory maintenance. *Neuroimage*. **44**, 1224–1238 (2009).

55. E. Fouragnan, C. Retzler, M. G. Philiastides, Separate neural representations of prediction error valence and surprise: Evidence from an fMRI meta-analysis. *Hum. Brain Mapp.* (2018), doi:10.1002/hbm.24047.

56. S. N. Haber, The primate basal ganglia: Parallel and integrative networks. *J. Chem. Neuroanat.* **26**, 317–330 (2003).

57. A. K. Engel, P. Fries, Beta-band oscillations—signalling the status quo? *Curr. Opin. Neurobiol.* **20**, 156–165 (2010).

58. J. Feingold, D. J. Gibson, B. Depasquale, A. M. Graybiel, Bursts of beta oscillation differentiate postperformance activity in the striatum and motor cortex of monkeys performing movement tasks. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 13687–13692 (2015).

59. T. U. Hauser, R. Iannaccone, P. Stämpfli, R. Drechsler, D. Brandeis, S. Walitza, S. Brem, The feedback-related negativity (FRN) revisited: New insights into the localization, meaning and

network organization. *Neuroimage*. **84**, 159–168 (2014).

60. J. R. Wessel, A. R. Aron, On the globality of motor suppression: Unexpected events and their influence on behavior and cognition. *Neuron*. **93**, 259–280 (2017).

61. N. Trudel, J. Scholl, M. C. Klein-Flügge, E. Fouragnan, L. Tankelevitch, M. K. Wittmann, M. F. S. Rushworth, Polarity of uncertainty representation during exploration and exploitation in ventromedial prefrontal cortex. *Nat. Hum. Behav.* (2020), doi:10.1038/s41562-020-0929-3.

62. P. Domenech, S. Rheims, E. Koechlin, Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex. *Science (80-.).* **369**, eabb0184 (2020).

63. T. U. Hauser, L. T. Hunt, R. Iannaccone, S. Walitza, D. Brandeis, S. Brem, R. J. Dolan, Temporally dissociable contributions of human medial prefrontal subregions to reward-guided learning. *J. Neurosci.* **35**, 11209–11220 (2015).

64. D. Foti, A. Weinberg, J. Dien, G. Hajcak, Event-related potential activity in the basal ganglia differentiates rewards from nonrewards: Temporospatial principal components analysis and source localization of the feedback negativity. *Hum. Brain Mapp.* **32**, 2207–2216 (2011).

65. M. X. Cohen, J. F. Cavanagh, H. A. Slagter, Event-related potential activity in the basal ganglia differentiates rewards from nonrewards: Temporospatial principal components analysis and source localization of the feedback negativity: Commentary. *Hum. Brain Mapp.* **32**, 2270–2271 (2011).

66. K. Amemori, S. Amemori, D. J. Gibson, A. M. Graybiel, Striatal microstimulation induces persistent and repetitive negative decision-making predicted by striatal beta-band oscillation. *Neuron*. **99**, 829-841.e6 (2018).

67. K. Amemori, S. Amemori, D. J. Gibson, A. M. Graybiel, Striatal beta oscillation and neuronal activity in the primate caudate nucleus differentially represent valence and arousal under approach-avoidance conflict. *Front. Neurosci.* **14**, 1–17 (2020).

68. J. F. Cavanagh, I. Eisenberg, M. Guitart-Masip, Q. J. M. Huys, M. J. Frank, Frontal theta overrides Pavlovian learning biases. *J. Neurosci.* **33**, 8541–8548 (2013).

69. D. J. Barr, R. Levy, C. Scheepers, H. J. Tily, Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* **68**, 255–278 (2013).

70. R. A. Rescorla, A. R. Wagner, in *Classical Conditioning II�:Current Research and Theory*, A. H. Black, W. F. Prokasy, Eds. (Appleton-Century-Crofts., New York, NY, 1972; http://homepage.mac.com/sanagnos/rescorlawagner1972.pdf), vol. 21, pp. 64–99.

71. P. Piray, A. Dezfouli, T. Heskes, M. J. Frank, N. D. Daw, Hierarchical Bayesian inference for concurrent model fitting and comparison for group studies. *PLOS Comput. Biol.* **15**, e1007043 (2019).

72. K. E. Stephan, W. D. Penny, J. Daunizeau, R. J. Moran, K. J. Friston, Bayesian model selection for group studies. *Neuroimage*. **46**, 1004–1017 (2009).

73. F. Krause, C. Benjamins, J. Eck, M. Lührs, R. Hoof, R. Goebel, Active head motion reduction in magnetic resonance imaging using tactile feedback. *Hum. Brain Mapp.* **40**, 4026–4037 (2019).

74. R. H. R. Pruim, M. Mennes, D. van Rooij, A. Llera, J. K. Buitelaar, C. F. Beckmann, ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. *Neuroimage*. **112**, 267–277 (2015).

75. R. C. Wilson, Y. Niv, Is model fitting necessary for model-based fMRI? *PLOS Comput. Biol.* **11**,

e1004237 (2015).

76. P. J. Allen, O. Josephs, R. Turner, A method for removing imaging artifact from continuous EEG recorded during functional MRI. *Neuroimage*. **12**, 230–239 (2000).

77. R. Oostenveld, P. Fries, E. Maris, J.-M. Schoffelen, FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* **2011**, 1–9 (2011).

78. L. T. Hunt, M. W. Woolrich, M. F. S. Rushworth, T. E. J. Behrens, Trial-type dependent frames of reference for value comparison. *PLoS Comput. Biol.* **9**, e1003225 (2013).

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

80. M. R. Nassar, M. J. Frank, Taming the beast: Extracting generalizable knowledge from computational models of cognition. *Curr. Opin. Behav. Sci.* **11**, 49–54 (2016).

81. M. X. Cohen, T. H. Donner, Midfrontal conflict-related theta-band power reflects neural oscillations that predict behavior. *J. Neurophysiol.* **110**, 2752–2763 (2013).

82. G. H. Proudfit, The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology*. **52**, 449–459 (2015).

83. K. Paul, E. Vassena, M. C. Severo, G. Pourtois, Dissociable effects of reward magnitude on fronto-medial theta and FRN during performance monitoring. *Psychophysiology* (2019), doi:10.1111/psyp.13481.

84. T. D. Sambrook, J. Goslin, Principal components analysis of reward prediction errors in a reinforcement learning task. *Neuroimage*. **124**, 276–286 (2016).

85. N. Yeung, A. G. Sanfey, Independent coding of reward magnitude and valence in the human brain. *J. Neurosci.* **24**, 6258–6264 (2004).

86. L. Kreussel, J. Hewig, N. Kretschmer, H. Hecht, M. G. H. Coles, W. H. R. Miltner, The influence of the magnitude, probability, and valence of potential wins and losses on the amplitude of the feedback negativity. *Psychophysiology*. **49**, 207–219 (2012).

87. A. Sato, A. Yasuda, H. Ohira, K. Miyawaki, M. Nishikawa, H. Kumano, T. Kuboki, Effects of value and reward magnitude on feedback negativity and P300. *Neuroreport*. **16**, 407–411 (2005).

88. D. Tanner, K. Morgan-Short, S. J. Luck, How inappropriate high-pass filters can produce artifactual effects and incorrect conclusions in ERP studies of language and cognition. *Psychophysiology*. **52**, 997–1009 (2015).

89. Y. Wu, X. Zhou, The P300 and reward valence, magnitude, and expectancy in outcome evaluation. *Brain Res.* **1286**, 114–122 (2009).

90. R. Scheeringa, P. Fries, K.-M. Petersson, R. Oostenveld, I. Grothe, D. G. Norris, P. Hagoort, M. C. M. Bastiaansen, Neuronal dynamics underlying high-and low-frequency EEG oscillations contribute independently to the human BOLD signal. *Neuron*. **69**, 572–583 (2011).

91. J. M. Zumer, R. Scheeringa, J.-M. Schoffelen, D. G. Norris, O. Jensen, Occipital alpha activity during stimulus processing gates the information flow to object-selective cortex. *PLoS Biol.* **12**, e1001965 (2014).

92. M. T. Jurkiewicz, W. C. Gaetz, A. C. Bostan, D. Cheyne, Post-movement beta rebound is generated in motor cortex: Evidence from neuromagnetic recordings. *Neuroimage*. **32**, 1281–1289 (2006).

93. P. Ritter, M. Moosmann, A. Villringer, Rolandic alpha and beta EEG rhythms' strengths are inversely related to fMRI-BOLD signal in primary somatosensory and motor cortex. *Hum. Brain Mapp.* **30**, 1168–1187 (2009).

94. W. Klimesch, EEG alpha and theta oscillations reflect cognitive and memory performance: A review and analysis. *Brain Res. Rev.* **29**, 169–195 (1999).

Acknowledgments

We thank Emma van Dijk for assistance with data collection, Michael J. Frank for helpful discussions, and the weekly Donders M/EEG meeting for discussions of these results and many helpful suggestions.

Funding

This work was supported by:
Netherlands Organization for Scientific Research (NWO) research talent grant 406-14-028 (JCS)
Netherlands Organization for Scientific Research (NWO) VENI grant 451-12-021 (RS)
Netherlands Organization for Scientific Research (NWO) VICI grant 453-14-005 (RC)
Netherlands Organization for Scientific Research (NWO) Ammodo KNAW Award 2017 (RC)
James S. McDonnell Foundation James McDonnell Scholar Award (RC)
Netherlands Organization for Scientific Research (NWO) VIDI grant 452-17-016 (HEMDO)

Author contributions

Conceptualization: JA, JCS, RC, HEMDO
Data curation: JA
Formal analysis: JA
Funding acquisition: JCS, RC, HEMDO
Investigation: JA, JCS
Methodology: JA, HEMDO
Project administration: JA, JCS, HEMDO
Resources: RC, HEMDO
Software: JA, JCS, HEMDO
Supervision: JCS, RS, RC, HEMDO
Validation: JA, JCS, RS, RC, HEMDO
Visualization: JA
Writing – original draft: JA, HEMDO
Writing – review & editing: JA, JCS, RS, RC, HEMDO

Competing interests

Authors declare that they have no competing interests.

Data availability statement

All data and code will be made available upon manuscript acceptance.
Group-level unthresholded fMRI z-maps are available on Neurovault (https://neurovault.org/collections/11184/).

Supplementary Materials to “Biased credit assignment in motivational learning biases arises through prefrontal influences on striatal learning”

S01: Behavioral, fMRI, and EEG analyses with only the 29 participants included in EEG-fMRI analyses

We repeated the behavioral, fMRI, and EEG analyses reported in the main text while excluding the seven participants that were also not included in the fMRI-inspired EEG analyses in the main text: (a) two participants due to fMRI co-registration failure, which were also not included in the fMRI-only analyses; (b) four further participants who exhibited excessive residual noise in their EEG data (> 33% rejected trials) and were thus also not included in the EEG-only analyses, and finally (c) one more participant who (together with four other participants already excluded) exhibited regression weights for every regressor about ten times larger than for other participants.

Participants in this subgroup learned the task, reflected in a significant main effect of required action on responses, b = 0.896, SE = 0.129, $\chi^2(1)$ = 28.398, p < .001, and exhibited motivational biases, reflected in a significant main effect of cue valence on responses, b = 0.439, SE = 0.084, $\chi^2(1)$ = 19.308, p < .001. The interaction between required action and cue valence was not significant, b = 0.025, SE = 0.085, $\chi^2(1)$ = 0.111, p = .739.

Participants in this subgroup also showed biased learning: They were more likely to repeat an action after a favorable outcome (main effect of outcome valence: b = .0553, SE = 0.059, $\chi^2(1)$ = 40.920, p < .001. After salient outcomes, they adjusted their responses more strongly after feedback on Go than on NoGo responses, in line with our model of biased learning and as reflected in a significant three-way interaction between action, salience, and valence, b = 0.266, SE = 0.055, $\chi^2(1)$ = 16.862, p < .001. When only analyzing trials with salient outcomes, outcome valence was more likely to affect response repetition following Go relative to NoGo responses, b = 0.324, SE = 0.079, $\chi^2(1)$ = 13.266, p < .001, with a stronger effect of outcome valence after Go responses, b = 1.342, SE = 0.120, $\chi^2(1)$ = 49.003, p = .001, than NoGo responses, b = 0.693, SE = 0.129, $\chi^2(1)$ = 18.988, p < .001.

In this subgroup of participants, Bayesian model selection clearly favored the full asymmetric pathways models featuring response and learning biases (M5, model frequency: 81.81%, protected exceedance probability: 100%). In sum, behavioral results were qualitatively identical when analyzing only this subgroup of only 29 participants.

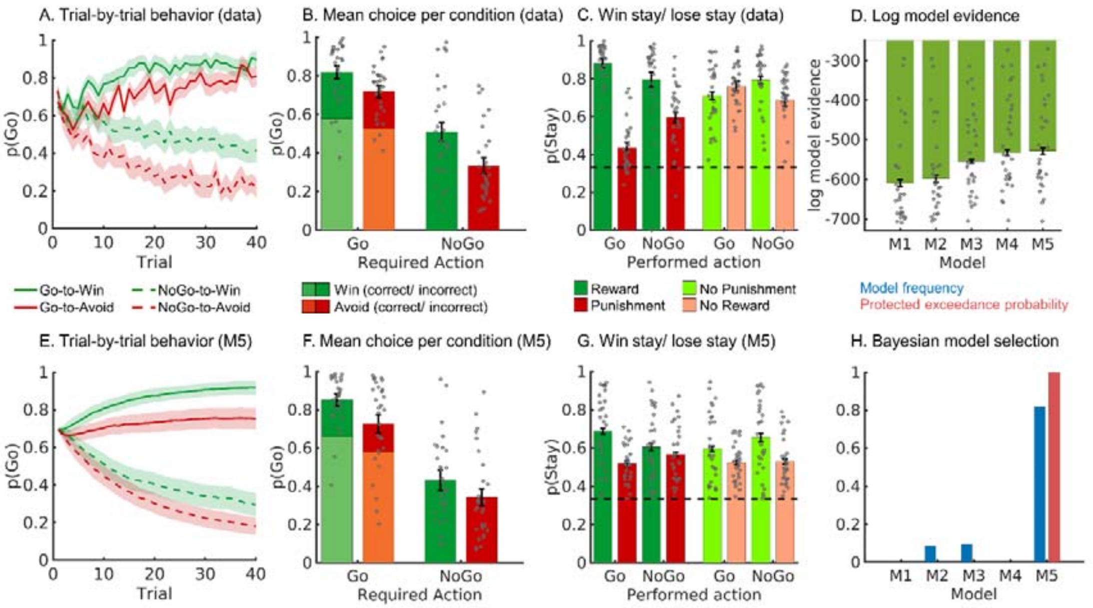

Figure S01A. Behavioral performance in the subgroup of 29 participants included in the fMRI-inspired EEG analyses. (A) Trial-by-trial proportion of Go responses (±SEM across participants) for Go cues (solid lines) and NoGo cues (dashed lines). The motivational bias is already present from very early trials onwards, as participants made more Go responses to Win than Avoid cues (i.e., green lines are above red lines). Additionally, participants clearly learn whether to make a Go response or not (proportion of Go responses increases for Go cues and decreases for NoGo cues). (B) Mean (±SEM across participants) proportion Go responses per cue condition (points are individual participants' means). C) Probability to repeat a response ("stay") on the next encounter of the same cue as a function of action and outcome. Learning is reflected in higher probability of staying after positive outcomes than after negative outcomes (main effect of outcome valence). Biased learning is evident in learning from salient outcomes, where this valence effect was stronger after Go responses than NoGo responses. Dashed line indicates chance level choice (p_{Stay} = 0.33). (D) Log-model evidence favors the asymmetric pathways model (M5 over simpler models (M1-M4). (E-G) Trial-by-trial proportion of Go responses, mean proportion Go responses, and probability of staying based on one-step-ahead predictions using parameters (hierarchical Bayesian inference) of the winning model (asymmetric pathways model, M5). (H) Model frequency and protected exceedance probability indicate best fit for model M5 (asymmetric pathways model), in line with log model evidence.

