Supplementary Materials

The PDF file includes:

Supplementary Notes 1-3

Figs. S1 to S5

Tables S1 to S4

Supplementary Notes1. Parameter recovery

We conducted a comprehensive parameter recovery analysis to validate our model fitting procedure. We generated synthetic data from 30 artificial subjects using the binary VKF combined with a softmax choice model. For parameter estimation, we employed the same Hierarchical Bayesian Inference (HBI) approach used in our empirical data analysis. To ensure robust estimation and minimize random variation effects, we repeated this procedure 20 times with different random seeds.

The recovery analysis revealed strong correlations between the true and recovered parameters, with median correlation coefficients across the 20 simulations: r_λ = 0.852, r_{v0} = 0.473, r_{σ^2} = 0.919, $r_{\beta 1}$ = 0.918, $r_{\beta 2}$ = 0.782.

For the visualization aim, we plotted relationships between simulated parameters and recovered parameters for one of the simulations (Fig. S1)

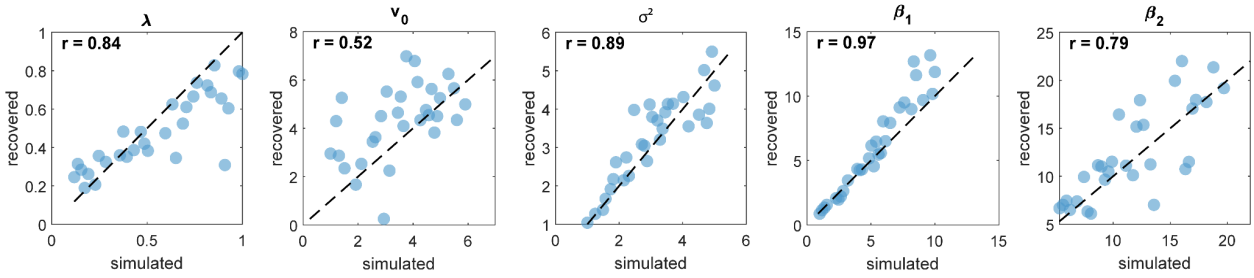

Fig. S1. Parameter recovery analysis.

The scatter plots show the relationship between simulated (true) and recovered parameters for the binary VKF model with softmax choice rule. Each point represents one artificial subject (n=30). The dashed lines indicate perfect recovery (y=x). Correlation coefficients (r) are shown for each parameter. Parameters shown are step size (λ), initial volatility parameter(v_0), observational noise (ω), and choice sensitivity parameters (β_1, β_2).

Note: Parameters were estimated using Hierarchical Bayesian Inference (HBI). Correlations indicate strong parameter recovery for most parameters, with moderate recovery for initial uncertainty (v_0).

Control analyses for initial volatility parameter(v_0)

While our main analyses treated the initial volatility parameter ($\underline{v_0}$) as a free parameter, we observed a relatively lower recovery rate for $\underline{v_0}$ compared to other parameters. Here we demonstrate that this lower recovery rate does not compromise the model's reliability or our main conclusions.

First, to validate our parameter estimation's robustness, we systematically analyzed parameter recovery across different fixed values ($\underline{v_0}$=[1, 3, 5, 7, 9]). Using the same set of simulated data, we found that the recovery rates for the key parameters (λ, σ^2, β1, β2) remained stable regardless of the fixed $\underline{v_0}$ value. Specifically, the correlations between true and recovered parameters maintained consistent levels (r_λ = 0.852, r_{σ^2}= 0.949, $r_{\beta 1}$ = 0.937, $r_{\beta 2}$= 0.831) across all $\underline{v_0}$ values, suggesting that $\underline{v_0}$ does not substantially interact with the recovery of other parameters.

Furthermore, when we compared model fits with $\underline{v_0}$ fixed at 5, the VKF-RVRU model still outperformed alternative models (BIC_{RW1}=13343, BIC_{RW2}=13303, BIC_{KF}=13481, BIC_{VKF}=13467, BIC_{VKF-RU}=13996, $BIC_{VKF-RVRU}$=**13243**), consistent with our main findings using the full model with free $\underline{v_0}$. This invariance to $\underline{v_0}$ demonstrates that while $\underline{v_0}$ shows lower recovery rates, this does not affect the model's ability to capture the key learning dynamics or our ability to reliably estimate the central parameters governing these dynamics. These results support our decision to retain $\underline{v_0}$ as a free parameter in the main analyses while providing evidence that its lower recovery rate does not impact the robustness of our primary findings.

