RESEARCH ARTICLE SUMMARY

NEUROSCIENCE

Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex

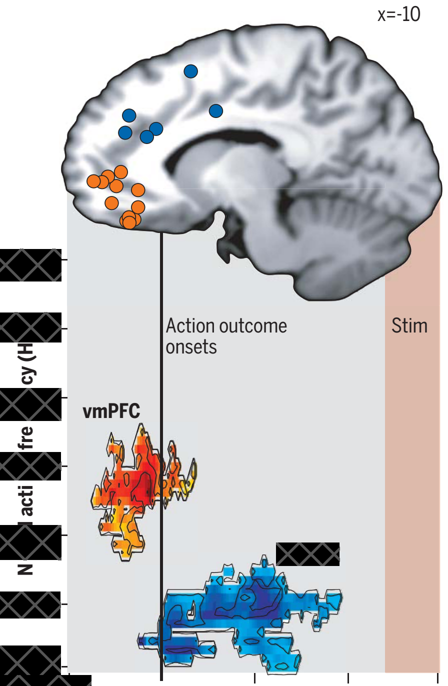
Philippe Domenech[1,3,4], Sylvain Rheims[5,6], Etienne Koechlin[1,2,7]*

[redacted]

[redacted]

[redacted]

[redacted]



[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[1]Institut National de la Santé et de la Recherche Médicale (INSERM), Paris, France. [2]Université Paris Sciences et Lettres (PSL) Research University, Ecole Normale Supérieure, Paris, France. [3]Paris Brain Institute, Paris, France. [4]APHP, Groupe Hospitalier Henri Mondor, DMU Psychiatry, Department of Neurosurgery, Université Paris Est Créteil, Créteil, France. [5]Department of Functional Neurology and Epileptology, Hospices Civils de Lyon, University of Lyon, Lyon, France. [6]Lyon's Neuroscience Research Center, INSERM-U1028, CNRS-UMR 5292, Lyon, France. [7]Université Pierre et Marie Curie, Paris, France.
***Corresponding author. Email: etienne.koechlin@upmc.fr**

Cite this article as P. Domenech *et al.*, *Science* 369, eabb0184 (2020). DOI: 10.1126/science.abb0184

S **READ THE FULL ARTICLE AT**
https://doi.org/10.1126/science.abb0184

RESEARCH ARTICLE

NEUROSCIENCE

Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex

Philippe Domenech[1,3,4], Sylvain Rheims[5,6], Etienne Koechlin[1,2,7]*

[redacted]

Everyday life frequently necessitates arbitration between the pursuit of an ongoing action plan with possible adjustments versus the exploring of a new plan. Resolving this so-called exploitation-exploration dilemma is critical for efficient adaptive behavior in uncertain, changing, and open-ended everyday environments (*1*, *2*) (supplementary note), and it primarily involves the medial prefrontal cortex (mPFC) (*3*–*8*). Human neuroimaging shows that activation in the ventromedial PFC (vmPFC) reflects the subjective value of the ongoing plan according to action outcomes, whereas the dorsomedial PFC (dmPFC) exhibits activation when this value drops and the plan is abandoned for exploring new ones (*6*). However, the neural mechanisms that resolve the dilemma and make the decision to exploit versus explore remain largely unknown. According to the classical view, the brain processes information in a feed-forward manner from stimuli to responses; namely, the vmPFC and dmPFC implement a bottom-up, reactive process evaluating the ongoing plan only after experiencing action outcomes to decide between further exploiting versus exploring. Alternatively, however, predictive coding that has been proposed for perception (*9*–*11*) could extend to the prefrontal executive function. Accordingly, the vmPFC and dmPFC might implement a top-down, proactive process wherein the ongoing plan is evaluated before experiencing action outcomes to prospectively code upcoming action outcomes as either learning signals to improve the ongoing plan or as potential triggers to explore new plans rather than exploit the ongoing plan. We addressed this issue by recording neuronal activity in six participants while they were performing a task that induced systematic exploitation-exploration dilemmas in an uncertain, changing, and open-ended environment (*2*, *6*) (Fig. 1).

Experimental paradigm

Participants were patients with drug-resistant focal epilepsy who had been stereotactically implanted with multilead electroencephalography (EEG) depth electrodes (*12*, *13*), which passed notably through the vmPFC and dmPFC (Fig. 1A and table S1). Participants were all eventually diagnosed with temporal or parietal lobe epilepsy with no impact observed on the PFC (see the Materials and methods section). Participants responded to visually, successively presented digits and searched for digit-response combinations by trial and error. Feedback was binary (positive versus negative) and stochastic (10% noise). Combinations changed episodically and unpredictably (every 33 to 57 trials), thereby dividing trial series into successive latent episodes associated with distinct combinations. Thus, the task induced participants to constantly arbitrate according to feedback between two options: (i) staying with the same presumed combination by possibly adjusting it or (ii) abandoning it to explore new combinations. In every trial, participants' responses could be correct (chance level, 25%), perseverating (incorrect in the current episode but correct in the preceding episode; chance level, 25%), or ancillary (neither correct nor perseverative; chance level, 50%) (Materials and methods). Overall, participants performed notably below the statistical optimum but similarly to healthy participants who had been previously tested in the same task (*2*, *6*). Correct response rates increased from ~5% at episode onsets to ~80%, perseverative response rates decreased from ~80% at episode onsets to ~5% after 25 trials, and ancillary response rates increased from ~15% at episode onsets, peaked to ~40% about 6 trials later, and returned to ~15% about 25 trials later (Fig. 2A).

Identifying covert switches into exploration

To determine when participants switched from exploiting presumed combinations to exploring new ones, we leveraged previous studies that have shown that these switches derive from an online algorithm approximating the optimal adaptive process (*2*, *6*). According to this model, presumed combinations form action plans associating digits, responses, and expected feedback. These plans are monitored online: The algorithm probabilistically infers the reliability value of presumed plans—i.e., the belief λ_i that plan i matches the current true combination—given observed action outcomes and the possibility that no monitored plans match. When plan i_0 is more likely matching than not matching the current combination—i.e., $\lambda_{i_0} > 1 - \lambda_{i_0}$ (or equivalently, $\lambda_{i_0} > 0.5$)—the plan is said to be reliable and the others are necessarily unreliable (because $\Sigma_i \lambda_i < 1$). The algorithm is then in the exploitation state: The reliable plan constitutes the actor that guides ongoing behavior and learns from feedback through reinforcement learning (RL) processes. When the actor becomes unreliable and all other monitored plans remain unreliable (all $\lambda_i < 0.5$), the algorithm switches into the exploration state: A new presumed plan is formed from mixing previously learned plans stored in long-term memory, and this new plan is used as a provisional actor guiding behavior and learning from feedback. The algorithm eventually returns to the exploitation state, when this provisional actor or another monitored plan becomes reliable. In the former case, the provisional actor is then consolidated in long-term memory, whereas it is disbanded in the latter case (see supplementary text for the full model description).

This model closely fitted and reproduced participants' performances (Fig. 2A), as has been explained and shown in previous studies (*2*, *6*, *14*). The model especially reveals when participants covertly switched from exploitation to exploration. After episode changes, the model switched, as expected, from exploitation to exploration at variable time points ranging from two to eight trials after episode onsets [95% confidence interval (CI)]. In these switch trials, feedback induced posterior actor reliability to decrease and drop below the reliability

[1] Institut National de la Santé et de la Recherche Médicale (INSERM), Paris, France. [2] Université Paris Sciences et Lettres (PSL) Research University, Ecole Normale Supérieure, Paris, France. [3] Paris Brain Institute, Paris, France. [4] APHP, Groupe Hospitalier Henri Mondor, DMU Psychiatry, Department of Neurosurgery, Université Paris Est Créteil, Créteil, France. [5] Department of Functional Neurology and Epileptology, Hospices Civils de Lyon, University of Lyon, Lyon, France. [6] Lyon's Neuroscience Research Center, INSERM-U1028, CNRS-UMR 5292, Lyon, France. [7] Université Pierre et Marie Curie, Paris, France.
*Corresponding author. Email: etienne.koechlin@upmc.fr