Regarding fMRI findings, we first repeated the model-free GLM just contrasting favorable and non-favorable outcomes. BOLD signal was higher for favorable than non-favorable outcomes in five clusters, namely in vmPFC, striatum, amygdala, and hippocampus (z_{max} = 5.65, p = 2.24e-25, 6110 voxels, MNI coordinates xyz = [6 30 -12]), left superior lateral occipital cortex (z_{max} = 4.40, p = .00144, 367 voxels, xyz = [-46 -68 46]), right occipital pole (z_{max} = 4.45, p = .00154, 363 voxels, xyz = [12 -92 -12]), posterior cingulate cortex (z_{max} = 4.36, p = .00181, 353 voxels, xyz = [-2 -48 28]), and left middle temporal gyrus (z_{max} = 4.63, p = .00548, 289 voxels, xyz = [-60 -10 -16]). The clusters in left slOCC, PCC, and left MTG emerged anew compared to the original analysis comprising 34 participants. Also, compared to the original analysis, clusters in left orbitofrontal cortex and left superior frontal gyrus were merged with the cluster in vmPFC. In sum, all clusters from the original analysis were found back, plus some additional clusters.

There was also one cluster in right orbitofrontal cortex (z_{max} = 4.37, p = .0209, 217 voxels, xyz = [30 62 -2]) in which BOLD signal was higher for non-favorable than favorable outcomes. Compared to the original analysis comprising 34 participants, clusters in precuneous and right superior frontal gyrus were not significant.

In the model-based GLM featuring regressors for standard PEs and the difference term towards biased PEs, BOLD signal correlated with standard PEs in ten clusters, namely in vmPFC, striatum, bilateral amygdala and hippocampus (z_{max} = 6.04, p = .4.78e-44, 8848 voxels, xyz = [12 14 -6]), left superior frontal gyrus (z_{max} = 5.58, p = 3.5e-10, 1043 voxels, xyz = [-18 34 52]), left occipital pole and lingual gyrus (z_{max} = 6.23, p = 7.18e-10, 998 voxels, xyz = [10 -92 -10]), posterior cingulate cortex (z_{max} = 5.12, p = 8.57e-10, 987 voxels, xyz = [4 -36 48]), left inferior temporal gyrus (z_{max} = 5.03, p = 7.07e-09, 859 voxels, xyz = [-52 -46 -10]), right anterior middle temporal gyrus (z_{max} = 5.32, p = .000292, 314 voxels, xyz = [62 -4 -16]), right cerebellum (z_{max} = 5.32, p = .002228, 231 voxels, xyz = [44 -72 -40]), left superior lateral occipital cortex (z_{max} = 4.69, p = .00322, 218 voxels, xyz = [-46 -74 -38]), right caudate (z_{max} = 4.33, p = .00538, 199 voxels, xyz = [20 12 22]), and right middle temporal gyrus (z_{max} = 4.09, p = .0129, 189 voxels, xyz = [54 -38 -12]). The clusters in left superior lateral occipital cortex, right caudate, and right posterior middle temporal gyrus emerged anew by splitting from larger clusters visible in the original analysis based on 34 participants. Vice versa, the cluster in left middle temporal gyrus reported for the original analysis was merged with a bigger cluster in the analysis of only 29 participants. The clusters in postcentral gyrus and ACC observed in the original analysis based on 34 participants were not significant anymore; however, they were still visible at a level of z > 3.1 uncorrected.

BOLD signal correlated significantly negatively with standard PEs in a single cluster in right superior frontal gyrus (z_{max} = 5.04, p = .00771, 186 voxels, xyz = [6 26 64]), similar to the respective cluster reported in the original analysis. In contrast, the clusters in right occipital pole, intracalcarine cortex, and left inferior lateral occipital cortex were not significant any more, though visible at a level of z > 3.1 uncorrected.

BOLD signal in six clusters correlated significantly positively with the difference term towards biased PEs, namely in large parts of cortex and subcortex including striatum (z_{max} = 6-54, p = 0, 29428 voxels, xyz = [34 -84 20]), dorsomedial prefrontal cortex (z_{max} = 5.94, p = 2.69e-40, 7001 voxels, xyz = [6 22 34]), right insula (z_{max} = 5.76, p = 7.84e-27, 3847 voxels, xyz = [34 20 -8]), thalamus and brainstem (z_{max} = 5.10, p = 4.06e-18, 2169 voxels, xyz = [4 -30 0]), left caudate (z_{max} = 4.71, p = .000188, 305 voxels, xyz = [-12 8 6]) and another cluster in brainstem (z_{max} = 4.05, p = .0151, 160 voxels, xyz = [4 -30 -30]). Clusters in dmPFC, right insula, and left caudate split from larger clusters reported in the original analysis. Vice versa, the cluster in left insula reported in the original analysis merged with the largest cluster. The clusters in right middle temporal gyrus and right insula were missing in the analysis of only 29 participants, but visible at a level of z > 3.1 uncorrected.

BOLD signal in three clusters correlated significantly negatively with the difference term towards biased PEs, namely in vmPFC (z_{max} = 4.23, p = .0051, 185 voxels, xyz = [-12 48 -6]), left hippocampus (z_{max} = 4.58, p = .00857, 168 voxels, xyz = [-26 -14 -22]), and left medial temporal gyrus (z_{max} = 4.30, p = .0172, 146 voxels, xyz = [-62 -4 -16]). Compared to the original analysis, the cluster in vmPFC emerged anew.

When computing the conjunction between both (positive) contrasts, BOLD signal encoded both the standard and the difference in four clusters, namely in vmPFC, bilateral striatum, bilateral ITG, and V1. Clusters in ACC, left motor cortex, and PCC were not significant any more (because they were z > 3.1, but not significant after cluster correction in the standard PE contrast). However, new (though rather small) clusters of biased PE encoding emerged in right insula, left amygdala, and left OFC. In sum, results when analyzing only this subgroup of only 29 participants were largely similar to results based on the full sample; however, clusters of biased PE encoding in left motor cortex, ACC, and PCC were small and thus did not survive cluster correction in this subgroup.

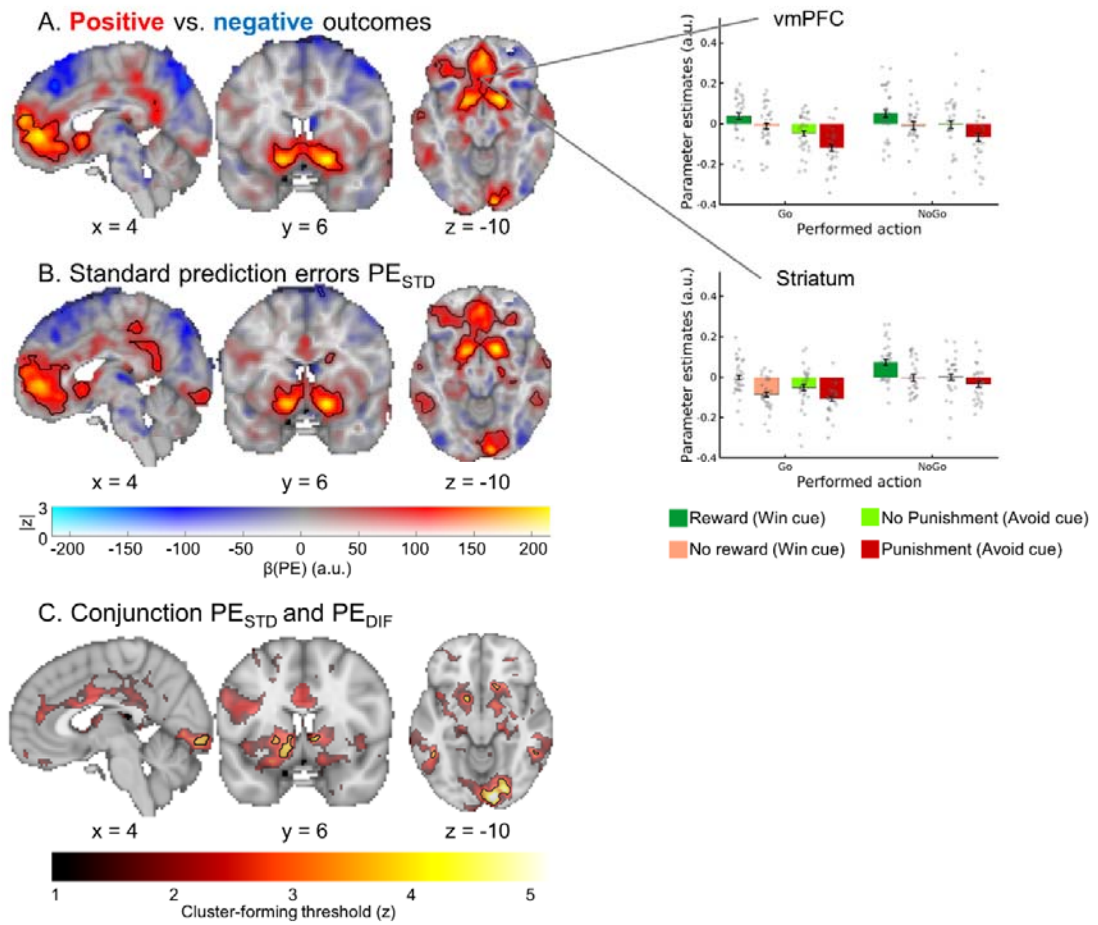

Figure S01B. BOLD signal reflecting outcome processing in the subgroup of 29 participants included in the fMRI-inspired EEG analyses. (A) BOLD signal was higher for favorable outcomes (rewards, no punishments) compared with unfavorable outcomes (no rewards, punishments) in a range of regions including bilateral ventral striatum and vmPFC. BOLD effects displayed using a dual-coding data visualization approach with color indicating the parameter estimates and opacity the associated z-statistics. Significant clusters are surrounded by black edges. Bar plots show parameter estimates per action x outcome condition (±SEM across participants) (B) When using the trial-by-trial PEs participants experienced as model-based regressors in our GLM, positive PE correlations occurred in several regions including importantly the ventral striatum, vmPFC, PCC and ACC. (C) Left panel: Regions encoding both the standard PE term and the difference term to biased PEs (conjunction) at different cluster-forming thresholds (color). Clusters significant at a threshold of z > 3.1 are surrounded by black edges. In bilateral striatum, vmPFC, bilateral ITG, and primary visual cortex, BOLD is significantly better explained by biased learning than by standard learning. Clusters in ACC, left motor cortex, and PCC are not significant any more.

Regarding EEG findings in this subgroup, both midfrontal theta and beta power reflected outcome valence: Theta power was higher for unfavorable than favorable outcomes (driven by a cluster around 225–500 ms, *p* = .002), while beta power was higher for favorable than unfavorable outcomes (driven by a cluster around 325–1000 ms, *p* = .002). When using PE terms as regressor for midfrontal EEG power while controlling for PE valence, delta power did not encode PE_{STD} positively, though not significant (p = .056), and also the positive encoding of PE_{DIF} was non-significant (*p* = .053). The positive correlation of beta power with PE_{STD} was not significant anymore (*p* = .059), while the negative correlation with PE_{DIF} remained (p = .001, 450–950 ms). When adding PE_{STD} and PE_{DIF} together to achieve PE_{BIAS}, theta/delta power indeed significantly encoded PE_{BIAS}, first positively (p = .032, 224–475 ms) and then negatively (p = .019, 600 – 1,000 ms; around 8 Hz and thus rather in the alpha band). Also, beta power was significantly negatively correlated with PE_{BIAS} (p = .008, 450 – 975 ms).

In sum, all findings reported in the main text also held when analyzing only this subgroup of only 29 participants. In addition, also late beta power and theta/alpha power appeared to negatively encode the PE_{BIAS} term.