Supplementary Notes 2. dmPFC did not signal subsequent decisions in feedback stage

We already know (from Fig.2), that the dmPFC did not differentiate between stay versus switch in the pre-selection stage. To further investigate whether dmPFC participates in action selection in the feedback stage, we analyzed its neural activity during the feedback stage using a linear mixed-effects model:

erd ~ NextTrials' (Stay vs. Switch) + reward (reward vs. non-reward) + previous trial feedback + (1|patientID) + (1|channelID:patientID)

Time-frequency analyses revealed no significant clusters differentiating between subsequent stay versus switch decisions in the dmPFC (Fig. S2, left panel). However, the dmPFC showed robust outcome-related activity, with distinct spectral signatures for reward versus non-reward feedback in both high-gamma (70-150 Hz) and theta (4-9 Hz) bands (Fig. S2, right panel). These results support our main finding that dmPFC primarily processes feedback information rather than directly encoding subsequent behavioral choices.

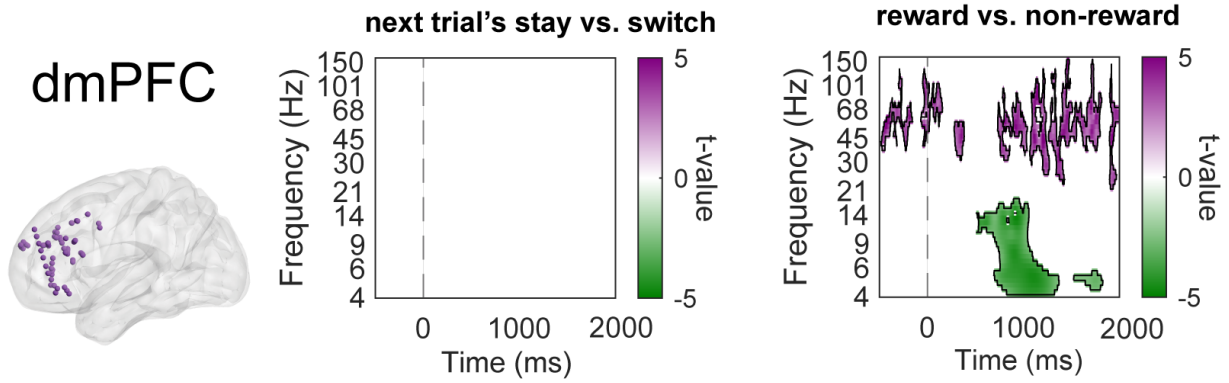

Fig. S2. dmPFC activity reflects feedback processing but not subsequent decisions

Time-frequency maps showing T-values from linear mixed-effects regression analyses of dmPFC local field potentials during feedback processing. Left: No significant differences between trials preceding stay versus switch decisions. Right: Significant differences between reward and non-reward feedback, particularly in high-gamma and theta bands. Color scales represent T-values, with warmer colors indicating higher values. Black outlines indicate significant clusters (cluster-based permutation tests, 5000 permutations, $p < 0.05$). Time 0 represents feedback onset. Frequency bands are displayed on the y-axis, ranging from 4 Hz to 150Hz.