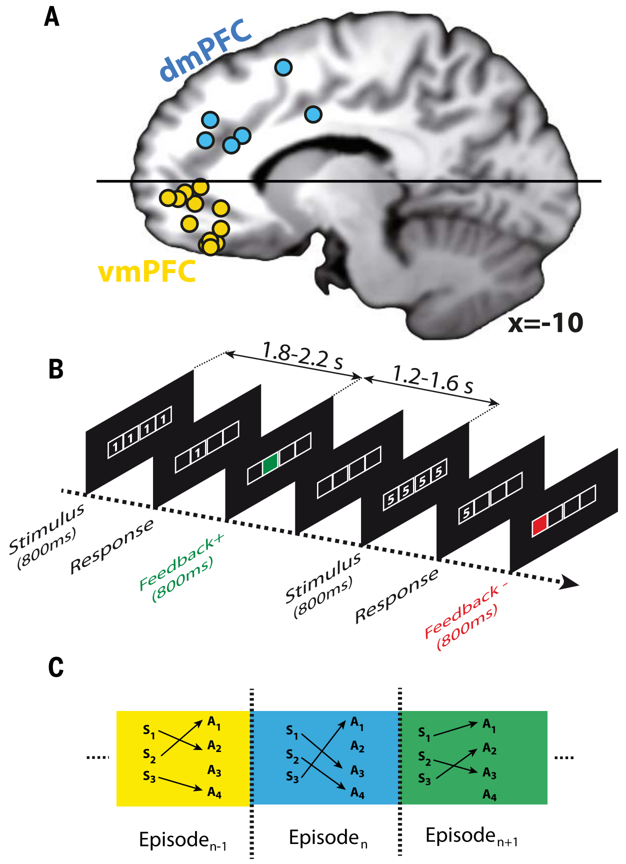


Fig. 1. Experimental protocol. (**A**) Localization of transverse electrodes within the medial PFC across participants, reconstructed on a canonical T1-weighted sagittal MRI brain slice [Montreal Neurological Institute (MNI) coordinate: x = −10]. Subgenual (yellow) and dorsogenual (blue) electrodes were ascribed to the vmPFC and dmPFC, respectively. The vmPFC and dmPFC comprised N = 13 and N = 12 electrode contacts, respectively. See fig. S1 for exact localizations in every participant. (**B**) Behavioral protocol. Trials started with the display of one out of three possible stimuli (digits; duration: 800 ms). Participants responded by pressing one out of four response buttons. Participants then received a positive or negative feedback (digits turned green or red, respectively; duration: 800 ms), depending on the current correct digit-button combination. Stimuli-feedback asynchrony was jittered (range: 400 ms). (**C**) Episode structure of the behavioral protocol. Current correct digit-button combinations episodically changed after an unpredictable number of trials, thereby defining successive latent episodes. There were no overlaps in digit-button associations ($S_i \rightarrow A_j$) between two successive episodes.

threshold (λ = 0.5), which led the model to start exploring in the subsequent trials. Realigning both the model's and participants' performances on switch trials rather than episode onsets (Fig. 2B) shows that both the model's and participants' responses were unaffected by episode changes up to these switch trials: Most responses remained perseverative (~85%), whereas residual responses were randomly distributed across ancillary and correct responses (~10 and ~5%, respectively). By contrast, in trials immediately after switch trials, both the model's and participants' perseverative responses abruptly dropped off to ~40%, whereas ancillary and correct responses abruptly increased close to their chance level (~40 and ~20%, respectively) (Fig. 2B). This abrupt algorithmic transition was also independently observed in participants' reaction times, which suddenly increased by ~100 ms in the trials immediately after switch trials (Fig. 2B)—an effect that could neither be ascribed to increases of response shifts after switch trials (fig. S5) nor to the occurrence of unexpected feedback (fig. S6). Thus, in switch trials, feedback induced participants to switch away from their ongoing plan to explore new plans in the next trials. In stay trials, by contrast, the model reveals that participants' behavior instead derived from RL processes—i.e., participants stayed with the same actor plan that adapted to external contingencies through RL processes—particularly when eventually returning from exploration to exploitation. Overall, switch trials occurred in the model contingent upon posterior rather than prior actor reliability relative to the actual feedback that participants observed. Switch trials predicted that participants behaviorally switched into exploration in the trials that immediately followed. Accordingly, participants covertly switched into exploration posterior to this feedback and before participants' responses that immediately followed.

Neural mechanisms inferring actor reliability

Using the standard time-frequency decomposition ([15]), we investigated whether local neural processing in the vmPFC and/or the dmPFC probabilistically infers and tracks actor reliability predicted by the model. We extracted neural activity in the high-gamma frequency band (50 to 150 Hz) reflecting local neuronal spiking ([16]) at each time point within the trials (Materials and methods). We entered this activity into a single multiple regression analysis performed over all trials that included model variables as within-subject regressors, including the following: (i) prior and posterior actor reliability relative to feedback, orthogonalized in that order so that the second regressor captures only reliability updates from feedback; (ii) the same reliability regressors for alternative plans monitored along the actor according to the model; and (iii) chosen values, i.e., the RL value of chosen actions, which, as feedback were binary, also measured positive feedback likelihoods involved in computing posterior from prior actor reliability. Additional regressors were included to remove potential confounding factors (Materials and methods).

vmPFC high-gamma activity linearly encoded prior actor reliability along virtually the whole trial epoch including intertrial intervals (Fig. 3). Moreover, ~300 ms after participants' responses, this activity further started encoding chosen values linearly. This encoding gradually increased up to feedback occurrences and then decreased afterward, which indicates that the vmPFC also anticipated when feedback occurred. Finally, ~350 ms after feedback occurrences, the activity further started encoding posterior actor reliability linearly. We detected no significant correlations with the reliability of alternative strategies (fig. S7), confirming the vmPFC specific role in monitoring the action plan driving ongoing behavior ([6]). In the dmPFC by contrast, high-gamma activity was uncorrelated with any preceding regressors (Fig. 3 and fig. S7). In both regions, finally, we found no significant correlations with neural activity in other frequency bands.

vmPFC and the prospect of exploration

To understand how monitoring of actor reliability leads to switch from exploitation to exploration, we compared neural activity locked on feedback onset between switch and stay trials. In the vmPFC, switch compared with neighboring stay trials elicited only a significant increase of neural activity in beta-band frequencies (13 to 30 Hz), which started at ~350 ms, peaked ~70 ms before feedback onset, and

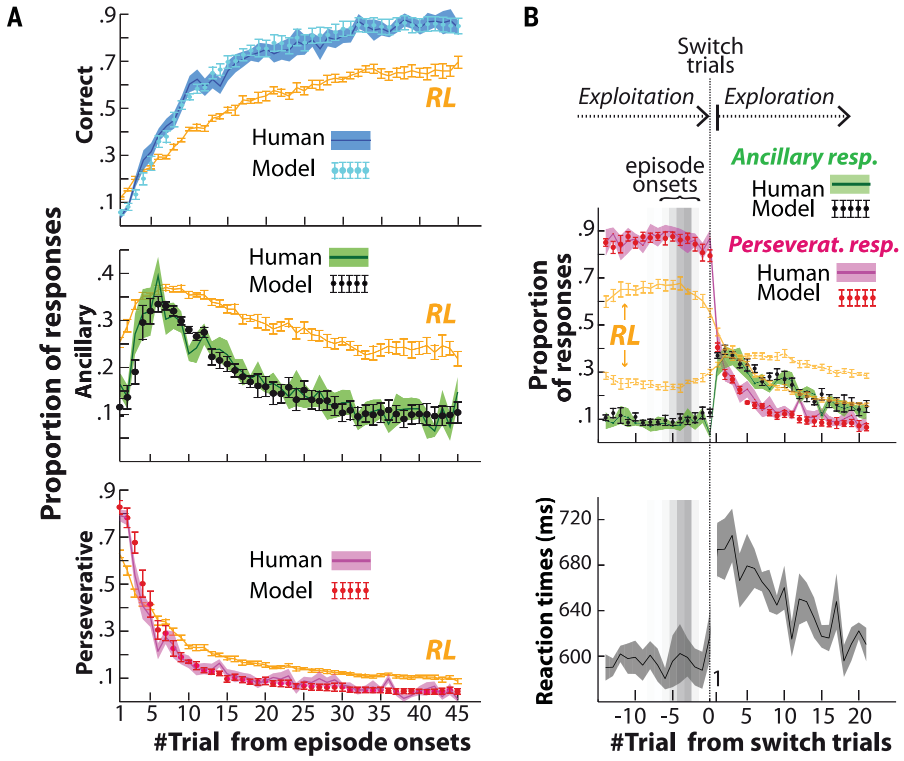


Fig. 2. Behavioral performances. (**A**) Proportion of correct, ancillary, and perseverative responses (summing up to 1) from participants and the model, according to the number of trials after episode onsets (i.e., combination changes). Ancillary and perseverative responses are both incorrect, but perseverative responses further correspond to correct responses in the preceding episode. Both the model and participants adapted much faster to combination changes than a Rescorla and Wagner's RL model that was fitted on participants' data (orange). (**B**) Participants' responses and model predictions (top) along with participants' reaction times (bottom) realigned on switch trials occurring in the model rather than on episode onsets. Orange lines show RL predictions. All model predictions are computed in every trial given participants' responses in previous trials. Error bars represent SEMs across participants. See Materials and methods for modeling details and tables S2 and S3 for model parameters.