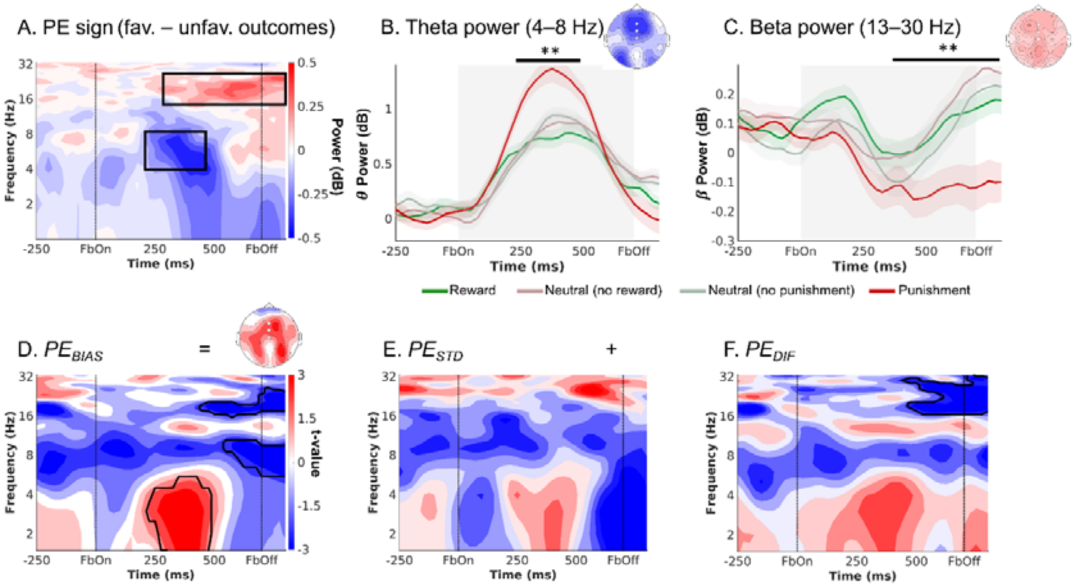

Figure S01C. EEG time-frequency power midfrontal electrodes (Fz/ FCz/ Cz) reflecting outcomes processing in the subgroup of 29 participants included in the fMRI-inspired EEG analyses. (A) Time-frequency plot (logarithmic y-axis) displaying high theta (4–8 Hz) power for unfavorable outcomes and higher beta power (16–32 Hz) for favorable outcomes. (B). Theta power transiently increases for any outcome, but more so for unfavorable outcomes (especially punishments) around 225–475 ms after feedback onset. (C) Beta is higher for favorable than unfavorable outcomes (especially punishments) over a long time period around 300–1,250 ms after feedback onset. (D-F). Correlations between midfrontal EEG power and trial-by-trial PEs. Solid black lines indicate clusters above threshold. Biased PEs were significantly positively correlated with midfrontal theta power, but also negatively correlated with later alpha and beta power (D). The correlations of theta with the standard PEs (E) and the difference term to biased PEs (F) were also positive, though not significant. Beta power only encoded the difference term to biased PEs (F). ** p < 0.01.** p < 0.01.

Regarding fMRI correlates of the past action, similar to the original analysis comprising 34 participants, there were no clusters with higher BOLD after Go than NoGo actions at the time of outcomes, but vice versa, large parts of cortex and subcortex showed higher BOLD after NoGo than Go actions, highly similar to the original analysis ($z_{\max}$ = 7.65, p = 0, 124629 voxels, xyz = [-58 18 22]).

Furthermore, there were four clusters with higher BOLD for Go than NoGo actions at the time of the response, namely one large cluster across lateral prefrontal cortex, anterior cingulate cortex, striatum, thalamus, angular gyrus, cerebellum, left operculum and motor cortex, intracalcarine cortex, and occipital pole ($z_{\max}$ = 7.45, p = 0, 61057 voxels, xyz = [32 -4 -4]), one in right middle temporal gyrus ($z_{\max}$ = 4.90, p = 8.66e-05, 493 voxels, xyz = [66 -32 -12]), one in left inferior temporal gyrus ($z_{\max}$ = 4.43, p = .00294, 293 voxels, xyz = [-60 -44 -18]), and one in precuneous ($z_{\max}$ = 2.39, p = .0041, 276 voxels, xyz = [-8 -70 38]). All these regions were also found in the original analysis comprising 34 participants. Vice versa, BOLD signal was higher NoGo than Go actions at the time of the response in two clusters in vmPFC and subcallosal cortex ($z_{\max}$ = 4.23, p = .00864, 239 voxels, xyz

= [-2 18 -6]) and right anterior temporal gyrus/ temporal pole (z_{max} = .4.14, *p* = .0193, 201 voxels, xyz = [48 -6 -8]), identical to the original analysis comprising 34 participants.

Finally, there was higher BOLD signal for left hand compared to right hand responses at the time of response in two clusters in right precentral and postcentral gyrus, superior parietal lobule, and operculum (z_{max} = 6.66, *p* = 0, 11597 voxels, xyz = [46 -24 64]) and left cerebellum (z_{max} = 6.76, *p* = 1.05e-18, 2672 voxels, xyz = [-18 -54 -16]), identical to the original analysis comprising 34 participants. Vice versa, there was higher BOLD signal for right hand than left hand responses at the time of responses in five clusters in left precentral and postcentral gyrus, superior parietal lobule, operculum, and thalamus (z_{max} = 6.4, *p* = 0, 12372 voxels, xyz = [-36 -20 66]), right cerebellum (z_{max} = 7.17, *p* = 3.41e-21, 3206 voxels, xyz = [20 -54 -20]), right superior lateral occipital cortex (z_{max} = 4.84, *p* = 2.28e-09, 988 voxels, xyz = [48 -86 -4]), right angular gyrus (z_{max} = 4.11, *p* = 7.68e-05, 396 voxels, xyz = [66 -50 28]), and left superior lateral occipital cortex (z_{max} = 5.03, *p* = .019, 164 voxels, xyz = [-18 -82 48]). The clusters in right occipital pole/ intracalcarine cortex and in right posterior cerebellum observed in the original analysis comprising 34 participants were not observed in this analysis. In sum, all major findings also held when analyzing only this subgroup of only 29 participants.

Regarding EEG time-frequency correlates of the past action, when testing for differences in broadband after outcome onset, there was no significant difference after Go and NoGo responses, *p* = .283. When restricting analyses to the low alpha range, the permutation test was marginally significant, *p* = .056, driven by a cluster around 0–100 ms around 7–10 Hz). When repeating the permutation test for the broadband signal including the last second before outcome onset, there was a significant difference after Go and NoGo responses, driven by clusters in the beta band. *p* = 0.002, -1000 – -275 ms, 13–32 Hz, and in the theta/ low alpha band, *p* = 0.020, -1000 – -525 ms, 4–10 Hz.

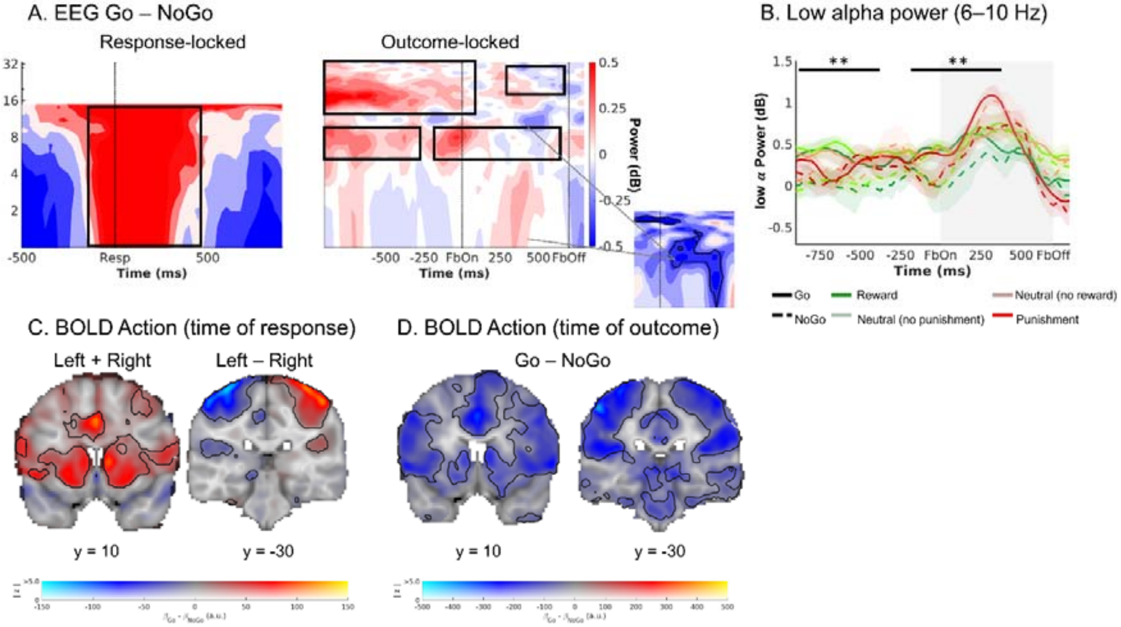

Figure S01D. Exploratory follow-up analyses on ACC BOLD signal and midfrontal low-alpha power in the subgroup of 29 participants included in the fMRI-inspired EEG analyses. (A) Midfrontal time-frequency response-locked (left panel) and outcome-locked (right panel). Before and shortly after outcome onset, power in the lower alpha band is higher on trials with Go actions than on trials with NoGo actions. The shape of this difference resembles the shape of ACC BOLD-EEG TF correlations (small plot; note that this plot depicts BOLD-EEG correlations, which are negative). Note that differences between Go and NoGo trials occurred already before outcome onset in the alpha and beta range, reminiscent of delay activity; but were not fully sustained since the actual response. (B) Midfrontal power in the lower alpha band per action x outcome condition. Lower alpha band power is consistently higher on trials with Go actions than on trials with NoGo

actions, starting already before outcome onset. (C) BOLD signal differences between Go and NoGo actions (left panel) and left vs. right hand responses (right panel) at the time or responses. Response-locked ACC BOLD is significantly higher for Go than NoGo actions. (D) BOLD signal differences between Go and NoGo actions at the time of outcomes. Outcome-locked ACC BOLD (and BOLD in other parts of cortex) is significantly lower on trials with Go than on trials with NoGo actions.

When linking trial-by-trial BOLD signal in selected ROIs as well as midfrontal EEG TF power to response repetition on the next trial with the same cue, ACC BOLD signal did not significantly predict the response repetition, *b* = -0.013, *SE* = 0.018, $\chi^2(1) = 0.524$, *p* = .469, and neither did PCC BOLD signal, *b* = -0.037, *SE* = 0.018, $\chi^2(1) = 2.079$, *p* = .149. However, participants in this subgroup were significantly more likely to repeat the sample action when striatal BOLD signal was high, *b* = 0.097, *SE* = 0.025, $\chi^2(1) = 12.043$, *p* < .001, but more likely to switch when vmPFC BOLD was high, *b* = -0.075, *SE* = 0.019, $\chi^2(1) = 13.170$, *p* < .001.

When linking trial-by-trial midfrontal EEG TF power to response repetition on the next trial with the same cue, participants in this subgroup were more likely to repeat the same response when beta power was high, *b* = 0.124, *SE* = 0.036, $\chi^2(1) = 3.502$, *p* < .001, or when low alpha power was high, *b* = 0.135, *SE* = 0.044, $\chi^2(1) = 8.789$, *p* = .003, but more likely to switch to another response when theta power was high, *b* = -0.090, *SE* = 0.040, $\chi^2(1) = 4.812$, *p* = .028.

S02: Stay behavior as a function of action, salience, and valence

Effect	χ^2	Df	*p*-value
Action	0.01	1	.924
Salience	5.15	1	.021
Valence	45.59	1	< .001
Action x Salience	0.12	1	.728
Action x Valence	3.24	1	.067
Salience x Valence	30.95	1	< .001
Action x Valence x Salience	19.73	1	< .001
Salient outcomes only:			
Action	0.01	1	.960
Valence	46.36	1	< .001
Action x Valence	17.80	1	< .001
Neutral outcomes only:			
Action	.102	1	.750
Valence	.830	1	.362
Action x Valence	12.32	1	< .001
Go with salient outcomes only:			
Valence	53.93	1	< .001
NoGo with salient outcomes only:			
Valence	18.23	1	< .001
Go with neutral outcomes only:			
Valence	0.13	1	.050
NoGo with neutral outcomes only:			
Valence	7.21	1	.007

Table S02. Full report of model of stay behavior. Mixed-effects logistic regression of stay vs. switch behavior (i.e., repeating vs. changing an action on the next occurrence of the same cue) as a function of performed action (Go vs. NoGo), outcome salience (salient: reward or punishment vs. neutral: no reward or no punishment), and outcome valence (positive: reward or no punishment vs. negative: no reward or punishment). Follow-up analyses are performed on trials with salient vs. neutral outcomes separately, and then separately based on Go vs. NoGo actions and salient vs. neutral outcomes. *P*-values are computed using likelihood ratio tests using the *mixed*-function (option "LRT") from package *afex*.

S03: Model parameters and fit indices for models M1-M6

	M1	M2	M3	M4	M5 (Asymmetric pathways)	M6 (Action priming)
Mean log model evidence	-609.30	-597.95	-554.46	-532.40	-528.13	-540.84
Model frequency	0	0.0278	0	0.0488	0.6815	0.2419
Protected exceedance probability	0	0	0	0	.9970	.0030
ρ	7.75	6.81	6.38	10.05	9.41	6.64
	[0.53 – 38.68]	[0.48 – 37.74]	[0.49 – 35.71]	[1.26 – 40.60]	[0.98 – 31.22]	[0.71 – 22.83]
ε_0	0.17	0.20	0.21	0.09	0.08	0.039
	[0.002 – 0.77]	[0.003 – 0.82]	[0.003 – 0.85]	[0.003 – 0.38]	[0.003 – 0.41]	[0.003 – 0.11]
b		-0.05	-0.01	0.13	0.14	0.16
		[-1.23 – 0.82]	[-1.23 – 1.09]	[-1.16 – 1.03]	[-1.18 – 1.10]	[-1.22 – 1.40]
π			0.77		0.17	-1.11
			[-0.78 – 3.73]		[-1.25 – 2.70]	[-3.29 – 1.23]
$\varepsilon_{\text{rewarded Go}}$ $(\varepsilon_0+\kappa)$				0.749	0.833	
				[0.29 – 0.99]	[0.43 – 0.99]	
$\varepsilon_{\text{punished NoGo}}$ $(\varepsilon_0-\kappa)$				0.001	0.003	
				[0.001 – 0.02]	[0.001 – 0.09]	
$\varepsilon_{\text{salient Go}}$						0.49
						[0.05 – 0.90]

Table S03. Model parameters for fitted models. Mean [minimum – maximum] of participant-level parameter estimates in model space, fitted with hierarchical Bayesian inference (only the respective model included in the fitting process). Model frequency and protected exceedance probability are based on a model comparison that involves models M1-M6. Note that Fig. 2 in the main text does not include M6.

S04: Simulations for asymmetric pathways and action priming model

Motivational learning biases are predicted by the *asymmetric pathways model* (*15*, *16*): Positive PEs, elicited by rewards, lead to long-term potentiation in the striatal direct "Go" pathway (and long term depression in the indirect pathway), allowing for a particularly effective acquisition of Go actions to obtain rewards. Conversely, negative PEs, elicited by punishments, lead to long term potentiation in the NoGo pathway, impairing the unlearning of NoGo actions in face of punishments.