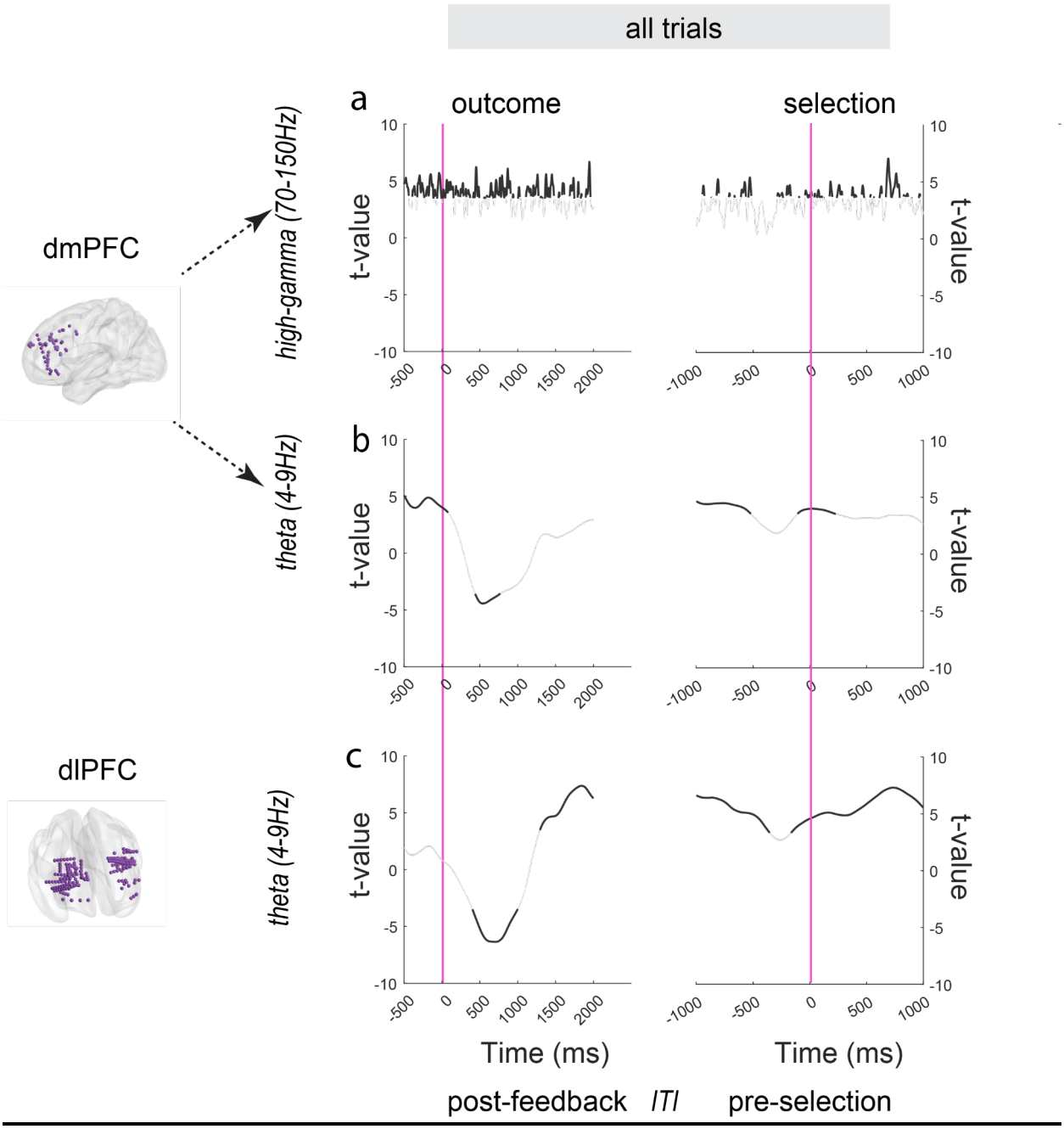

Fig. S3. Neural representations of prediction error in dlPFC and dmPFC across all trials.

(a) dmPFC high-gamma band (70-150 Hz), (b) dmPFC theta band (4-9 Hz), and (c) dlPFC theta band showing robust PE representation. Lines represent t-values from linear mixed-effects regression, with shaded areas indicating SEM. Black segments of the lines indicate time periods with significant prediction error representation ($p < 0.001$). Time 0 represents outcome onset (left) and next selection onset (right).