vanished ~150 ms after feedback onset (Fig. 4A). This prefeedback effect is consistent with the documented role of beta-band activity in anticipating upcoming stimuli (*17*–*19*). The effect abruptly appeared in switch trials and was neither present in previous and subsequent neighboring trials (Fig. 4B) nor in stay trials that lead to either negative or positive feedback (Fig. 5). Thus, the effect was unlikely to reflect differences in reward or feedback expectations between stay and switch trials: In switch and immediately preceding stay trials, positive feedbacks were equally frequent and chosen values (or feedback likelihoods) were virtually identical (Fig. 4D). The effect was also unlikely to reflect actor reliability, which gradually decreased, whereas beta-band activity remained constant along stay trials preceding switch trials (Fig. 4D). We also dismissed the possibility that this prefeedback effect reflects the commitment to switch into exploration. The analysis of behavioral performances reported above indicates that participants covertly switched into exploration contingent upon and, consequently, posterior to these feedbacks. However, this analysis indicates that in switch compared with stay trials, the following prefeedback event occurred: Prior actor reliability monitored in the vmPFC (see Fig. 3) approached the 0.5 reliability threshold closely enough that upcoming feedback could cause posterior actor reliability to cross the threshold and trigger exploration. We therefore concluded that the prefeedback effect observed in the vmPFC reflects this event—namely, the sudden possibility that upcoming feedback may cause posterior actor reliability to cross the threshold or, equivalently, the fact that prior actor reliability is close enough to the threshold. Accordingly, the vmPFC appears to evaluate actor reliability relative to the threshold before feedback occurrences. This prospectively flags upcoming feedback in switch trials as potential triggers committing to exploration rather than as regular learning signals serving to adjust the ongoing actor plan.

dmPFC responses to exploration triggers versus learning signals

In the dmPFC, by contrast, switch compared with neighboring stay trials exhibited a significant differential neural activity only after feedback onset. The activity started at feedback onset and lasted up to ~200 ms after feedback offset (i.e., lasting ~1000 ms). This postfeedback effect occurred in theta-band frequencies (4 to 8 Hz) and slightly extended to alpha-band frequencies (8 to 12 Hz) (Fig. 4C). Although this postfeedback activity remained unchanged in stay trials preceding and after switch trials, it abruptly decreased in switch trials (Fig. 4B) (*20*, *21*). For the same reasons as above, this abrupt postfeedback effect could neither be ascribed to any variations in reward-feedback expectations or in actor reliability across stay and switch trials (Fig. 4D) nor to differences in reward prediction errors (RPEs), because these errors were virtually identical in switch and immediately preceding stay trials (Fig. 4D). Moreover, actor reliability was unrelated to dmPFC neural activity and was updated in the vmPFC only ~350 ms after feedback onset. Consequently, the effect emerging at feedback onset was unlikely to reflect a bottom-up, reactive process comparing posterior actor reliability with the 0.5 reliability threshold and leading to the decision to explore in switch trials. However, when this dmPFC postfeedback effect emerged at feedback onset, the vmPFC prefeedback effect started declining. This suggests that the vmPFC proactively configured the dmPFC to process feedback differently in stay and switch trials—i.e., as learning signals versus exploration triggers, respectively. Consistent with this interpretation, the vmPFC prefeedback effect arose in beta-band frequencies known to convey top-down, predictive neural processing (*22*–*24*), whereas the dmPFC postfeedback effect arose in theta-band frequencies considered as reflecting the configuration of PFC neural networks underpinning behavioral control (*25*–*28*).

In stay trials, dmPFC neural responses to positive and negative feedback exhibited a single common feature starting at feedback onset—namely, a strong increased activity in alpha-band frequencies that vanished ~600 ms later (Fig. 6A). dmPFC-centered alpha-band activities are thought to drive the inhibition of neural representations that are irrelevant to ongoing behavior (*29*–*33*), thereby favoring the maintenance of the ongoing actor plan and its adjustment in response to feedback. Consistently, dmPFC neural responses to positive and negative feedback exhibited a second common feature from ~200 to ~600 ms after feedback onset, which corresponded to the signature of RL processes—namely, a strong increase of high-gamma activity that correlated with unsigned RPEs (i.e., the discrepancy between chosen values and actual feedback) (*34*) (Fig. 6, A, C, and D). Thus, dmPFC neuronal responses to feedback in stay trials transiently encoded unsigned RPEs scaling RL processes. As expected, this postfeedback RPE encoding

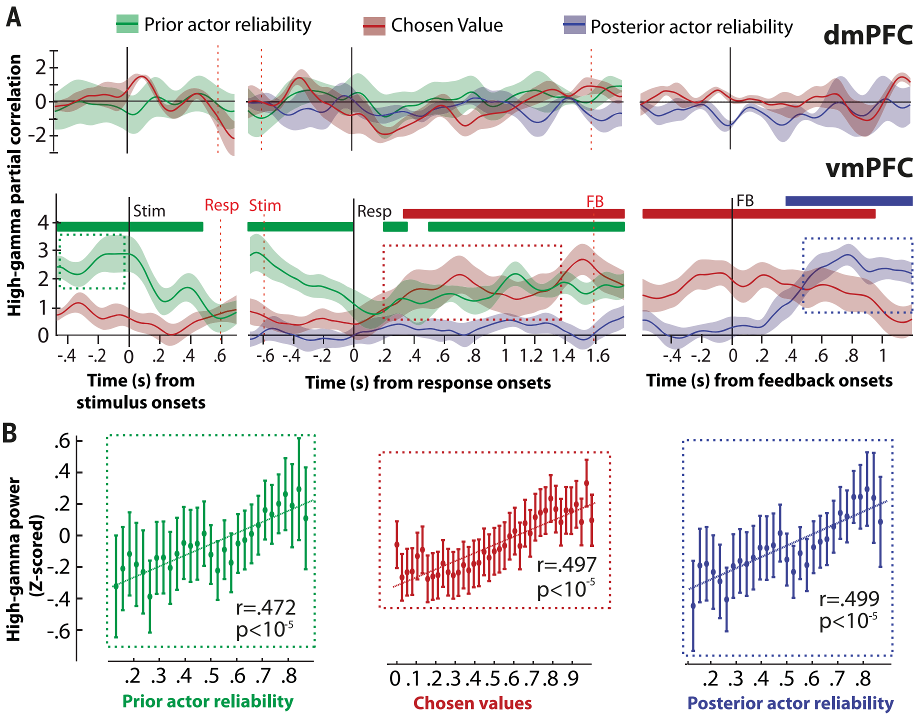


Fig. 3. Neural encoding of model variables in the medial PFC. (**A**) Time courses of partial correlation coefficients (betas) at each time point within trials between high-gamma neural activity (local field potentials >40 Hz), averaged over electrode contacts within the dmPFC and vmPFC and model variables (prior and posterior actor reliability, chosen values) derived from multiple regression analyses across all trials. Time courses are locked on stimulus onsets (left), participants' responses (middle), and feedback onset (right). Red vertical dashed lines show the average onsets of trial events. Shaded areas represent SEMs across contacts. Thick horizontal colored bars indicate statistical significance at $P < 0.05$, corrected for multiple comparisons at cluster level. The dmPFC exhibited no significant correlations. Resp, response; Stim, stimulus; FB, feedback. (**B**) vmPFC high-gamma activities (Z-scored for each electrode contact across trials before averaging) averaged over specific time windows [dashed boxes in (A)] and plotted against prior actor reliability (left), chosen values (middle), and posterior actor reliability (right). Error bars are SEMs across electrode contacts. Linear regression coefficients, r, are shown with P values. See fig. S2 for data from one vmPFC individual contact.

in the dmPFC was functionally connected with the prefeedback encoding of chosen values that we observed in the vmPFC (see the Psychophysiological interactions section in the Materials and methods; Fig. 7). vmPFC prefeedback and dmPFC postfeedback high-gamma activities were indeed correlated, and this cross-temporal correlation increased with chosen values (Fig. 7). This indicates that the chosen value representation in the vmPFC is transmitted to the dmPFC to subserve RPE computations. Consistently, the cross-temporal correlation further decreased when unsigned RPEs increased [i.e., decreased with the discrepancy between expected feedbacks (chosen values) encoded in the vmPFC and actual feedbacks processed in the dmPFC] (Fig. 7). By contrast, the cross-temporal correlation was unrelated to actor reliability (Fig. 7), which is in agreement with the absence of any dmPFC activity related to action plan reliability (Fig. 3 and fig. S7).