An alternative account has recently suggested that self-generated (Go) actions lead to preferential learning (relative to non-self-generated actions, including inaction), more generally (henceforth called "action priming model")(*33*). A self-generated action could "prime" basal ganglia circuits and lead to subsequently larger PEs and thus faster learning. The main differential prediction between these two models is how they account for the failure to learn "Go" actions to avoid punishment: In the first model, this is due to a failure to unlearn punished "NoGo" actions, while in the second model, this is due increased unlearning of punished "Go" actions.

Here, we directly tested both models against each other. As an alternative model M6 (Cockburn et al. 2014), we specified a model with two separate learning rates, one learning rate for trials where self-generated (Go) action selection should prime the processing of any following salient outcome (i.e., Go actions followed by rewards/ punishments), and one learning rate for any other action-outcome combination. In this model, equation (6) is substituted by equation (7):

$$\varepsilon = \begin{cases} \varepsilon_{salGo} & for\ any\ Go\ action\ with\ salient\ outcomes \\ \varepsilon_0 & else \end{cases} \qquad (7)$$

When comparing all models M1–M6 using Bayesian model selection, M5 (the asymmetric pathways model) received highest support (model frequency: 68.15%; protected exceedance probability: 99.70%), also compared to M6 (the action priming model; model frequency: 24.19%; protected exceedance probability: 0.30%). In fact, as visible in Fig. S04E-H, the action priming did not reproduce the motivational biases in learning curves and bar plots, which constitutes a case of qualitative model falsification (*79*, *80*). If anything, it seems that the action priming model trades off both biases, leading to negative response biases for a majority of participants. In contrast, the asymmetric pathways model (M5) was well able to capture the qualitative patterns observed in the data (Fig. S04A-D). We conclude that only the asymmetric pathways model is able to qualitatively reproduce core characteristics of our data.

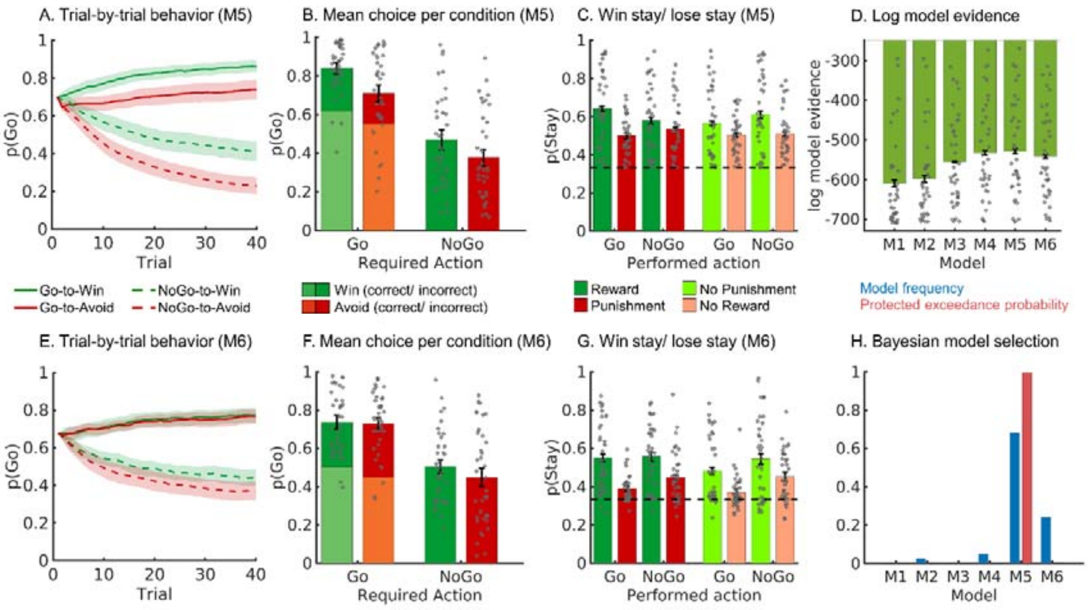

Figure S04. Model comparison and validation of asymmetric pathways (M5) and action priming (M6) model. (A-C) One-step-ahead predictions using parameters (hierarchical Bayesian inference) of the winning model asymmetric pathways model (M5). (A) Trial-by-trial proportion of Go responses (±SEM across participants) for Go cues (solid lines) and NoGo cues (dashed lines); (B) Mean (±SEM across participants) proportion Go responses per cue condition (points are individual participants' means); (C) Probability to repeat a response ("stay") on the next encounter of the same cue as a function of action and outcome. The asymmetric pathways model is well able to capture core characteristics of the empirical data (see Fig. 2 in the main text). (D) Log-model evidence favors the asymmetric pathways model (M5), even over the action priming model (M6). (E-G) Trial-by-trial proportion of Go responses, mean proportion Go responses, and probability of for the action priming model (M6). This model does not reproduce motivational biases (i.e., the difference between green and red lines and bars) well. (H) Model frequency and protected exceedance probability indicate best fit for model M5 (asymmetric pathways model), in line with log model evidence.

S05: Anatomical masks and conjunctions of anatomical and functional masks

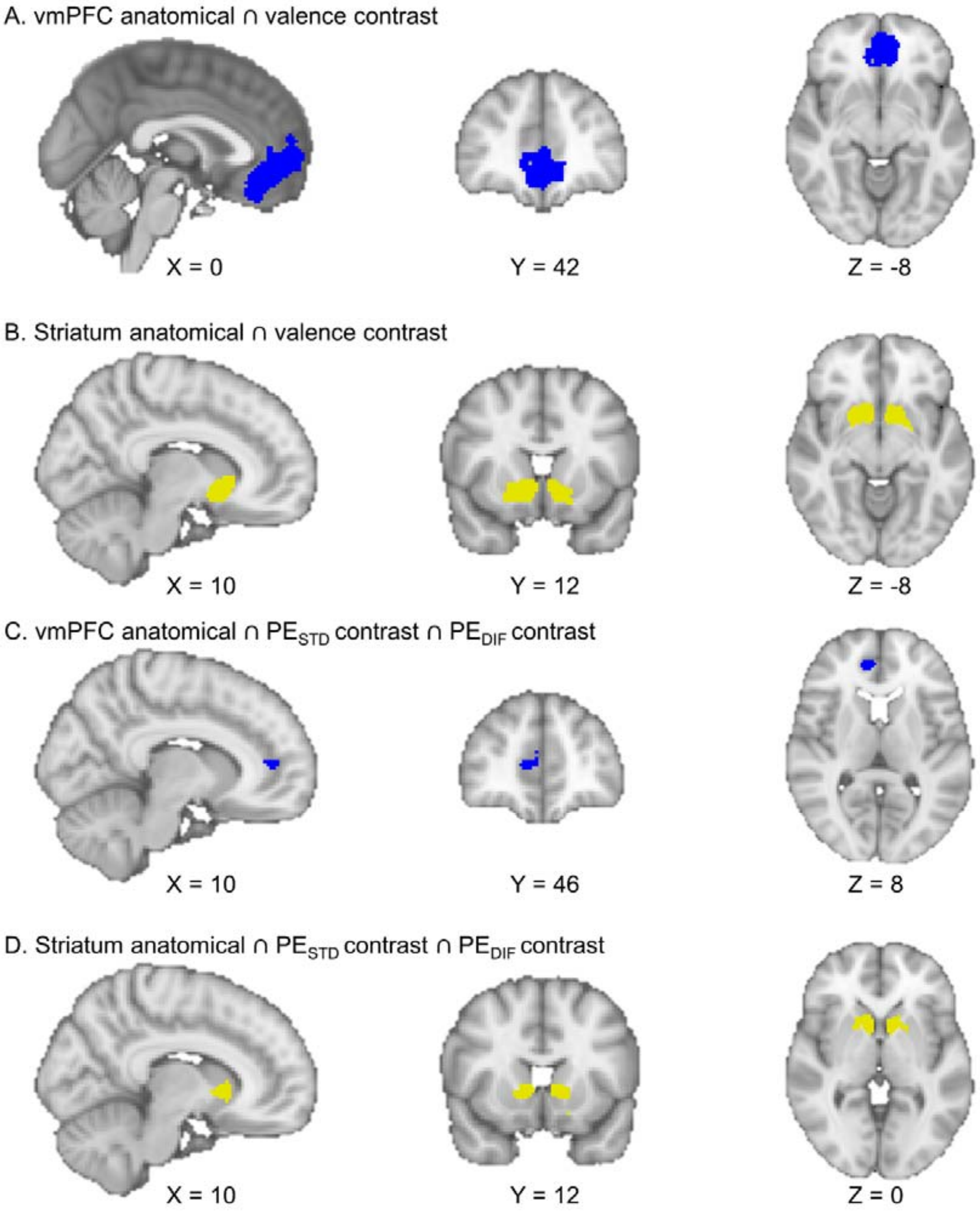
A. vmPFC anatomical ∩ valence contrast
X = 0
Y = 42
Z = -8
B. Striatum anatomical ∩ valence contrast
X = 10
Y = 12
Z = -8
C. vmPFC anatomical ∩ PE$_{STD}$ contrast ∩ PE$_{DIF}$ contrast
X = 10
Y = 46
Z = 8
D. Striatum anatomical ∩ PE$_{STD}$ contrast ∩ PE$_{DIF}$ contrast
X = 10
Y = 12
Z = 0

Figure S05A. Conjunctions of anatomical masks with functional contrasts from fMRI GLM analyses used for fMRI-informed EEG analyses. Anatomical masks are based on the Harvard-Oxford Atlas. Functional contrasts involve outcome Valence and conjunction of PE$_{STD}$ and PE$_{DIF}$. (A) Anatomical AAC contrast (pink, cingulate gyrus, anterior division); (B) vmPFC outcome valence contrast (dark blue, conjunction of frontal pole, frontal medial cortex, and paracingulate gyrus); (C) striatum outcome valence contrast (yellow, conjunction of bilateral nucleus accumbens, caudate, and putamen); (D) vmPFC PE$_{STD}$ ∩ PE$_{DIF}$ contrast (dark blue); and (E) and striatum PE$_{STD}$ ∩ PE$_{DIF}$ contrast (yellow). All anatomical masks were extracted from the probabilistic Harvard-Oxford Atlas, thresholded at 10%. Note that images are in radiological orientation (i.e., left brain hemisphere presented on the right and vice versa).

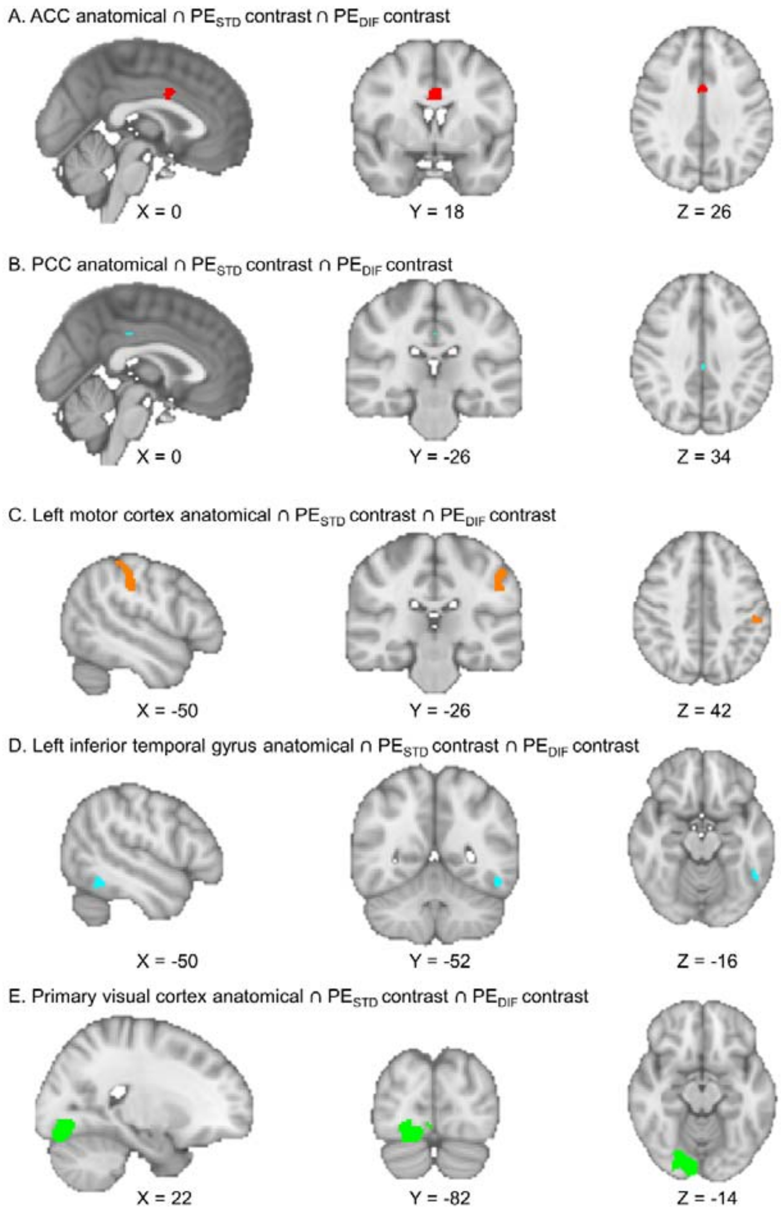
A. ACC anatomical ∩ PE_{STD} contrast ∩ PE_{DIF} contrast
X = 0
Y = 18
Z = 26
B. PCC anatomical ∩ PE_{STD} contrast ∩ PE_{DIF} contrast
X = 0
Y = -26
Z = 34
C. Left motor cortex anatomical ∩ PE_{STD} contrast ∩ PE_{DIF} contrast
X = -50
Y = -26
Z = 42
D. Left inferior temporal gyrus anatomical ∩ PE_{STD} contrast ∩ PE_{DIF} contrast
X = -50
Y = -52
Z = -16
E. Primary visual cortex anatomical ∩ PE_{STD} contrast ∩ PE_{DIF} contrast
X = 22
Y = -82
Z = -14

Figure S05B. Conjunctions of anatomical masks with functional contrasts from fMRI GLM analyses used for fMRI-informed EEG analyses: (A) AAC PE_{STD} ∩ PE_{DIF} contrast (red, cingulate gyrus, anterior division); (B) PCC PE_{STD} ∩ PE_{DIF} contrast (light blue, cingulate gyrus, posterior division); (C) left motor cortex PE_{STD} ∩ PE_{DIF} contrast (orange, conjunction of precentral and postcentral gyrus); (D) Left inferior temporal gyrus PE_{STD} ∩ PE_{DIF} contrast (turquoise, conjunction of inferior temporal gyrus, posterior division, and inferior temporal gyrus, temporooccipital part); and (E) primary visual cortex PE_{STD} ∩ PE_{DIF} contrast (green, conjunction of lingual gyrus, occipital fusiform gyrus, occipital pole). All anatomical masks were extracted from the probabilistic Harvard-Oxford Atlas, thresholded at 10%. Note that images are in radiological orientation (i.e., left brain hemisphere presented on the right and vice versa).