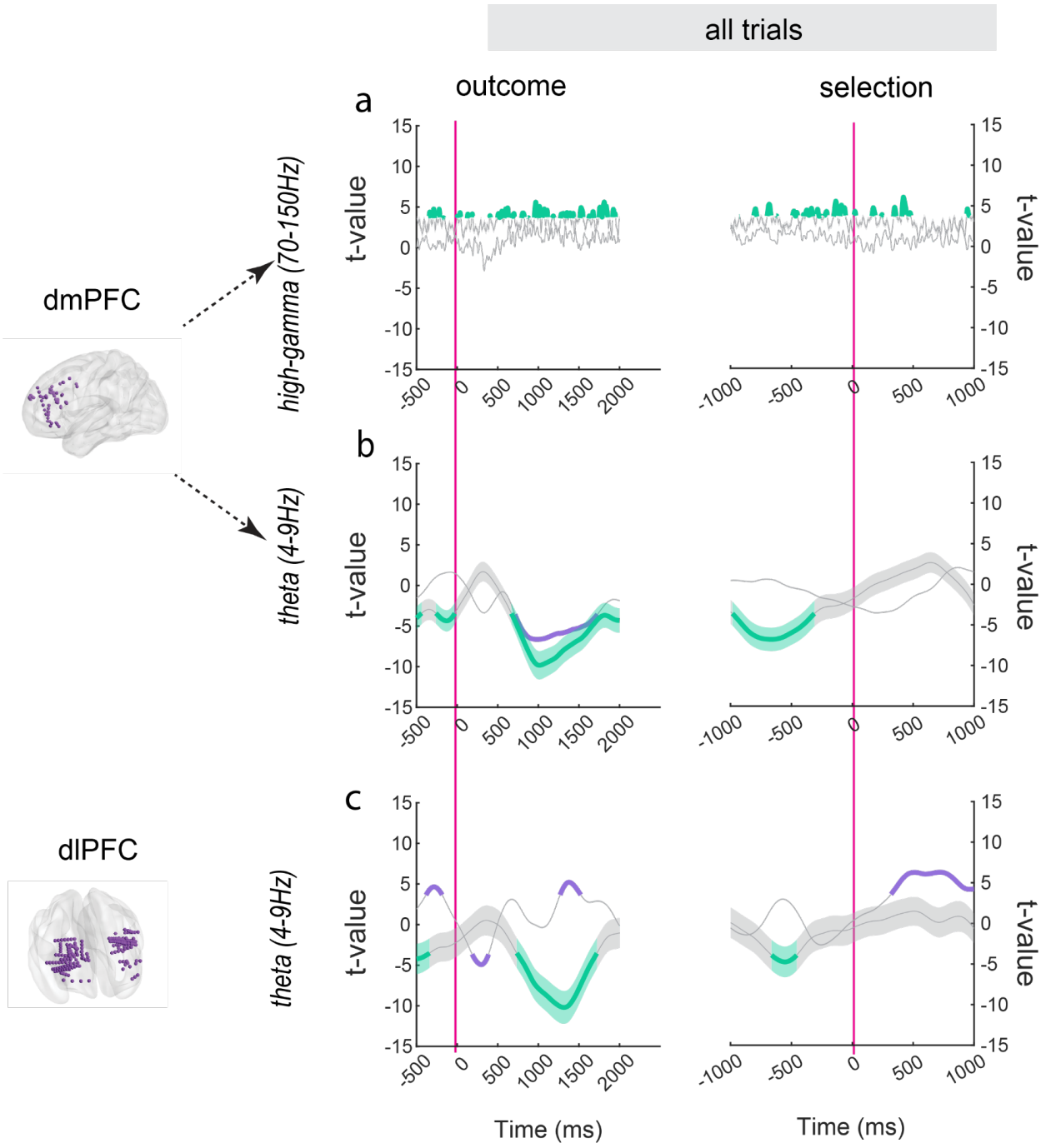

Fig. S4. Neural representations of relative value and uncertainty in dlPFC and dmPFC across all trials.

(a) dmPFC high-gamma band (70-150 Hz), (b) dmPFC theta band (4-9 Hz), and (c) dlPFC theta band showing robust PE representation. Lines represent t-values from linear mixed-effects regression, with shaded areas indicating SEM. Horizontal bars beneath each plot indicate periods of significant encoding ($p < 0.05$, cluster-corrected).