In switch trials, this cascade of dmPFC responses to feedback was disrupted. First, dmPFC theta-band activity in switch compared with stay trials decreased from feedback onset; second, the dmPFC alpha-band activity was notably reduced and rapidly vanished ~250 ms later (Fig. 6, A and B); third, this suppression was accompanied by a drop-off of dmPFC high-gamma activity observed in stay trials from ~200 to ~600 ms after feedback onset, which further ceased coding for any RPEs (Fig. 6, A and D). Thus, the dmPFC processed feedback in switch trials unlike learning signals, with the alpha-band response suppression presumably releasing the inhibition bearing upon neural representations that are irrelevant to the ongoing actor plan and preventing RL processes from adjusting this action plan. The dmPFC thus appears to process feedback in switch trials as favoring the covert emergence of neural representations forming new action plans to explore from ~250 ms after feedback onset.

Resolving exploitation-exploration dilemmas through predictive coding

vmPFC neural activity in gamma-band frequencies infers and tracks the reliability of the ongoing action plan according to action outcomes. After the action, it proactively flags (through beta-band frequencies) upcoming outcomes as either learning signals to better exploit this plan or potential triggers to explore new ones. According to this functional construct, dmPFC activity in theta-band frequencies appears to reflect the dmPFC configuration to respond to action outcomes. The dmPFC response to outcomes that are flagged as actual triggers then appears to realize the switch into exploration through the suppression of neural activity in alpha-band frequencies. This favors the emergence of neural representations that form new action plans and, through inhibiting dmPFC high-gamma activity–scaling RL processes, prevents the ongoing plan from adjusting through RL. Thus, the medial PFC resolves exploitation-exploration dilemmas through a top-down, predictive coding process from the vmPFC to the dmPFC. This predictive coding process has the advantage of speeding up the abandonment of the ongoing action plan and preventing action outcomes that trigger exploration from inappropriately acting as learning signals.

Predictive coding, which was originally developed to describe perceptual cortical processes (*9*–*11*, *35*), may also play a role in prefrontal executive processes. In perceptual predictive coding, observers' prior beliefs about a scene alter how they perceive the scene. Our findings suggest that within the prefrontal executive system, predictive coding proactively alters the functional signification of behavioral events according to the agents' beliefs about their own behavior.

Materials and methods

Participants

Six patients with drug-resistant focal epilepsy (one female; age range: 25 to 49 years old; see table S1) from the Department of Functional Neurology and Epileptology at University Hospital of Lyon participated in the present study. The participants belonged to a group of patients, which were stereotactically implanted with intracranial EEG depth electrodes to locate epileptic foci because noninvasive methods were unsuccessful (*13*, *36*). Implantation sites were selected according to clinical requirements, independent of the present study. We recruited the patients with at least one electrode implanted in the vmPFC and who were eventually diagnosed with temporal lobe (n = 5) or parietal lobe (n = 1) epilepsy with no electrophysiological impacts observed in the PFC. The study (DSI-SEEG protocol, NCT02869698) was approved by the Institutional Review Board (ANSM no. 2009-A00239-48) and National French

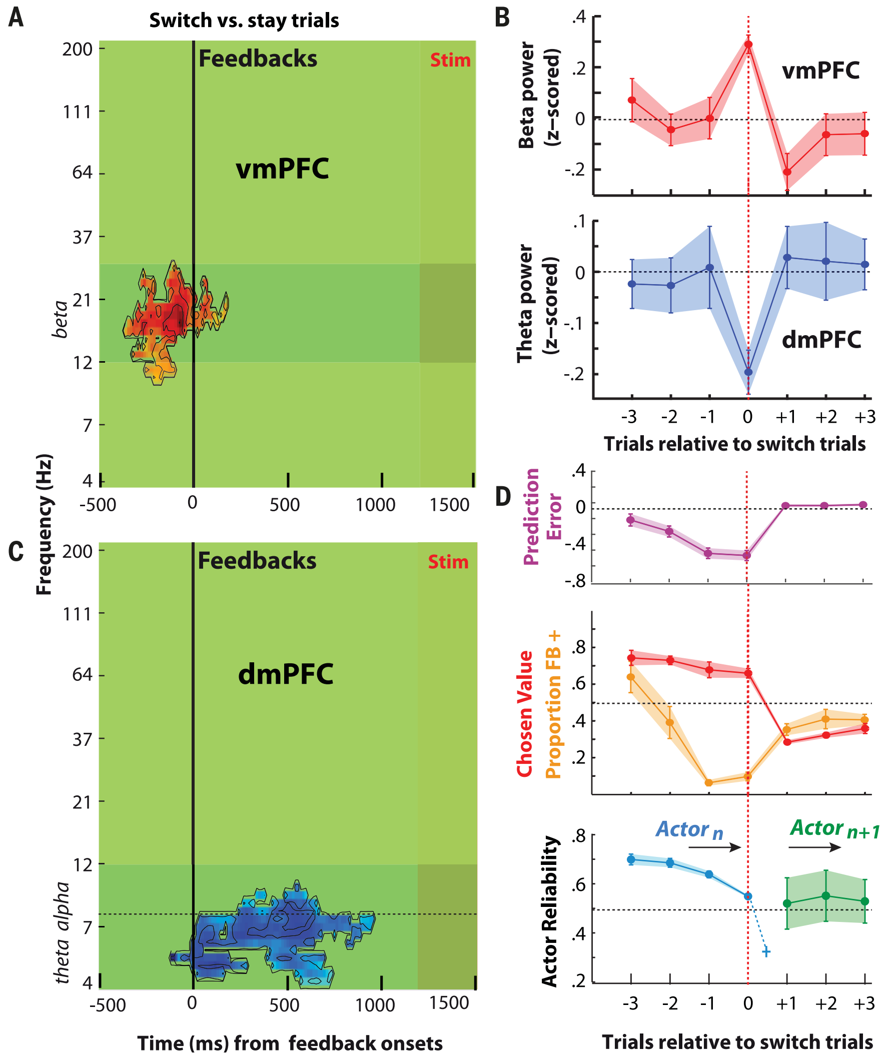


Fig. 4. Medial PFC neural activity associated with switch compared with stay trial feedback. (**A** and **C**) Time-frequency analyses (T-value maps) of neural local field potentials in switch compared with neighboring stay trials (from −2 to +2 trials relative to switch trials) locked on feedback and averaged over vmPFC (A) and dmPFC (C) electrode contacts. Power increases (positive T values) and decreases (negative T values) are shown in red and blue, respectively. T-value maps are thresholded at $P < 0.05$, corrected for multiple comparisons (cluster-level, family-wise error corrections). Black contours delimit statistical thresholds from $P < 0.05$ to $P < 5.0 \times 10^{-6}$. Vertical shaded areas indicate onset windows of stimuli from the next trial. See fig. S3 for the unthresholded maps. (**B**) Power amplitudes in switch and neighboring stay trials averaged over the vmPFC beta-band and dmPFC theta-band clusters shown in (A) and (C), respectively. Power amplitudes were Z-scored in each electrode contact before averaging. Error bars are SEMs over trials. (**D**) Signed prediction errors, chosen values, proportion of positive feedback (FB+), and prior actor reliability in switch and neighboring stay trials. In switch trials, posterior actor reliability dropped below the 0.5 reliability threshold (blue cross), and a new actor is formed to guide exploration in subsequent trials. Error bars are SEMs over participants.