S06: Regressors and contrast in fMRI analyses

Model-based GLM with PE_{STD} and PE_{DIF} regressor:

- WinGoOnset: for every trial with Win cue and Go action, at cue onset, duration 1, value +1
- AvoidGoOnset: for every trial with Avoid cue and Go action, at cue onset, duration 1, value +1
- WinNoGoOnset: for every trial with Win cue and NoGo action, at cue onset, duration 1, value +1
- AvoidNoGoOnset: for every trial with Avoid cue and NoGo action, at cue onset, duration 1, value +1
- Handedness: for every trial, at cue onset, duration 1, value +1 for left hand response, 0 for NoGo 10 response, -1 for right hand response 11
- Error: for every trial, at cue onset, duration 1, value +1 for incorrect response, 0 for correct response
- OutcomeOnset: for every trial, at outcome onset, duration 1, value +1 for every trial
- PE_{STD}: for every trial, at outcome onset, duration 1, value is demeaned PE times learning rate for model M1
- PE_{DIF}: for every trial, at outcome onset, duration 1, value is demeaned difference between (PE times learning rate) for model M1 and (PE times learning rate) for model M5
- Invalid: for trials where uninstructed button was pressed, at outcome onset, duration 1, value 1

	Regressor	1	2	3	4	5	6	7	8	9	10
	Contrast	WinGoOnset	AvoidGoOnset	WinNoGoOnset	AvoidNoGoOnset	Handedness	Error	Outcome Onset	PE_{STD}	PE_{DIF}	Invalid
1	PE_{STD}								1		
2	PE_{DIF}									1	

Model-free GLM using response-locked and outcome-locked response regressors:

- GoReward: for every trial with Go action and reward obtained, at outcome onset, duration 1, value +1
- GoNoReward: for every trial with Go action and no reward obtained, at outcome onset, duration 1, value +1
- GoNoPunishment: for every trial with Go action and no punishment obtained, at outcome onset, duration 1, value +1
- GoPunishment: for every trial with Go action and punishment obtained, at outcome onset, duration 1, value +1
- NoGoReward: for every trial with NoGo action and reward obtained, at outcome onset, duration 1, value +1
- NoGoNoReward: for every trial with NoGo action and no reward obtained, at outcome onset, duration 1, value +1
- NoGoNoPunishment: for every trial with NoGo action and no punishment obtained, at outcome onset, duration 1, value +1
- NoGoPunishment: for every trial with NoGo action and punishment obtained, at outcome onset, duration 1, value +1
- LeftHand: for very trial with left hand response, at response onset, duration 1, value + 1
- RightHand: for very trial with right hand response, at response onset, duration 1, value + 1
- Error: for every trial, at cue onset, duration 1, value +1 for incorrect response, 0 for correct response
- OutcomeOnset: for every trial, at outcome onset, duration 1, value +1 for every trial
- Invalid: for trials where uninstructed button was pressed, at outcome onset, duration 1, value 1

	Regressors	1	2	3	4	5	6	7	8	9	10	11	12	13
	Contrast	GoReward	GoNoReward	GoNoPunishment	GoPunishment	NoGoReward	NoGoNoReward	NoGoNoPunishment	NoGoPunishment	LeftHand	RightHand	Error	OutcomeOnset	Invalid
1	Valence	1	-1	1	-1	1	-1	1	-1					
2	Action	1	1	1	1	-1	-1	-1	-1					
3	Hand Sum									1	1			
4	Hand Dif									1	-1			

S07: Significant clusters in BOLD-GLMs with behavioral regressors only

Model-based GLM with PE_{STD} and PE_{DIF} regressor:

	Contrast				Peak coordinates		
No	Brain region	Maximal Z-value	Cluster size (voxels)	Corrected p	x	y	z
	PE_{STD} Positive						
1	Ventromedial prefrontal cortex, Nucleus accumbens, caudate, putamen, bilateral amygdala, bilateral hippocampus	6.47	8762	1.02e-43	12	14	-6
2	Occipital pole, lingual gyrus, occipital fusiform gyrus	6.64	1012	6.10e-10	10	-92	-10
3	Posterior cingulate cortex	4.72	985	9.40e-10	4	-50	18
4	Left superior frontal gyrus	5.56	910	3.19e-09	-18	34	50
5	Right middle temporal gyrus, anterior division	5.48	381	6.47e-05	62	-4	-18
6	Left inferior temporal gyrus, temporooccipital part	5.16	360	.000103	-52	-46	-10
7	Left middle temporal gyrus, anterior division	4.70	329	.000209	-60	-10	-14
8	Left postcentral gyrus	4.33	271	.000838	-52	-28	48
9	Right cerebellum	4.89	147	.0239	44	-72	-40
10	Anterior cingulate cortex	4.27	146	.0247	2	6	34
	PE_{STD} Negative						
1	Right superior frontal gyrus	5.20	351	.000127	6	26	62
2	Right occipital pole, right inferior lateral occipital cortex	4.76	211	.00391	30	-94	4
3	Left lingual gyrus	4.21	186	.00776	-22	-64	2
4	Left inferior lateral occipital cortex	4.28	147	.0239	-44	-86	-10
	PE_{DIF} Positive						
1	Bilateral superior frontal gyrus, paracingulate gyrus, anterior cingulate cortex, posterior cingulate cortex, ventromedial frontal cortex, bilateral frontal orbital cortex, bilateral frontal pole, bilateral supramarginal gyrus, bilateral middle temporal gyrus, bilateral inferior temporal gyrus, bilateral fusiform gyrus, bilateral inferior occipital cortex, bilateral superior occipital cortex, precuneous, bilateral cerebellum	7.11	35109	0	34	-84	20
2	Right insula, right frontal operculum, right inferior frontal gyrus, right middle frontal gyrus, right frontal orbital cortex, bilateral caudate, bilateral Nucleus accumbens, bilateral thalamus, brainstem	6.36	10364	0	34	20	-8
3	Left insula, left frontal operculum,	6.51	10132	0	-36	20	-6

	left inferior frontal gyrus, left middle frontal gyrus, left frontal orbital cortex						
4	Right middle temporal gyrus, posterior division	4.66	307	.0003	56	-32	-4
5	Right insula, right planum polare	4.72	143	.0248	40	-8	-12
	PE$_{DIF}$ Negative						
1	Left middle temporal gyrus, anterior division	4.22	191	.00607	-64	-6	-14
2	Left hippocampus	4.49	158	.0158	-26	-14	-22

Model-free GLM using response-locked and outcome-locked response regressors:

	Contrast				Peak coordinates		
No	Brain region	Maximal Z-value	Cluster size (voxels)	Corrected p	x	y	z
	Favorable > Unfavorable						
1	Ventromedial prefrontal cortex, left lateral orbitofrontal cortex, Nucleus accumbens, caudate, putamen, bilateral amygdala, bilateral hippocampus	5.65	3999	2.86e-19	8	12	-4
2	Left superior frontal gyrus	4.03	331	0.00239	-18	28	60
3	Left lateral orbitofrontal cortex	4.31	288	0.00512	-34	40	-8
4	Right occipital pole	4.59	213	0.0212	18	-92	-16
	Unfavorable > Favorable						
1	Right lateral orbitofrontal cortex	4.59	367	0.00142	30	62	-2
2	Precuneous	4.58	356	0.00170	8	-66	58
3	Right superior frontal gyrus	4.32	340	0.00223	12	14	72
	Go > NoGo **outcome-locked** *No significant clusters*						
	NoGo > Go **outcome-locked**						
1	Bilateral lateral orbitofrontal cortex, Bilateral superior frontal gyrus, anterior cingulate cortex, posterior cingulate cortex, pre-SMA, bilateral precentral gyrus, bilateral postcentral gyus, bilateral supramarginal gyrus, bilateral operculum, bilateral planum temporale, bilateral superior temporal gyrus, bilateral middle temporal gyrus, bilateral inferior temporal gyrus, bilateral superior lateral occipital cortex, bilateral inferior lateral occipital cortex, bilateral thalamus	7.32	114090	0	-42	-6	12
	Go (left + right hand response) > NoGo **response-locked**						
1	Cerebellum, bilateral thalamus, bilateral putamen, bilateral caudate, bilateral Nucleus Accumbens, posterior cingulate cortex, right operculum, right angular gyrus, right superior parietal lobule. anterior cingulate cortex, paracingulate gyrus, bilateral ventrolateral frontal cortex, right middle frontal gyrus	7.08	46437	0	32	-4	-6
2	Left operculum, left angular gyrus, left superior parietal lobule	5.88	3936	3.13e-17	-46	-24	26
3	Intracalcarine cortex	3.79	374	0.00248	-12	-88	6
4	Right middle temporal gyrus	4.63	287	0.00956	68	-32	-12

	NoGo > Go (left + right hand response) response-locked						
1	Right medial temporal gyrus, right temporal pole	4.09	465	0.000636	50	-8	-16
2	vmPFC, subcallosal cortex	3.95	435	0.000973	0	40	-12
	Left Hand > Right Hand Response response-locked						
1	Right precentral gyrus, right postcentral gyrus, right superior parietal lobule, right operculum	7.05	9460	9.41e-39	46	-24	64
2	Left cerebellum	7.18	2208	2.1e-14	-18	-54	-18
	Right Hand > Left Hand Response response-locked						
1	left precentral gyrus, left postcentral gyrus, left superior parietal lobule, left operculum, left thalamus	7.06	14870	0	-36	-20	66
2	Right anterior cerebellum	7.90	3735	1.44e-20	18	-54	-20
3	Right inferior lateral occipital cortex, right superior lateral occipital cortex	4.96	1452	9.66e-11	48	-86	-4
4	Right angular gyrus	4.98	551	2.06e05	66	-50	28
5	Left occipital pole, right intracalcarine cortex	3.93	409	0.000236	-4	-96	26
6	Right posterior cerebellum	4.64	200	0.0157	48	-78	-32

S08: EEG time-frequency results after ERPs are removed

Given that differences in theta power between favorable and unfavorable outcomes as well as differences in lower alpha band power after Go and NoGo responses occurred quite soon after cue onset, we aimed to test whether these effects reflected differences in evoked rather than induced activity. For this purpose, we removed evoked components from our data by computing the ERP for each of the eight conditions (action x outcome) for each participant and then subtracting the condition-specific ERP from the trial-by-trial data (*81*). Only afterwards, we performed time-frequency decomposition.

In line with the results reported in the main text, power was higher for unfavorable compared to favorable outcomes in the theta band (*p* =.018, driven by cluster at 225–475 ms; Fig. S08B), but higher for favorable than unfavorable outcomes in the beta band (*p* < .001, driven by cluster at 0–1250 ms; Fig. S08C). Notably, unlike the results reported in the main text (Fig. 4A), the cluster of high power for unfavorable compared to favorable outcomes was constrained to the theta range, and did not extend further into the delta range (Fig. S08A).

When using the trial-by-trial PEs (both the standard PE and the difference term to a biased PE) as predictors in a multiple linear regression at each time-frequency-channel bin while controlling for PE valence, delta power encoded PE_{STD} positively, though not significantly (*p* = .198). However, at a later time point around outcome offset, delta (and theta) power in fact correlated negatively with PE_{STD} (575–800 ms, *p* = .002; Fig. S08E). The correlation between delta and the PE_{DIF} term was still positive, but not significant (p = .228, Fig. S08F). Similarly, the correlation of the PE_{BIAS} term with delta power was positive, but not significant (*p* = .084; Fig. S08D).

Regarding beta power, there was a positive, though non-significant correlation of beta power with PE_{STD} (*p* = .096). There was again a significantly negative correlation of beta power with PE_{DIF} (425–875 ms, *p* < .001, Fig. S08B). Likewise, beta power correlated significantly negatively with PE_{BIAS} (450–800 ms, *p* = .018), driven by the correlation with PE_{DIF}.

In sum, after subtracting the condition-wise ERP from each trial before time-frequency decomposition, supposedly removing the phase-locked aspect of power, both beta and theta still encoded PE valence. However, the encoding of PE magnitude by delta power was attenuated and not significant any more.

This reduction in magnitude encoding might occur of several reasons. Firstly, it might be that this correlation in the delta range is in fact (partly) reflecting correlations with phase-locked, i.e., evoked activity (ERPs), especially in the N2 (FPN)/ P3 (RewP) time range (see S09) (*26*, *28*–*30*, *82*–*87*). Nonetheless, a positively correlation between delta power and biased PEs is still visible in Fig. S08D, suggesting that at least part of the signal encoding biased PEs is not phase-locked. Secondly, it might be that the removal of the condition-wise ERPs has introduced additional noise in the data, attenuating any true correlation. Thirdly, there was a negative correlation between PE_{STD} and theta/ delta power at later time points which was visible, though not significant in the results reported in the main text (Fig. 4D). Subtraction of an ERP-like template acts like a high-pass filter. High-pass filtering at relatively high cut-offs (> 0.5 Hz) can artificially postpone or induce effects at later points (*88*). It is possible that in this case, ERP subtraction attenuated a positive correlation in the theta/ delta range, but enhanced a later negative correlation.

Taken together, it is possible that part of the PE magnitude encoding in the theta/ delta range is due to correlations with the phase-locked (ERP) signal. However, this finding does not compromise the conclusion that overall, theta/delta power seemed to be more strongly associated with the PE_{BIAS} term than the PE_{STD} term. Our primary goal is not to pinpoint the precise nature of electrophysiological correlates of biased learning, but rather test the relative temporal order of when different regions exhibiting biased learning signals become active.