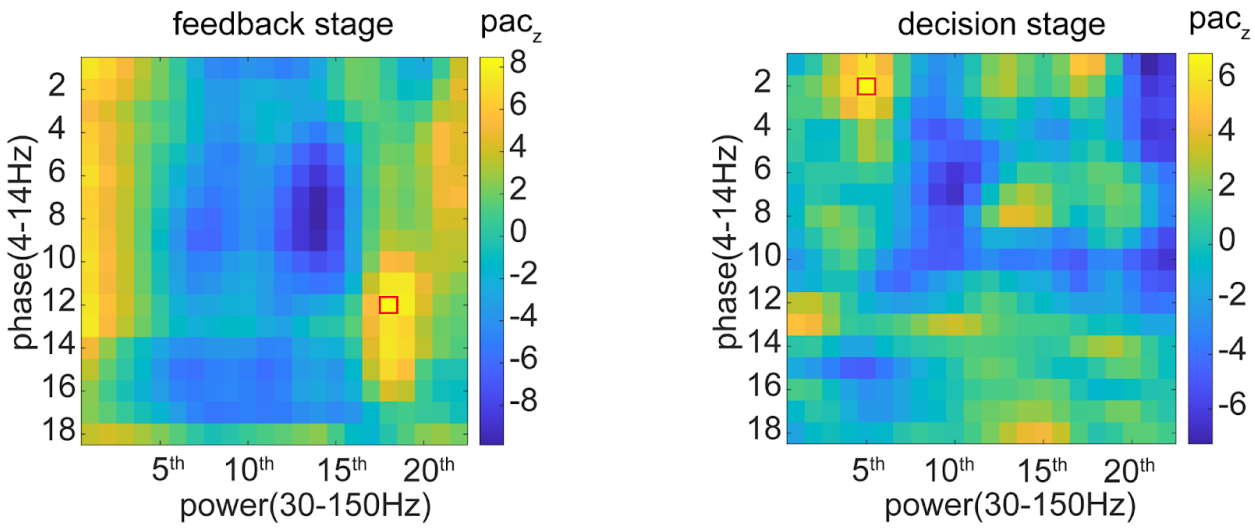

Fig. S5. Phase-amplitude coupling analysis between dmPFC high-frequency power and dlPFC low-frequency phase

Each panel shows a grid map of PAC z-values (permutation, 1000 times) for different frequency combinations. The x-axis shows the indices (5th, 10th, 15th, 20th) from the logarithmically spaced high-frequency power bands (30-150 Hz), while the y-axis shows the indices (2nd, 4th, 6th, ..., 18th) from the logarithmically spaced low-frequency phase bands (4-14 Hz). Each cell in the grid represents the PAC z-value for that specific phase-amplitude frequency combination, averaged across all trials and patients. Color intensity indicates the strength of coupling, with warmer colors representing stronger PAC.

Table S1.Patients demographics

ID	Gender	Age	handedness	Completed trial numbers
P01	Male	35	R	509
P02	Male	29	R	274
P03	Female	25	R	161
P04	Male	35	R	699
P05	Female	55	R	524
P06	Male	51	R	914
P07	Male	37	R	864
P08	Female	43	R	689
P09	Male	57	R	300
P10	Male	32	R	764
P11	Male	34	L	758
P12	Male	32	R	192
P13	Female	61	R	90
P14	Male	33	R	453

Notes: This table summarizes the patient demographics in the study. The table includes information on each patient's ID, gender, age, handedness, and the number of completed trials.

Table S2. Behavioral indices and response times

	Stay %	Switch %	Win. Stay	Lose.shift	RT_{stay}	RT_{switch}	$RT_{win\text{-}stay}$	$RT_{lose\text{-}switch}$
Mean	0.446	0.516	0.599	0.749	1.046	1.06	1.014	1.01
SD	0.183	0.174	0.249	0.148	0.451	0.384	0.424	0.374

Notes: Model-free behavioral measures and response times (RT) across participants. Stay % = percentage of trials where participants repeated their previous choice; Switch % = percentage of trials where participants changed their choice; Win-Stay = percentage of trials where participants repeated their choice following reward; Lose-Switch = percentage of trials where participants changed their choice following no reward; RT_{stay} = response time for stay decisions (in seconds); RT_{switch} = response time for switch decisions (in seconds); $RT_{win\text{-}stay}$ = response time for stay decisions following reward (in seconds); $RT_{lose\text{-}switch}$ = response time for switch decisions following no reward (in seconds). Values are shown as mean ± standard deviation across participants.

Table S3. More time information

	The interval between selection and outcome onset within the same trial	Inter-trial interval
Mean	0.461	0.626
SD	0.216	0.072

Notes: Values are shown as mean ± standard deviation across participants.

Table S4. Model performance

All models	XP (exceedance probabilities)	BIC
RW1	0.169	13343
RW2	0.006	13303
KF	0.001	13481
VKF	0.001	13544
VKF-RU	0.001	14043
VKF-RVRU	**0.822**	**13294**

Notes: Model comparison results using exceedance probabilities (XP) and Bayesian Information Criterion (BIC). Lower BIC values indicate better model fit. RW1 = standard Rescorla-Wagner model with single learning rate; RW2 = Rescorla-Wagner model with separate learning rates for reward and no-reward; KF = standard Kalman filter; VKF = volatile Kalman filter; VKF-RU = volatile Kalman filter with relative uncertainty; VKF-RVRU = volatile Kalman filter incorporating both relative value and relative uncertainty. The VKF-RVRU model showed the highest XP and lowest BIC, indicating it best explains participants' choice behavior.