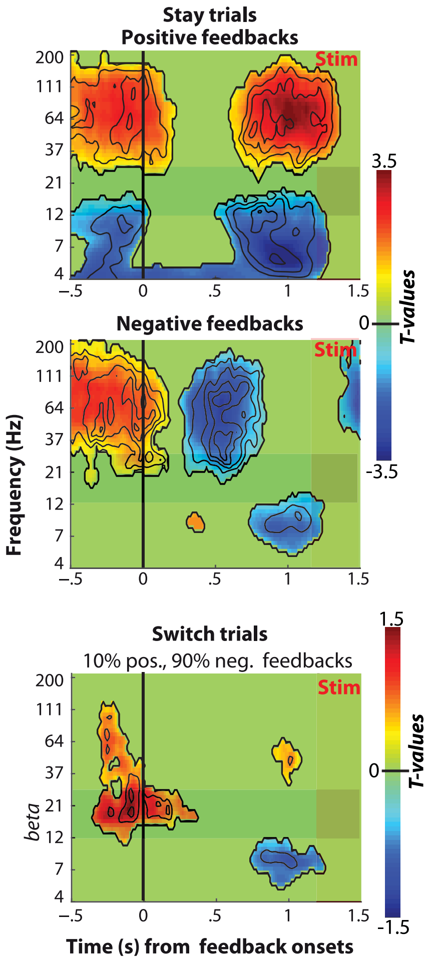


Fig. 5. vmPFC neural activity associated with response feedback. Time-frequency analyses (T-value maps) of vmPFC local field potentials relative to trial grand averages (from −2 s to +5 s relative to stimulus onsets), averaged across electrode contacts and locked on feedback onset in stay (top: positive feedback; middle: negative feedback) and switch (bottom) trials. Switch trials comprised 10% positive and 90% negative feedback. Maps are thresholded at $P < 0.05$, corrected for multiple comparisons (cluster-wise, family-wise error corrections). Increases (positive T values) and decreases (negative T values) are shown in red and blue, respectively. Black contours delimit statistical thresholds from $P < 0.05$ to $P < 0.000005$. Vertical shaded areas indicate onset windows of stimuli from the next trial. See fig. S4 for the unthresholded maps.

science ethic committee (CPP 09-CHUG-12, no. 0907). All participants volunteered to participate and provided written informed consent before participation.

Intracranial electroencephalography

We collected intracranial electroencephalography (iEEG) recordings from the six patients. They were chronically implanted with 12 to 15 stereotactic multilead depth electrodes for 2 to 3 weeks. These semirigid electrodes (DIXI Medical Instrument) have a diameter of 0.8 mm and, depending on the target structure, consisted of 5 to 15 linearly arranged contact leads (2 mm wide), with a 1.5-mm gap between two consecutive leads. Overall, we recorded from 929 contacts distributed across 81 depth electrodes, among which 21 depth electrodes (185 contacts) were located in the PFC. All these electrodes were implanted orthogonal to the interhemispheric plane with the deepest contacts located in the medial PFC. Over the six patients, 13 electrode contacts were localized in the vmPFC and 12 in the dmPFC (see Fig. 1A and fig. S1). T1-weighted anatomical magnetic resonance imaging (MRI) [three-dimensional gradient-recalled echo (3D GRE); resolution, 1 mm^3; matrix size, 256 voxels by 256 voxels] were acquired before and after surgery. Exact contact locations were manually determined

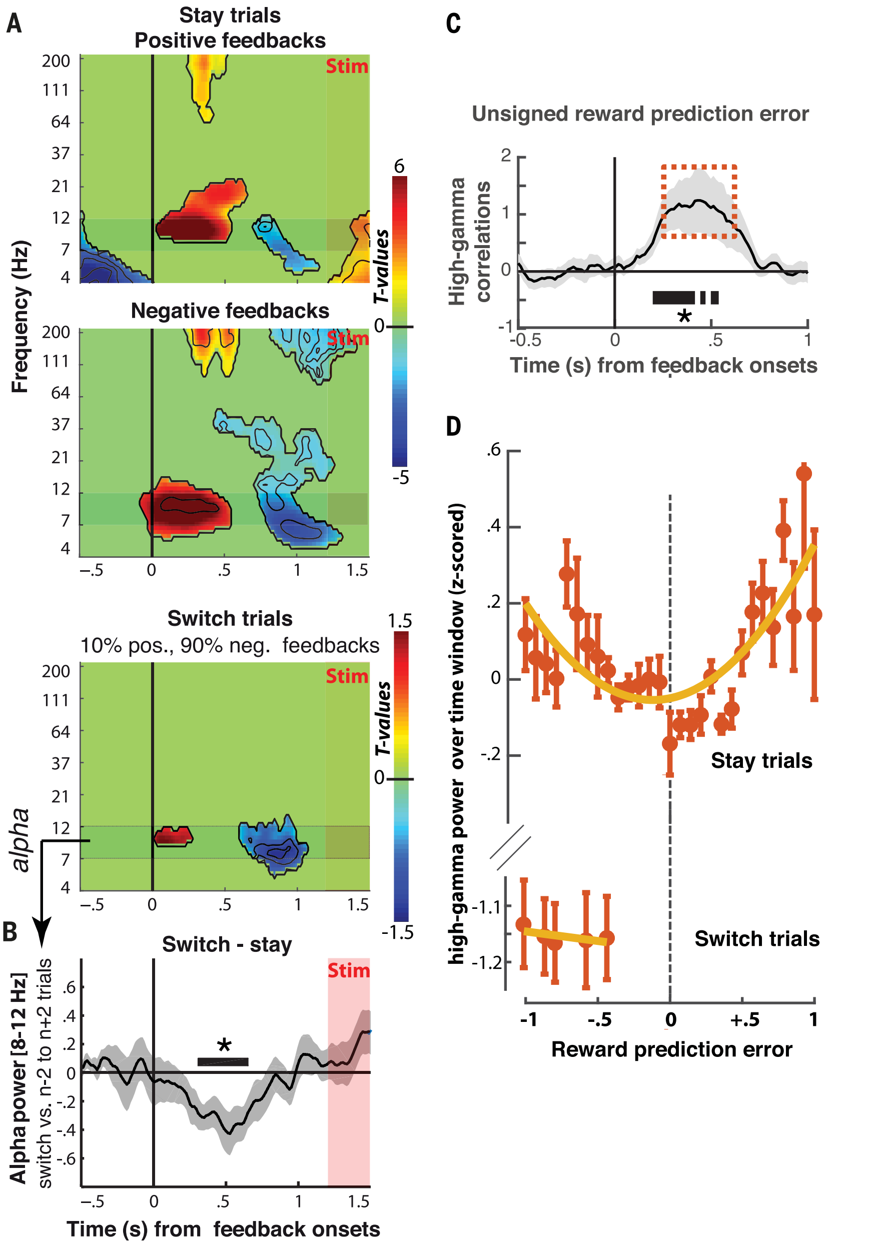


Fig. 6. dmPFC neural activity associated with response feedback. (**A**) Same as in Fig. 5 but for local field potentials averaged across dmPFC electrode contacts. See fig. S4 for the unthresholded maps. (**B**) Time courses of dmPFC alpha-band power in switch compared with neighboring stay trials (from −2 to +2 trials relative to switch trials) locked on feedback onset. (**C**) Correlation between dmPFC high-gamma neural activity and unsigned RPEs plotted against time from feedback onset. Shaded areas in (B) and (C) are SEMs across dmPFC electrode contacts. Horizontal black bars in (B) and (C) indicate statistical significance at $P < 0.05$, corrected for multiple comparisons (cluster-wise, family-wise error correction). (**D**) High-gamma activity over time window shown in (C) (orange) plotted against RPEs in stay and switch trials. Error bars are SEMs across binned trials. Lines show second-order polynomial regressions in stay trials (df = 340; linear $P = 0.34$; quadratic $P < 0.0001$) and in switch trials (df = 212; both linear and quadratic $P > 0.89$).

from the position of the corresponding artifact relative to the main anatomical landmarks on the postsurgery MRI. We recorded iEEG ~8 days after surgery (8.6 days ±1.4 SEM; see table S1) using a 128-channels video-EEG monitoring system (Micromed; sampling rate, 512 Hz). iEEG data were bandpass filtered online (0.1 to 200 Hz).

Experimental paradigm

The participants performed a variant of the Wisconsin Card Sorting Test, in which they learned combinations between digits and response buttons by trial and error. Combinations changed episodically and unpredictably. The experimental paradigm is identical to that used in previous studies testing healthy participants within and outside MRI scanners (*2*, *6*). In the present study, we only adjusted the event timing and jittering to iEEG constraints.