Finally, we tested whether after ERP subtraction, low alpha (and beta power) still encoded the previously performed action. When testing for differences in broadband power after Go and NoGo responses, power was indeed significantly different between conditions, driven by clusters in beta band (*p* = 0.002, 0.125 – 625 ms; *p* = 0.052, 700 - 1000 ms, 23 - 29 Hz) and theta/ low alpha band (*p* = 0.024, 575 – 1000 ms, 5–9 Hz; *p* = 0.056, 0 –225 ms, 6–11 Hz). For power before outcome onset, there were again broadband differences between Go and NoGo (*p* = 0.002, -1000 – +225 ms, 1 –33 Hz), but note that there was no ERP subtracted before outcome onset. We thus conclude that the differences between Go and NoGo responses were attributable to differences in induced rather than evoked activity.

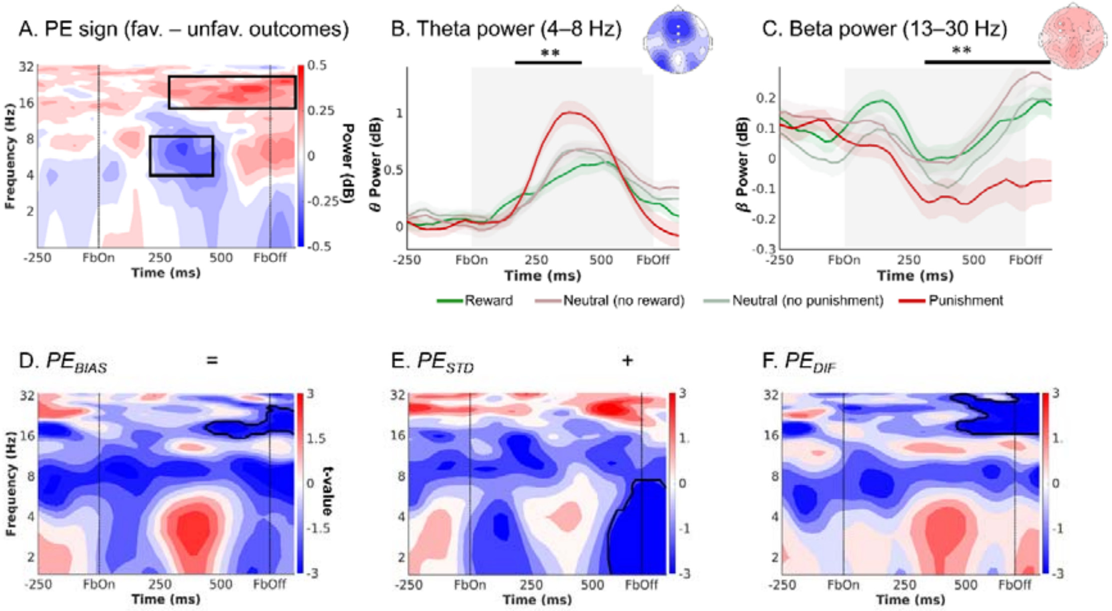

Figure S08. EEG time-frequency power over midfrontal electrodes (Fz/ FCz/ Cz) after the (action x outcome) condition-wise ERPs has been removed. (A) Time-frequency plot (logarithmic y-axis) displaying high theta (4–8 Hz) power for unfavorable outcomes and higher beta power (16–32 Hz) for favorable outcomes. (B). Theta power transiently increases for any outcome, but more so for unfavorable outcomes (especially punishments) around 225–475 ms after feedback onset. (C) Beta is higher for favorable than unfavorable outcomes (especially punishments) over a long time period around 300–1,250 ms after feedback onset. (D-F). Correlations between midfrontal EEG power and trial-by-trial PEs. Solid black lines indicate clusters above threshold. There still was a visible positive correlation between biased PEs and midfrontal delta power, but this correlation was not significant (D). The correlation of delta with the standard PEs (E) was also positive, though not significant; in fact, at a later time point around stimulus offset, delta power correlated significantly negatively with standard PEs. The difference term to biased PEs (F) also correlated positively, though not significantly with delta power. Beta power encoded the difference term and biased PEs themselves (F). ** p < 0.01.

S09: ERPs as a function of action and outcome

In addition to the induced activity in time-frequency power reported in the main text, we also analyzed the data in the time domain to test for differences in evoked activity. These analyses were particularly motivated given that differences in time-frequency power between favorable and unfavorable outcomes (theta/delta range) and after Go and NoGo responses (lower alpha/ theta range) occurred soon after outcome onset, warranting the assumption that differences might also occur in evoked activity. A large range of previous research has reported a modulation of evoked potentials by outcome valence in form of the feedback-reduced negativity (*29*, *64*, *82*–*87*), i.e., a stronger N2 component for negative compared to positive outcomes around ~ 250 post-cue over midfrontal electrodes, recently also characterized as rather constituting a reward positivity (RewP) (*82*). Also, some studies have reported a modulation of the P3 by outcome valence, which has been attributed to outcome magnitude or salience rather than valence (*85*–*87*, *89*).

Similar to the analysis of time frequency power, we sorted trials into the eight conditions spanned by the performed action (Go/ NoGo) and the obtained outcome (reward/ no reward/ no punishment/ punishment), computed the average ERP for each condition per participant, and tested for differences between favorable (reward/ no punishment) and unfavorable (no reward/ punishment) outcomes as well as conditions of relative stronger (rewarded Go and punished Go) vs. relatively weaker learning (rewarded NoGo and punished NoGo). We used cluster-based permutation tests on the average signal over midfrontal electrodes (Fz/ FCz/ Cz) in the time range of 0–700 ms after outcome onset (where evoked potentials visible in condition-averaged plot).

First, midfrontal ERPs were significantly different between favorable and unfavorable outcomes, driven by two separate clusters of differences above threshold (Cluster 1: around 246 – 294 ms, p = .034; Cluster 2: around 344 – 414 ms, p =.004, Fig. S09A panel A, C). The first cluster the classical feedback-related negativity, i.e., a stronger N2 component for unfavorable compared to favorable outcomes. The second cluster reflected weaker P3 component for unfavorable compared to favorable outcomes, similar the reward positivity reported before. In fact, the N3 was rather absent for unfavorable outcomes (Fig. S09B). Both effects were clearly focused on midfrontal electrodes. These findings replicate previous findings of outcome valence modulating N2 (feedback-related negativity) and P3 components, and complement our time-frequency findings of theta and beta power reflecting outcome valence.

Second, when contrasting trials with Go vs. NoGo responses, no significant difference was observed (p = .358; Fig. S09A panel D). Visual inspection of the topoplot yielded that, if anything, differences emerged over right occipital electrodes. If one performed a test over those right occipital electrodes (O2, 04, PO4; Fig. S09A panel F; note that this procedure constitutes double-dipping because the test was informed by first looking at the data), this test would have yielded significant results (p = .016) driven by cluster around 423–466 ms, reflecting a slightly larger P3 after Go than NoGo responses (Fig. S09A panel E). This finding appears to be the strongest (if any) difference in amplitude after outcome onset between Go and NoGo actions. Given that this difference was not hypothesized and occurred far away from our a-priori selected channels of interest, we are careful not to over-interpret those differences.

Third, contrasting trials with favorable and unfavorable at the same right occipital electrodes yielded a significant difference, driven by clusters around 46–103 ms (p = 0.034), 141–255 ms (p = .002), and 519 – 580 ms (p = .034). Most notably, the P1 amplitude was much larger for favorable than unfavorable outcomes (Fig. S09A panel B). However, given that these differences were not hypothesized and occurred far away from our a-priori selected channels of interest, we are careful not to over-interpret those differences.

Taken together, we found a bigger midfrontal N2/ FRN for unfavorable compared to favorable outcomes, and a bigger midfrontal P3/ RewP for favorable compared to unfavorable outcomes, in line with a vast literature of previous findings (*29*, *64*, *82–87*, *89*). Midfrontal voltage did not significantly differ after Go or NoGo responses. If anything, differences after Go and NoGo responses were maximal over right occipital electrodes, with a larger P3 after Go than after NoGo responses. Signal at these channels also differed between favorable and unfavorable outcomes, most notably with a bigger P1 after favorable than unfavorable outcomes. In sum, we replicate classical reward learning ERP effects, which shows that the motivational Go/NoGo learning task taps into reward learning processes reported before, but these processes appeared to be unaffected by the previously performed action.

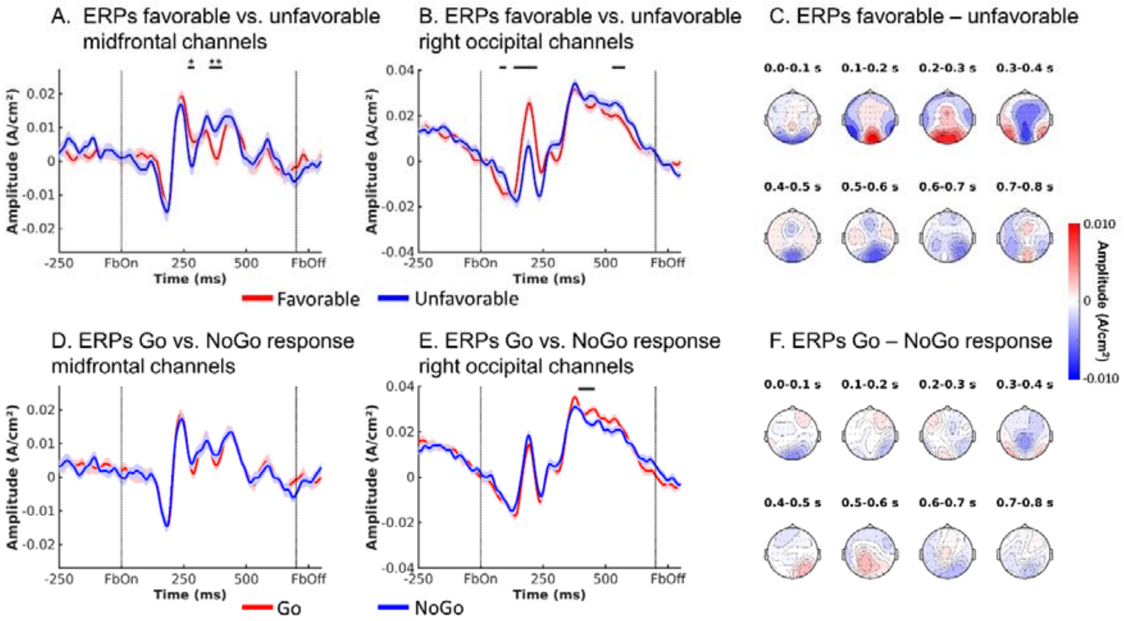

Figure S09A. ERPs reflecting outcome valence and performed action. (A) Voltage (±SEM) over midfrontal electrodes (Fz/FCz/Cz) was lower for unfavorable than favorable outcomes around 246–294 ms (stronger N2, FRN) and higher for favorable than unfavorable outcomes around 344 – 414 ms (stronger P3/ RewP). (B) Over right occipital electrodes, the P3 was slightly bigger for favorable than unfavorable outcomes. ** $p < 0.01$. * $p < .05$ (C) Topoplots of difference in voltage between trials with favorable and unfavorable outcomes over selected time windows. (D) There was no difference in voltage over midfrontal electrodes between trials with Go and NoGo responses. (E) Over right occipital electrodes, the P3 was slightly stronger after Go than NoGo actions (no *p*-value because ROI selected based on visual inspection). (F) Topoplots of difference in voltage between trials with Go and NoGo actions over selected time windows.

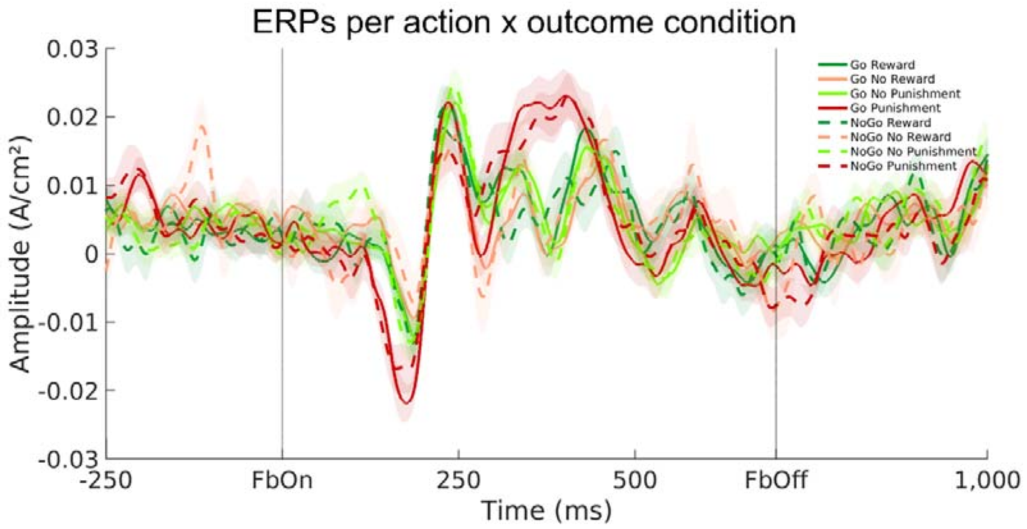

Figure S09B. ERPs per action x outcome condition. Biggest differences occurred around the time of the N2 (FRN) and P3 (RewP). N2 and P3 exhibited larger amplitudes on trials with punishments. There was no apparent modulation by the previous action (Go/ NoGo).

S10: Model-based EEG analyses in the time domain

In addition to testing whether midfrontal time-frequency power reflected signatures of biased learning (see main text), we also tested whether the midfrontal time domain signal reflected biased learning. Again, we used the standard PE term and the difference term to biased PEs as regressors in a multiple linear regression on each channel-time bin.

Focusing on midfrontal electrodes, and controlling for outcomes valence, first, the PE_{STD} term was negatively correlated with midfrontal voltage around 529–575 ms (p = .039; Fig. S10B). Note that so late after outcome onset, signal was not part of any "classical" ERP component any more. Second, the PE_{DIF} correlated negatively with midfrontal voltage around 123–166 ms (*p* = .029) in the time range of the N1 and later positively around 365–443 ms (*p* < .001; Fig S10C) in the time range of the P3/ RewP. Third, a similar pattern of correlations occurred for the PE_{BIAS} term (Cluster 1: negative, 111–184 ms, *p* = .004; Cluster 2: positive, 346–449 ms, *p* < .001; Fig. S10A). Fourth, around these same time windows, midfrontal voltage also encoded outcome valence itself, but with opposite sign (Cluster 1: positive, 99–184 ms, *p* < .001; Cluster 2: negative, 308–448 ms, *p* < .001; see S09).