Four white boxes representing four response buttons were displayed on a black background at the center of a screen (Fig. 1B). Each trial started with the display of a white digit (out of three possible digits) in every box during 800 ms. The patients responded to this stimulus by pressing one of four buttons (response box: Cedrus Lumina, LU444-RH). Patients had to use the same finger to press the same button throughout the experiment. If a response occurred within 1500 ms from the stimulus onset, all displayed digits disappeared between 1800 and 2200 ms after the stimulus onset, except the digit displayed in the box related to the pressed button: If the participant pressed the correct button, this digit instead turned green with 90% probability (positive feedback) and turned red with 10% probability (negative feedback). If the participant pressed another button, positive and negative feedback probabilities were reversed. Response feedbacks were thus stochastic. If no responses occurred within 1500 ms from the stimulus onset, all digits disappeared and were replaced by an uninformative, neutral feedback (dashes in every box). Response feedbacks were presented during 800 ms. After 800 ms, the feedback disappeared, leaving the four boxes empty. The next trial started after a delay of 1200 to 1600 ms from the feedback onset. Participants were explicitly instructed that in every trial, each digit was associated with only one correct response button, and that distinct digits were associated with distinct responses (see Fig. 1C). Participants were also told that response feedbacks were not fully reliable (they were however not informed about the exact feedback probabilities). Finally, participants were informed that digit-response combinations could change episodically and unpredictably. No additional instructions were provided to participants. We refer to series of trials with no combination changes as episodes. Episode lengths pseudorandomly ranged between 33 and 57 trials. When combinations changed, every digit-response association was changed (Fig. 1C)

Overall, the experiment included two behavioral sessions administered on successive days. Each session included 1011 trials comprising 24 episodes. Each session included five short breaks occurring within episodes. After each break, the last digit-response combination was used again during six to nine trials, so that breaks were unrelated to combination changes (participants were explicitly instructed that breaks were unrelated to combination changes). Stimuli were pseudorandomly drawn from the set {1,3,5} for one session and {2,4,6} for the other session, counterbalanced across participants. In one session, three distinct digit-response combinations were pseudorandomly

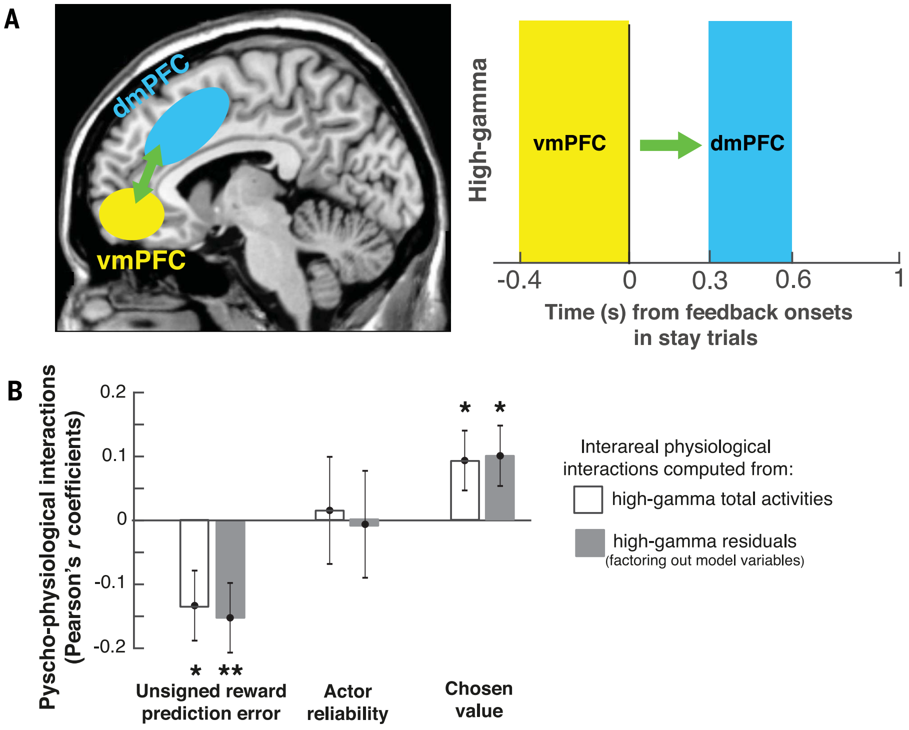


Fig. 7. Functional connectivity analysis from vmPFC to dmPFC high-gamma activities in stay trials. (**A**) The analysis investigated the physiological correlations between vmPFC high-gamma activity recorded from −400 ms to feedback onsets when this activity maximally encoded chosen values (Fig. 3A, right) and dmPFC high-gamma activity recorded from 300 to 600 ms after feedback onsets when this activity maximally encoded unsigned RPEs (Fig. 6C). Note that Fig. 3A shows vmPFC high-gamma activity to still encode chosen values and actor reliability over this postfeedback 300- to 600-ms time window. However, the analysis was restricted to prefeedback vmPFC activity to estimate the directional connectivity from vmPFC to dmPFC. Only the three participants with electrodes implanted in both the vmPFC and dmPFC were included in the analysis (see fig. S1). Physiological correlations were computed within every pair comprising one vmPFC and one dmPFC electrode contact. (**B**) Psychophysiological interactions (PPIs)—i.e., variations of physiological correlations with the model variables of interest, namely actor reliability, chosen value, and (unsigned) RPE. PPIs were computed for physiological correlations between high-gamma total activities as well as for correlations between high-gamma residual activities factoring out the influence of model variables on local neural activities. The graph shows that these physiological correlations increased with chosen values (*$P < 0.039$), decreased when unsigned RPEs increased (*$P < 0.016$ and **$P < 0.006$), and were unrelated to actor reliability ($P < 0.85$). Error bars are SEMs across contact pairs. As only the vmPFC encoded chosen values and actor reliability (Fig. 3A), the results indicate that vmPFC high-gamma activity conveys chosen value rather than actor reliability information to the dmPFC, with this chosen value information serving to compute RPEs that dmPFC high-gamma activity encoded (Fig. 6C).

repeated over episodes. In the other session, 24 distinct digit-response combinations were used across the 24 episodes (no combination repetitions). Order of sessions were counterbalanced across participants.

The day before the first session, the participants performed a short training session under supervision of the experimenter with a simplified protocol, using only two shapes (instead of three numbers) and two buttons (instead of four). This was done to ensure that participants correctly understood the task and experienced stochastic feedbacks and episode changes.

Computational model

In two previous studies (2, 6), we proposed a computational model describing how the human prefrontal function guides adaptive behavior in uncertain, changing, and open-ended environments. Collins and Koechlin (2) have described how the proposed model derives and approximates optimal adaptive processes (Dirichlet processes mixture) and, using experimental protocols comprising the present one, accounts for human adaptive performance better than various alternative models. Donoso *et al.* (6) have provided additional behavioral and neuroimaging evidence supporting the model and have revealed the central role of the medial PFC in arbitrating between exploiting the actor action plan versus exploring alternative plans. The present study relies on the exact same model, which is described in the supplementary text.

Model fitting procedures

The model has seven continuous parameters and one discrete parameter (buffer size N comprising monitored action plans, see model description in the supplementary text). To fit the model to each participant choice behavior, we used the slice-sampling Markov chain Monte Carlo method (37). Accordingly, we drew 7 million samples from the parameter posterior (with uniform priors over parameter ranges; four independent sets of 400,000 samples for each buffer size ranging from 1 to 6; 50,000 burn-in). We computed the model log-likelihood for each participant by summing the log-likelihoods provided by the model's epsilon-softmax function over all trials. To obtain robust individual parameter estimates, we computed the log-likelihood weighted average over all the samples drawn using the buffer size with maximal posterior probability (table S2). Finally, we computed trial-by-trial Monte Carlo estimates of model variables (prior and posterior reliabilities, choice values, etc.) by randomly resampling from our full sample set (70,000 resamples) and computing their log-likelihood weighted average.

Preprocessing of iEEG data

In the present study, we only report data from electrode contacts located in the mPFC. We performed all preprocessing steps using EEGLAB (38). First, we removed electrical artifacts from power lines and medical equipment using frequency-domain regressions and a threshold based on a Thompson F statistic for artifact detection (Cleanline, http://chronux.org; bandwidth 2 Hz, sliding window length of 2- and 1.5-s steps). Then, iEEG data were epoched from −2000 ms before stimulus onset to +5000 ms after stimulus onset, and detrended using third degree polynomial fits (overall participants performed 2022 trials over two sessions leading to 2022 epochs). Each contact trace was subsequently re-referenced with respect to its nearest neighbor along the same electrode (bipolar derivation). We used bipolar rather than unipolar derivations because it allows for better signal artifact removal and achieves high spatial resolution (~3 mm^3) by canceling out contributions of distant sources which spread equally to recording sites (39). Epochs with possible epileptic spikes, electrical artefacts along with those corresponding to missed trials and to the first six to nine trials after breaks were removed. Overall, we excluded ~5% of epochs.

Time-frequency analyses were carried out using the FieldTrip toolbox for MATLAB ([15]). Spectral powers were estimated using an adaptive multi-tapering time-frequency transform ([40]) (Slepian tapers; lower frequency range: 4 to 32 Hz, six cycles, and three tapers per window; higher frequency range: 32 to 200 Hz, fixed time windows of 240 ms, 4 to 31 tapers per window). This approach uses a constant number of cycles across frequencies <32 Hz (time window durations decrease when frequencies increase) and above 32 Hz, a fixed time window with an increasing number of tapers to obtain more precise power estimates (smoothing adaptively increases with frequencies).

iEEG data analyses

Model-based analyses of gamma-band activity

We first computed local activity reflecting neural spiking by averaging spectral powers (in decibels) across gamma and high-gamma frequencies at each time point in each trial. For each contact, we characterized the lower and upper boundaries of the gamma band on the full time-frequency map averaged across trials. On average, the lower bound was 41.9 ± 5.6 Hz (mean ± SEM; minimum, 40 Hz; maximum, 60 Hz), and the upper bound was 140.4 ± 30.8 Hz (mean ± SEM; minimum, 110 Hz; maximum, 200 Hz). Extracting the gamma-band activity using fixed boundaries (50 to 150 Hz) instead of contact specific gamma bands changed neither the pattern nor the overall significance of the reported results (Fig. 3 and fig. S7).

Second, we regressed across trials the gamma-band amplitude computed as above against key parametric variables from the model. More specifically, we estimated a general linear model predicting the trial-to-trial variability in gamma-band amplitude independently at each time step (13.8 ms) within time windows locked on events of interest (from −700 ms to +700 ms relative to stimulus onsets, from −700 ms to +1800 ms relative to responses, from −600 ms to +1200 ms relative to feedback onsets). The general linear model included the following parametric regressors: (i) the posterior reliability of the actor plan in the previous trial, its prior and posterior reliability in the current trial, orthogonalized in that order so that the current prior and posterior reliability regressor properly reflects the sequences of reliability updates; (ii) the same regressors for the most reliable alternative plan in the monitoring buffer; (iii) the RL chosen value $Q^t_{\text{actor}}(s_t, a_{\text{chosen}})$ in the previous and current trial, orthogonalized in that order; and (iv) two potentially confounding factors, namely choice uncertainty [entropy over current actor Q values $Q^t_{\text{actor}}(s_t, a_t)$] and reaction time. For model variable regressors, we used the Monte Carlo trial-by-trial estimates of model variables described in section model fitting procedure above. These regressions were carried out separately for each of contact, then averaged across contacts and participants to generate group-level averages. Fig. 3A shows the time courses of betas associated with the regressors of interest—i.e., the partial correlation slope between gamma-band activity and model variables at every time step. Statistical inferences were performed with a statistical threshold of $P < 0.05$ corrected for multiple comparisons [Bonferroni-Holmes Family-Wise-Error corrections at the cluster level ([41]); n = 150 permutations within contacts, n = 50,000 permutations between contacts].

Analyses of switch compared with stay trials

For each contact, we computed the paired difference in frequency powers at each time step (from −500 ms to +1500 ms relative to feedback onsets) for every frequency (4 to 200 Hz) between switch trials and the two neighboring stay trials (from −2 to +2 trials relative to switch trials). We then computed the corresponding T values, which we averaged across contacts and participants (Fig. 4A). Statistical inferences were performed with a statistical threshold of $P < 0.05$ corrected for multiple comparisons [Bonferroni-Holmes Family-Wise-Error corrections at the cluster level ([41]); n = 150 permutations within contacts, n = 50,000 permutations between contacts].

Neural activity associated with response feedbacks

To analyze the vmPFC and dmPFC local field potentials associated with response feedbacks (Fig. 5 and Fig. 6), we computed in every trial the frequency power at each time step (from −500 ms to +1500 ms relative to feedback onset) relative to the trial grand average (from −2 s to +5 s around stimulus onset) and we averaged across trials. We then computed the mean across contacts and participants. Statistical inferences were performed with a statistical threshold of $P < 0.05$ corrected for multiple comparisons [Bonferroni-Holmes Family-Wise-Error corrections at the cluster level ([41]); n = 150 permutations within contacts, n = 50,000 permutations between contacts].

Psychophysiological interactions

We performed psychophysiological interaction (PPI) analyses (Fig. 7) in stay trials ([42]) between prefeedback high-gamma activities in the vmPFC (averaged from −400 ms to feedback onsets), which encoded chosen values and prior actor reliability (Fig. 3A) and post-feedback high-gamma activities in the dmPFC (averaged from +300 ms to +600 ms after feedback onsets), which encoded unsigned prediction errors (Fig. 6C). We thus tested whether in stay trials, the correlation between these vmPFC and dmPFC activities varied with each of these three model variables (chosen values, actor reliability, and unsigned prediction errors). We performed the three corresponding PPI analyses across all pairs of vmPFC-dmPFC contacts in participants with electrodes implanted in both the vmPFC and dmPFC (three participants, fig. S1, for a total of n = 18 contact pairs). For each analysis, we first sorted the stay trials into 30 bins according to the model variable of interest. Across the trials within each bin, we then computed for each contact pair the correlation between vmPFC and dmPFC contact activities, namely the Fisher Z-transformed correlation coefficient. We then estimated the PPI for every contact pair by computing the Pearson's r correlation between these Fisher Z-transformed correlation coefficients and the model variable of interest. PPI statistical significances were assessed by entering the Pearson's r correlations across contact pairs into one-sample two-sided t tests. The results are reported in Fig. 7 (white bars). For completeness, we also performed the same three analyses from residual rather than total high-gamma activities, i.e., after regressing out the joint contribution of the three model variables to every electrode contact activity (Fig. 7, gray bars). PPI results were robust to changes in the number of bins and in the exact boundaries of time windows used for extracting high-gamma activities.

REFERENCES AND NOTES

1. J. D. Cohen, S. M. McClure, A. J. Yu, Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Phil. Trans. R. Soc. B* **362**, 933–942 (2007). doi: 10.1098/rstb.2007.2098; pmid: 17395573
2. A. Collins, E. Koechlin, Reasoning, learning, and creativity: Frontal lobe function and human decision-making. *PLOS Biol.* **10**, e1001293 (2012). doi: 10.1371/journal.pbio.1001293; pmid: 22479152
3. B. Y. Hayden, J. M. Pearson, M. L. Platt, Neuronal basis of sequential foraging decisions in a patchy environment. *Nat. Neurosci.* **14**, 933–939 (2011). doi: 10.1038/nn.2856; pmid: 21642973
4. N. Kolling, T. E. Behrens, R. B. Mars, M. F. Rushworth, Neural mechanisms of foraging. *Science* **336**, 95–98 (2012). doi: 10.1126/science.1216930; pmid: 22491854
5. H. Seo, X. Cai, C. H. Donahue, D. Lee, Neural correlates of strategic reasoning during competitive games. *Science* **346**, 340–343 (2014). doi: 10.1126/science.1256254; pmid: 25236468
6. M. Donoso, A. G. Collins, E. Koechlin, Human cognition. Foundations of human reasoning in the prefrontal cortex. *Science* **344**, 1481–1486 (2014). doi: 10.1126/science.1252254; pmid: 24876345
7. T. C. Blanchard, S. J. Gershman, Pure correlates of exploration and exploitation in the human brain. *Cogn. Affect. Behav. Neurosci.* **18**, 117–126 (2018). doi: 10.3758/s13415-017-0556-2; pmid: 29218570
8. M. Sarafyazd, M. Jazayeri, Hierarchical reasoning by neural circuits in the frontal cortex. *Science* **364**, eaav8911 (2019). doi: 10.1126/science.aav8911; pmid: 31097640
9. R. P. Rao, D. H. Ballard, Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999). doi: 10.1038/4580; pmid: 10195184