In sum, similar to analyses of midfrontal power reported in the main text, PE sign and magnitude were encoded in midfrontal voltage around the same time, but with opposite polarity: Signal around the time of the N1 encoded PE sign positively, but PE magnitude negatively. Vice versa, signal around the time of the P3/ RewP encoded PE sign negatively, but PE magnitude positively. The same phenomenon of separate valence and magnitude encoding in midfrontal EEG signal has been reported before (*28*–*30*). Notably, magnitude encoding in midfrontal voltage emerged for the PE_{BIAS} term, but not the PE_{STD}, indicating that this correlation was driven by the PE_{DIF} term and that biased learning described midfrontal voltage better than standard learning. These results complement our findings of theta/delta power encoding outcome valence and magnitude with opposite polarities (see main text).

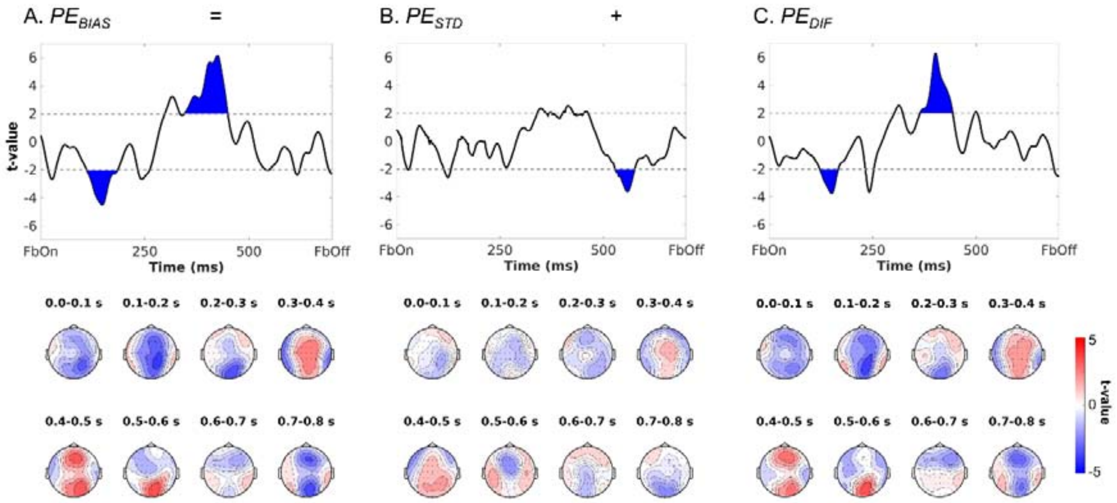

Figure S10. Modulation of EEG voltage by biased PEs and decomposition into the standard PE term and the difference term to biased PEs. (A) Mean EEG voltage over midfrontal electrodes (Fz, FCz, Cz) was significantly modulated by biased PEs around 111–184 (negatively) and 353–414 ms (positively) after outcome onset. (B) Correlations with the standard PE term only emerged around 529 – 575 ms (negatively). (C) Correlations with the difference term to biased PEs were similar to correlations for the biased PE term itself, i.e., around 123–166 (negatively) and 365–443 ms (positively).
Bottom row: Topoplots displaying *t*-values of beta-weights for the respective regressor over the entire scalp in steps of 100 ms from 0 to 800 ms.

S11: Supplementary fMRI-inspired EEG results in time-frequency space

Besides the results for striatum, ACC, and PCC reported in the main text, there were also significant EEG correlates over midfrontal electrodes for trial-by-trial BOLD signal from left motor cortex (*p* = .002, around 0–625 ms, 16–27 Hz; Fig. S11A). There were however no significant EEG correlates over midfrontal electrodes for BOLD signal from vmPFC/ subgenual ACC (*p* = .174; Fig. S11B), left inferior temporal gyrus (*p* = .097; Fig. S11C), and primary visual cortex (*p* = .017; Fig. S11D).

As quality checks, we checked whether visual cortex BOLD correlated negatively with alpha over occipital electrodes (*90*, *91*) and whether motor cortex BOLD correlated negatively with beta power over central electrodes (*92*, *93*). Both was the case (see Fig. S11E and F), showing that our data was of sufficient quality to detect these well-established associations.

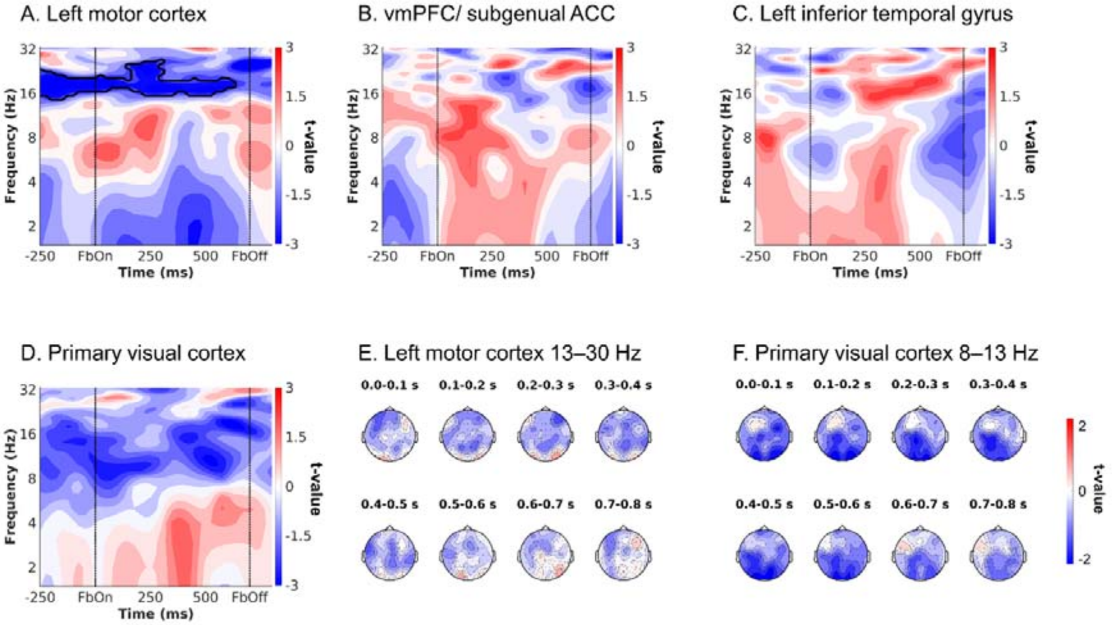

Figure S11. Supplementary fMRI-informed EEG results in the time-frequency domain. Unique temporal contributions of BOLD signal in (A) left motor cortex, (B) vmPFC, (C) left ITG and (D) primary visual cortex to midfrontal EEG power. Group-level *t*-maps display the modulation of the EEG power over midfrontal electrodes (Fz/ FCz/ Cz) by trial-by-trial BOLD signal in the selected ROIs. There significant correlations between midfrontal EEG TF power in the beta range and left motor cortex BOLD signal (*p* = .002), but no significant midfrontal EEG correlates for BOLD signal from other ROIs. (E) Topoplot displaying *t*-values of left motor cortex BOLD over the entire scalp between 13 and 30 Hz (beta band) in steps of 100 ms from 0 to 800 ms. There are significant negatively correlates over central electrodes, especially round 300–500 ms. (F) Topoplot displaying *t*-values of primary visual cortex BOLD over the entire scalp between 8 and 13 Hz (alpha band) in steps of 100 ms from 0 to 800 ms. There are significant negatively correlates over occipital electrodes throughout outcome presentation.

S12: Supplementary fMRI-inspired EEG results in the time domain

For fMRI-inspired analysis of the EEG signal in the time domain (voltage), we applied the same approach as reported in main text, but with voltage signal (time-domain) instead of time-frequency power as dependent variable. As independent variables, we entered the trial-by-trial BOLD signal from all seven regions encoding biased PEs plus the trial-by-trial standard PE and the different term towards the biased PE (exact same procedure as for EEG TF analyses), all in one single multiple linear regression. On a group-level, we again focused on the mean signal over midfrontal electrodes (Fz/ FCz/ Cz) in a time range of 0–700 ms, for which ERPs had been visible in the condition-averaged plots (see S09).

First, trial-by-trial striatal BOLD correlated significantly with midfrontal voltage at two time points, namely positively around 152–196 ms (*p* = .017) in the time range of the N1 and again negatively around 316–383 ms (*p* < .001, see Fig. S12A) in the time range of the N2/ FRN and P3/RewP. Second, trial-by-trial vmPFC BOLD correlated significantly positively with midfrontal voltage around 347–412 ms (*p* = .006, see Fig. S12A) in the time range of the N2/ FRN and P3/RewP. Third, trial-by-trial BOLD from primary visual cortex correlated significantly positively with midfrontal voltage around 307–367 ms (*p* = .011, see Fig. S12B), overlapping with (but slightly earlier than) correlations from vmPFC BOLD, i.e., in the time range of the N2/ FRN and P3/RewP. For midfrontal voltage split up per high vs. low BOLD signal (revealing which ERP components are respectively modulated), see Fig. S12C-E. There were no significantly correlations between midfrontal voltage and trial-by-trial BOLD from ACC (*p* = .927, see Fig. S12A), left motor cortex (*p* = .649, see Fig. S12B), PCC (*p* = .796, see Fig. S12A), or left inferior temporal gyrus (*p* = .649, see Fig. S12B). For further details on BOLD-EEG voltage correlations in the time domain, see Fig. S12F–L.

Taken together, trial-by-trial BOLD signal in striatum, vmPFC, and V1 all correlated with FRN/ RewP amplitude, which is the dominant phenomenon over midfrontal electrodes reflecting outcome valence (see S09 and S10). Notably, correlations with striatal and vmPFC BOLD were of opposite signs, which aligns with the finding that striatal and vmPFC BOLD predicted opposite behavioral tendencies on future trials (see main text; see S15). However, crucially, the time domain signal did not allow for a temporal dissociation of these different regions. Possibly, the midfrontal evoked signal (i.e., the part of the signal that is phase-locked to outcome onset) is so stereotyped that only the FRN/RewP complex shows enough variation across trials to allow for substantial correlations with trial-by-trial BOLD signal. This finding demonstrates that the time-frequency domain signal (i.e., the part of the signal that is not necessarily phase-locked to outcome onset) might be more suited for dissociating the activity of different regions in time.

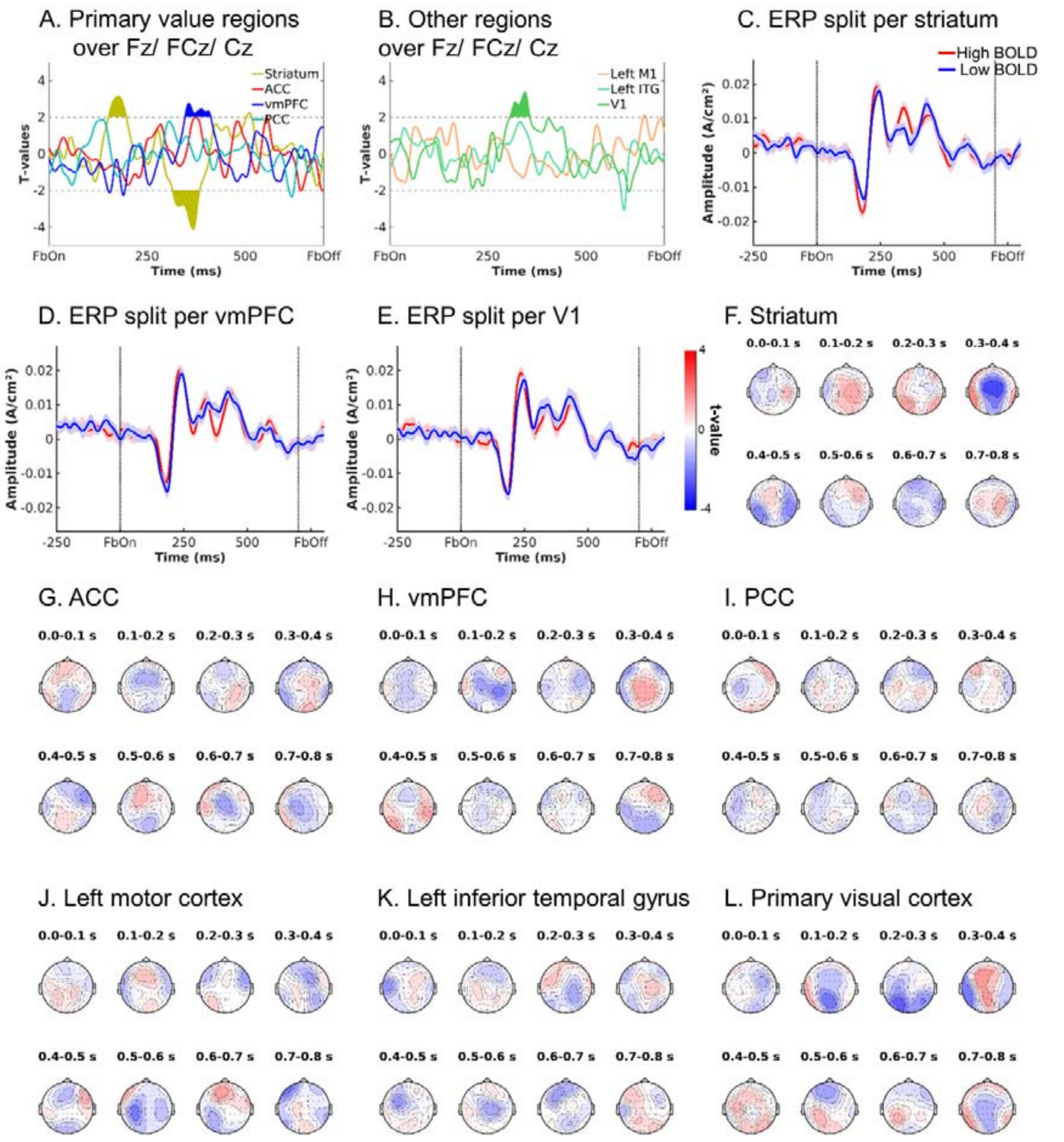

Figure S12. fMRI-informed EEG analyses in the time-domain. Group-level *t*-value time courses display the modulation of the EEG voltage over midfrontal electrodes (Fz/ FCz/ Cz) by trial-by-trial BOLD signal in the selected ROIs. (A) Correlations between midfrontal voltage and trial-by-trial BOLD signal from core value regions, i.e., striatum, ACC, vmPFC, and PCC. Striatal BOLD modulates the amplitude of the N1 and P3, while the P3 amplitude is also modulated by vmPFC BOLD. (B) Correlations between midfrontal voltage and trial-by-trial BOLD signal from other regions, i.e., left motor cortex, left inferior temporal gyrus, and primary visual cortex. Visual cortex BOLD modulates the amplitude of the P3, as well. (E-F) Midfrontal voltage split up for high vs. low BOLD signal (median split) from regions significantly modulating voltage. Striatal BOLD modulates N1 and P2 amplitude, while vmPFC BOLD and visual cortex BOLD modulate N2 (FRN) amplitude. (G-L) Topoplots displaying *t*-values of correlations between midfrontal voltage and trial-by-trial BOLD for all regions in steps of 100 ms from 0 to 800 ms.