10. K. Friston, A theory of cortical responses. *Phil. Trans. R. Soc. B* **360**, 815–836 (2005). doi: 10.1098/rstb.2005.1622; pmid: 15937014
11. T. Lochmann, S. Deneve, Neural processing as causal inference. *Curr. Opin. Neurobiol.* **21**, 774–781 (2011). doi: 10.1016/j.conb.2011.05.018; pmid: 21742484
12. J. Parvizi, S. Kastner, Promises and limitations of human intracranial electroencephalography. *Nat. Neurosci.* **21**, 474–483 (2018). doi: 10.1038/s41593-018-0108-2; pmid: 29507407
13. P. Ryvlin, J. H. Cross, S. Rheims, Epilepsy surgery in children and adults. *Lancet Neurol.* **13**, 1114–1126 (2014). doi: 10.1016/S1474-4422(14)70156-5; pmid: 25316018
14. T. Hare, Exploiting and exploring the options. *Science* **344**, 1446–1447 (2014). doi: 10.1126/science.1256862; pmid: 24970062
15. 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**, 156869 (2011). doi: 10.1155/2011/156869; pmid: 21253357
16. S. Ray, N. E. Crone, E. Niebur, P. J. Franaszczuk, S. S. Hsiao, Neural correlates of high-gamma oscillations (60-200 Hz) in macaque local field potentials and their potential implications in electrocorticography. *J. Neurosci.* **28**, 11526–11536 (2008). doi: 10.1523/JNEUROSCI.2848-08.2008; pmid: 18987189
17. H. Liang, S. L. Bressler, M. Ding, W. A. Truccolo, R. Nakamura, Synchronized activity in prefrontal cortex during anticipation of visuomotor processing. *Neuroreport* **13**, 2011–2015 (2002). doi: 10.1097/00001756-200211150-00004; pmid: 12438916
18. J. Gross *et al*., Anticipatory control of long-range phase synchronization. *Eur. J. Neurosci.* **24**, 2057–2060 (2006). doi: 10.1111/j.1460-9568.2006.05082.x; pmid: 17067302
19. Y. Zhang, X. Wang, S. L. Bressler, Y. Chen, M. Ding, Prestimulus cortical activity is correlated with speed of visuomotor processing. *J. Cogn. Neurosci.* **20**, 1915–1925 (2008). doi: 10.1162/jocn.2008.20132; pmid: 18370597
20. J. F. Cavanagh, Cortical delta activity reflects reward prediction error and related behavioral adjustments, but at different times. *Neuroimage* **110**, 205–216 (2015). doi: 10.1016/j.neuroimage.2015.02.007; pmid: 25676913
21. J. F. Cavanagh, M. J. Frank, T. J. Klein, J. J. Allen, Frontal theta links prediction errors to behavioral adaptation in reinforcement learning. *Neuroimage* **49**, 3198–3209 (2010). doi: 10.1016/j.neuroimage.2009.11.080; pmid: 19969093
22. X. J. Wang, Neurophysiological and computational principles of cortical rhythms in cognition. *Physiol. Rev.* **90**, 1195–1268 (2010). doi: 10.1152/physrev.00035.2008; pmid: 20664082
23. A. K. Engel, P. Fries, Beta-band oscillations—Signalling the status quo? *Curr. Opin. Neurobiol.* **20**, 156–165 (2010). doi: 10.1016/j.conb.2010.02.015; pmid: 20359884
24. A. Oswal, V. Litvak, P. Sauleau, P. Brown, Beta reactivity, prospective facilitation of executive processing, and its dependence on dopaminergic therapy in Parkinson's disease. *J. Neurosci.* **32**, 9909–9916 (2012). doi: 10.1523/JNEUROSCI.0275-12.2012; pmid: 22815506
25. G. Buzsáki, A. Draguhn, Neuronal oscillations in cortical networks. *Science* **304**, 1926–1929 (2004). doi: 10.1126/science.1099745; pmid: 15218136
26. J. F. Cavanagh, M. J. Frank, Frontal theta as a mechanism for cognitive control. *Trends Cogn. Sci.* **18**, 414–421 (2014). doi: 10.1016/j.tics.2014.04.012; pmid: 24835663
27. B. Voytek *et al*., Oscillatory dynamics coordinating human frontal networks in support of goal maintenance. *Nat. Neurosci.* **18**, 1318–1324 (2015). doi: 10.1038/nn.4071; pmid: 26214371
28. R. F. Helfrich, R. T. Knight, Oscillatory Dynamics of Prefrontal Cognitive Control. *Trends Cogn. Sci.* **20**, 916–930 (2016). doi: 10.1016/j.tics.2016.09.007; pmid: 27743685
29. W. Klimesch, P. Sauseng, S. Hanslmayr, EEG alpha oscillations: The inhibition-timing hypothesis. *Brain Res. Rev.* **53**, 63–88 (2007). doi: 10.1016/j.brainresrev.2006.06.003; pmid: 16887192
30. K. E. Mathewson *et al*., Pulsed out of awareness: EEG alpha oscillations represent a pulsed-inhibition of ongoing cortical processing. *Front. Psychol.* **2**, 99 (2011). doi: 10.3389/fpsyg.2011.00099; pmid: 21779257
31. S. Sadaghiani, A. Kleinschmidt, Brain Networks and α-Oscillations: Structural and Functional Foundations of Cognitive Control. *Trends Cogn. Sci.* **20**, 805–817 (2016). doi: 10.1016/j.tics.2016.09.004; pmid: 27707588
32. T. J. Buschman, E. L. Denovellis, C. Diogo, D. Bullock, E. K. Miller, Synchronous oscillatory neural ensembles for rules in the prefrontal cortex. *Neuron* **76**, 838–846 (2012). doi: 10.1016/j.neuron.2012.09.029; pmid: 23177967
33. O. Jensen, M. Bonnefond, Prefrontal α- and β-band oscillations are involved in rule selection. *Trends Cogn. Sci.* **17**, 10–12 (2013). doi: 10.1016/j.tics.2012.11.002; pmid: 23176827
34. J. M. Pearce, G. Hall, A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* **87**, 532–552 (1980). doi: 10.1037/0033-295X.87.6.532; pmid: 7443916
35. L. H. Arnal, A. L. Giraud, Cortical oscillations and sensory predictions. *Trends Cogn. Sci.* **16**, 390–398 (2012). doi: 10.1016/j.tics.2012.05.003; pmid: 22682813
36. J. Isnard, M. Guénot, K. Ostrowsky, M. Sindou, F. Mauguière, The role of the insular cortex in temporal lobe epilepsy. *Ann. Neurol.* **48**, 614–623 (2000). doi: 10.1002/1531-8249(200010)48:4<614::AID-ANA8>3.0.CO;2-S; pmid: 11026445
37. R. M. Neal, Slice sampling. *Ann. Stat.* **31**, 705–767 (2003). doi: 10.1214/aos/1056562461
38. A. Delorme, S. Makeig, EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9–21 (2004). doi: 10.1016/j.jneumeth.2003.10.009; pmid: 15102499
39. J.-P. Lachaux, D. Rudrauf, P. Kahane, Intracranial EEG and human brain mapping. *J. Physiol. Paris* **97**, 613–628 (2003). doi: 10.1016/j.jphysparis.2004.01.018; pmid: 15242670
40. P. P. Mitra, B. Pesaran, Analysis of dynamic brain imaging data. *Biophys. J.* **76**, 691–708 (1999). doi: 10.1016/S0006-3495(99)77236-X; pmid: 9929474
41. E. Maris, R. Oostenveld, Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164**, 177–190 (2007). doi: 10.1016/j.jneumeth.2007.03.024; pmid: 17517438
42. K. J. Friston *et al*., Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage* **6**, 218–229 (1997). doi: 10.1006/nimg.1997.0291; pmid: 9344826
43. P. Domenech, S. Rheims, E. Koechlin, PROBE iEEG, version 1.0.0, OpenNeuro (2020); https://openneuro.org/datasets/ds003078/versions/1.0.0.

ACKNOWLEDGMENTS

We thank V. Wyart and J. Drugowitsch for their help. **Funding:** This work was supported by a European Research Council grant (ERC-AdG no. 250106) to E.K. and the Institut Hospitalo-Universitaire CESAME (ANR-10-IBHU-0003). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. **Author contributions:** E.K. and P.D. conceived the study and designed the experimental protocol. P.D. and S.R. collected the data. P.D. and E.K. analyzed the data. E.K. and P.D. wrote the paper. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Intracranial EEG and MRI data are available in the OpenNeuro public repository database (*43*).

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/369/6507/eabb0184/suppl/DC1
Supplementary Text
Figs. S1 to S7
Tables S1 to S3
References (*44*–*49*)

View/request a protocol for this paper from *Bio-protocol*.

23 January 2020; accepted 4 June 2020
10.1126/science.abb0184

Science

Neural mechanisms resolving exploitation-exploration dilemmas in the medial prefrontal cortex

Philippe Domenech, Sylvain Rheims and Etienne Koechlin

Science **369** (6507), eabb0184.
DOI: 10.1126/science.abb0184

To continue or to switch strategy?

Su[redacted]

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.