S13: Full list of significant clusters with EEG regressors in fMRI GLMs

	Contrast				Peak coordinates		
No	Brain region	Maximal Z-value	Cluster size (voxels)	Corrected p	x	y	z
	Central Lower Alpha Band Positive						
	No significant clusters						
	Central Lower Alpha Band Negative						
1	Precuneous, cuneal cortex, right superior lateral occipital cortex	5.78	8346	2.50e-33	6	-60	66
2	Anterior cingulate gyrus, right superior frontal gyrus	4.77	2449	1.75e-14	24	12	66
3	Left middle frontal gyrus,	5.59	1828	7.63e-12	-38	8	34
4	Right insula, right central opercular cortex	4.71	1794	1.08e-11	42	2	28
5	Right frontal pole, right middle frontal gyrus, right inferior frontal gyrus, pars triangularis	5.43	1300	2.37e-09	30	40	20
6	Left supramarginal gyrus, anterior division	4.61	959	1.19e-07	-64	-36	42
7	Left angular gyrus	5.83	916	2.38e-07	-48	-52	18
8	Right cerebellum, anterior	4.79	480	.000131	42	-38	-38
9	Posterior cingulate cortex, parahippocampal gyrus, right thalamus	4.41	424	.000328	14	-38	-2
10	Left temporal pole, left inferior frontal gyrus, pars opercularis left insula	4.08	413	.000394	-56	16	-6
11	Left cerebellum, anterior	5.44	263	.00598	-30	-40	-42
12	Right lingual gyrus	3.43	235	.0104	10	-74	-10
13	Left cerebellum, posterior	5.74	215	.0158	-14	-76	-42
14	Brainstem	4.35	207	.0186	8	-34	-20
	Frontal Theta Band Positive						
1	Right bilateral precentral gyrus	4.82	394	.000577	12	-16	80
2	Left bilateral precentral gyrus	5.25	357	.0011	-20	-28	78
	Frontal Theta Band Negative						
1	Right supramarginal gyrus, posterior division, right superior lateral occipital cortex	3.94	1002	1.10e-07	-54	-50	44
2	Left supramarginal gyrus, posterior division, Left superior lateral occipital cortex	4.39	508	8.96e-05	56	-50	20
3	Posterior cingulate cortex	4.58	419	.000378	-6	-30	38
4	Ventromedial prefrontal cortex	4.03	342	.00143	0	42	4
	Central Beta Band Positive						
1	Right caudate	4.19	258	.00481	16	30	6
2	Left parahippocampal gyrus, posterior divison	4.86	221	.0106	-38	-36	-8
	Central Beta Band Negative						
1	Right frontal pole, right middle frontal gyrus, right superior frontal gyrus	5.49	6599	7.06e-30	-32	8	28
2	Left frontal pole, left middle frontal gyrus, Left superior frontal gyrus	5.51	6144	1.82e-28	40	38	36
3	Left supramarginal gyrus, posterior	5.51	5175	2.43e-25	-66	-44	28

	division, left superior parietal lobule, left superior lateral occipital cortex, Left middle temporal gyrus, temporooccipital part						
4	Right supramarginal gyrus, posterior division, Right superior parietal lobule, right superior lateral occipital cortex	5.13	3264	1.62e-18	30	-74	54
5	Left superior frontal gyrus, paracingulate gyrus, precuneous	4.54	1235	1.80e-09	-4	12	52
6	Right superior temporal gyrus, posterior division	4.59	1076	1.33e-08	48	-14	-10
7	Left temporal pole, left planum temporale	4.96	320	.00139	-46	4	-18

S14: Go/NoGo difference in alpha (and beta) over time

We observed differences between trials with Go responses and trials with NoGo responses in the low alpha power before and shortly after outcome onset (Fig. 6A, B main text). Alpha typically increases over the time course of an experiment, potentially related to fatigue and decreasing arousal (*94*). If the ratio of Go and NoGo responses changed over time, as well, such an increase over time could spuriously lead to a difference between Go and NoGo responses (though note that this ratio did not noticeably change over time; Fig. S14D). To exclude this possibility, we extracted trial-by-trial time-frequency power from the three significant clusters report in the main text in which power differed between Go and NoGo responses: a) lower alpha band power after outcome onset, b) lower alpha band power before and after outcome onset, c) beta band power before outcome onset. We transformed this data to decibel and analyzed it as a function of the performed response (factor), block number (1–6; z-standardized), and the interaction between both. We reasoned that if power differences occurred merely due to fatigue effects, the main effect of performed response should not be significant when accounting for time on task (i.e., block number).

For lower alpha band power after outcome onset, there was a significant main effect of performed response, b = 0.035, SE = 0.015, $\chi^2(1)$ = 5.350, p = .021, with higher power for Go than NoGo responses, a significant main effect of block number with lower alpha band power increasing over time, b = 0.052, SE = 0.019, $\chi^2(1)$ = 6.645, p = .010, but no significant interaction, b = 0.003, SE = 0.008, $\chi^2(1)$ = 0.156, p = .693. As Fig. S14A reveals, lower alpha band power was consistently higher after Go than after NoGo responses for every block of the task, suggesting that differences in lower alpha band power were not merely due to time on task.

For lower alpha band power before and after outcome onset, as well, there was a significant main effect of performed response, b = 0.068, SE = 0.030, $\chi^2(1)$ = 5.010, p = .025, with higher power after Go than NoGo responses, a significant main effect of block number with lower alpha band power increasing over time, b = 0.072, SE = 0.029, $\chi^2(1)$ = 6.757, p = .016, but no significant interaction, b = 0.010, SE = 0.009, $\chi^2(1)$ = 1.184, p = .277 (Fig. S14B), leading to identical conclusions.

For beta band power before and after outcome onset, there was a significant main effect of performed response, b = 0.083, SE = 0.032, $\chi^2(1)$ = 6.301, p = .012, with higher power after Go than NoGo responses, a significant main effect of block number with beta power decreasing over time, b = -0.042, SE = 0.021, $\chi^2(1)$ = 4.007, p = .045, but no significant interaction, b = 0.001, SE = 0.007, $\chi^2(1)$ = 0.030, p = .864 (Fig. S14C). In sum, even in presence of changes in power over the time course of the task, lower alpha band and beta band power were consistently higher after Go responses than after NoGo responses, suggesting that these effects were not due to time on task.

Furthermore, we asked whether differences in ACC BOLD between trials with Go and trials with NoGo response at the time of the outcome were due to outcome-related activity or might rather the reflect action on the next trial. We thus plotted the “raw” BOLD signal per action x outcome condition. We used the first eigenvariate of the BOLD in signal in the ACC cluster that reflected biased learning, upsampled the BOLD signal, epoched it into trials relative to outcome onset (same procedure as for fMRI-informed EEG analyses), and averaged the signal across trials and participants separately per performed action (Go/NoGo) and outcome valence (positive/ negative). This plot yielded higher ACC BOLD signal on trials with NoGo responses than on trials with Go responses at the time of outcomes (Fig. S14E). However, this difference could potentially be driven by the response on the following task, so we further split the data according to whether the action on the following trial was a Go or a NoGo response. Irrespective of the action on the following trial, ACC BOLD signal was higher when the action on the current trial was a NoGo response compared to a Go response (Fig. S15F). In sum, these analyses corroborate that ACC BOLD signal was indeed higher after NoGo than Go responses at the time of outcomes.

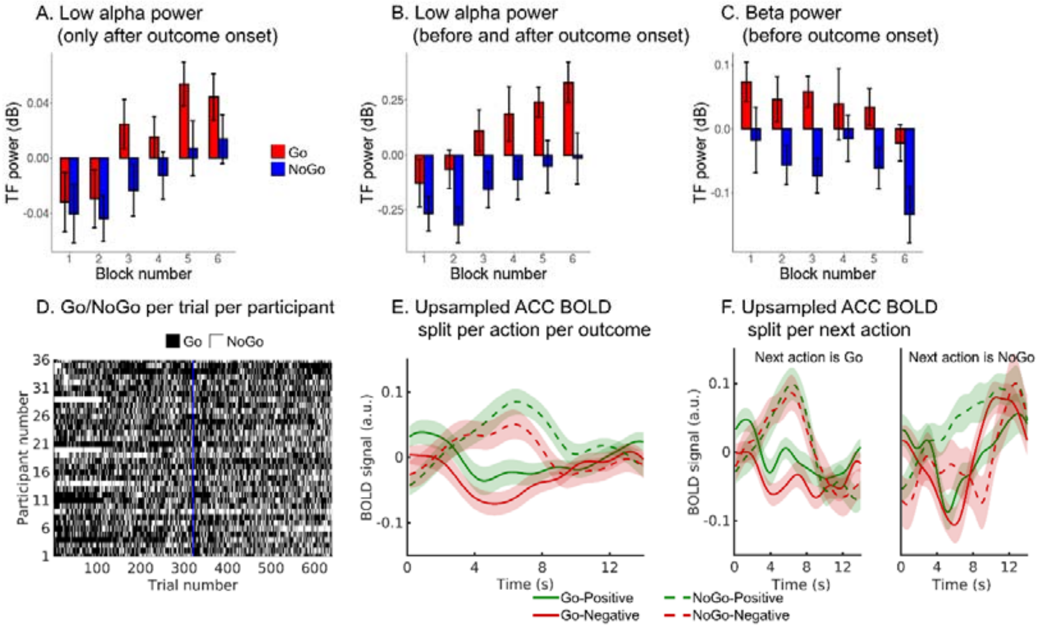

Figure S14. Control analyses excluding temporal confounds in midfrontal lower alpha band power and ACC BOLD. (A) Mean midfrontal low alpha power (±SEM across participants) after outcome onset, (B) before and after outcome onset, and (C) beta power before outcome onset as a function of the performed action and block number (i.e., time on task). While low alpha power increases and beta power decreases over the time course of the task, power is always consistently higher for trials with Go than trials with NoGo responses, suggesting that action effects are not reducible to time on task. (D) Response for each participant (rows) on each trial (columns). There is no noticeable change in the overall ratio of Go to NoGo responses over time. The vertical blue line indicates the start of the second session featuring new stimuli. (E) Mean upsampled ACC BOLD signal (±SEM across participants) at the time of the outcome, split per performed action (Go/NoGo) and outcome valence (positive/negative). BOLD signal is higher after NoGo than Go responses. (F) Same plot as (E), but split based on whether the next action is a Go (left panel) or an NoGo (right panel) response. Even if the next response is NoGo, BOLD signal is higher for trials with NoGo responses (on the current trial) than trials Go responses.

S15: Stay behavior as a function of BOLD and EEG TF power

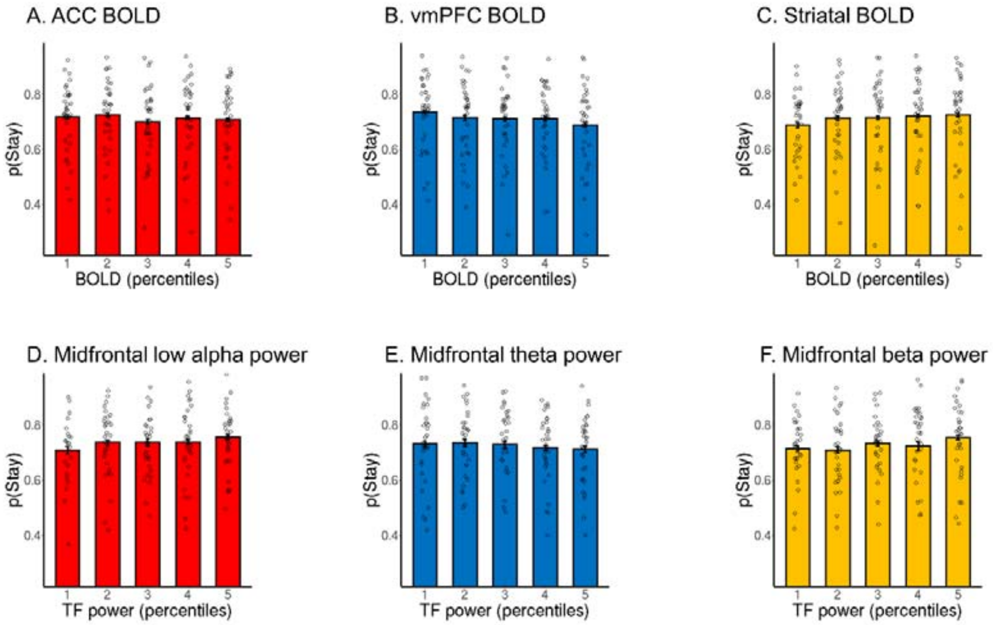

Figure S15. Probability of repeating the same response (“stay”) on the next cue encounter as a function of outcome-related BOLD and EEG signal. (A-C) Probability of repeating the same action (“staying”) as a function of BOLD signal from (A) ACC, (B) vmPFC, and (C) striatum (split into 5 bins). While ACC BOLD was not significantly linked to the probability to stay, high BOLD signal in vmPFC predicted a higher chance to switch to another action, while high BOLD signal in striatum predicted a higher probability of staying with the same action. (D-E) Probability of staying as a function of midfrontal time-frequency power in the (A) low alpha, (B) theta/delta, and (C) beta range. Higher low alpha power and higher beta power predict a higher probability of staying with the same action, while higher theta power predicts a higher chance to switch to another action. Grey circles represent individual per condition-per-participant means. Error bars are very narrow (and thus hardly visible) and computed based on the Cousineau-Morey methods based on per-condition-per-participant means.