

# Thalamic regulation of reinforcement learning strategies across prefrontal-striatal networks

---

Received: 3 September 2024

---

Accepted: 3 September 2025





---

Published online: 16 October 2025

---

 Check for updates

---

Bin A. Wang<sup>1</sup>, Mien Brabeeba Wang<sup>2</sup>, Norman H. Lam <sup>3</sup>, Liu Mengxing <sup>3</sup>, Shumei Li<sup>4</sup>, Ralf D. Wimmer<sup>3</sup>, Pedro M. Paz-Alonso<sup>5,6</sup>, Michael M. Halassa <sup>3,7,10</sup> ✉ & Burkhard Pleger <sup>8,9,10</sup>

---

Human decision-making involves model-free and model-based reinforcement learning (RL) strategies, largely implemented by prefrontal-striatal circuits. Combining human brain imaging with neural network modelling in a probabilistic reversal learning task, we identify a unique role for the mediodorsal thalamus (MD) in arbitrating between RL strategies. While both dorsal PFC and the striatum support rule switching, the former does so when subjects predominantly adopt model-based strategy, and the latter model-free. The lateral and medial subdivisions of MD likewise engage these modes, each with distinct PFC connectivity. Notably, prefrontal transthalamic processing increases during the shift from stable rule use to model-based updating, with model-free updates at intermediate levels. Our CogLinks model shows that model-free strategies emerge when prefrontal-thalamic mechanisms for context inference fail, resulting in a slower overwriting of prefrontal strategy representations - a phenomenon we empirically validate with fMRI decoding analysis. These findings reveal how prefrontal transthalamic pathways implement flexible RL-based cognition.

Under conditions of uncertainty, human choice behavior is controlled by habitual and goal-directed systems, which can be well formalized by the algorithmic framework of model-free and model-based reinforcement learning (RL)<sup>1,2</sup>. Model-free strategies prioritize computational efficiency, determining the value of each action and guiding behavior based on reward prediction errors. Conversely, model-based control strategically computes optimal actions by incorporating contextual details, thereby facilitating adaptable, outcome-specific behaviors<sup>3-5</sup>.

Although the two systems have been considered as competitors for behavioral control<sup>1</sup>, evidence indicates their cooperative interactions in response to concurrent cognitive demands<sup>6,7</sup>.

The prefrontal cortico-striatal circuits are considered to be the neural substrates underlying distinct components of the RL process<sup>8</sup>. Within this system, model-free control is most strongly associated with the dorsolateral striatum<sup>9-11</sup>, while the ventromedial (vmPFC) and dorsolateral prefrontal cortex (dlPFC) is essential for model-based

---

<sup>1</sup>School of Psychology, Center for Studies of Psychological Application, Guangdong Key Laboratory of Mental Health and Cognitive Science, Key Laboratory of Brain Cognition and Educational Science, Philosophy and Social Science Laboratory of Reading and Development in Children and Adolescents, South China Normal University, Guangzhou, China. <sup>2</sup>Computer Science & Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>3</sup>Department of Neuroscience, Tufts University School of Medicine, Boston, MA, USA. <sup>4</sup>Department of Medical Imaging, The Affiliated Guangdong Second Provincial General Hospital of Jinan University, Guangzhou, China. <sup>5</sup>BCBL, Basque Center on Cognition, Brain and Language, Donostia-San Sebastián, Spain. <sup>6</sup>Ikerbasque, Basque Foundation for Science, Bilbao, Spain. <sup>7</sup>Department of Psychiatry, Tufts University School of Medicine, Boston, MA, USA. <sup>8</sup>Department of Neurology, BG University Hospital Bergmannsheil, Ruhr-University Bochum, Bochum, Germany. <sup>9</sup>Collaborative Research Centre 874 "Integration and Representation of Sensory Processes", Ruhr-University Bochum, Bochum, Germany. <sup>10</sup>These authors jointly supervised this work: Michael M. Halassa, Burkhard Pleger. ✉ e-mail: [michael.halassa@tufts.edu](mailto:michael.halassa@tufts.edu)

processes as it combines rewards with contextual information<sup>12,13</sup>. Focal brain lesions targeting dlPFC can shift the control of behavior from one RL system to another<sup>4</sup>, indicating that each system has its own distinct representation. Disruptions in the prefrontal-striatal circuits are associated with recurring pathological behaviors, such as those seen in obsessive-compulsive disorders<sup>14</sup>, and with the behavioral rigidity observed in schizophrenia<sup>15</sup>. Despite the recognized importance of RL strategies in adaptive behavior, their precise neural implementation and mechanisms governing their arbitration remain incompletely understood.

Unlike first-order thalamic nuclei that primarily relay peripheral sensory information to the cortex, the mediodorsal thalamus (MD) regulates excitatory/inhibitory balance and effective connectivity within and across frontal areas<sup>16,17</sup>. This regulation is thought to mediate executive functions that underlie adaptive behavior<sup>18–20</sup>. Research in non-human animals has clarified the microcircuit substrates involved in these processes, demonstrating, for example, that flexible switching of attention upon cue changes involves a subset of MD neurons that activate task-relevant prefrontal ensembles, while another MD subset suppresses task-irrelevant ones<sup>21,22</sup>. More recently, experiments have shown that covert rule reversals engage a subset of MD neurons that encode context prediction errors, which switch PFC state underlying strategy updating<sup>23</sup>. In human neuroimaging studies, the MD has been shown to integrate inputs from various prefrontal regions when dynamically selecting between competing behavioral strategies, with more widespread interactions between the MD and large-scale learning networks as task demands increase<sup>24–26</sup>. However, the precise role of the human MD in adaptive behavior generally and RL strategies specifically remains largely unknown.

Here, we address this critical gap by pursuing a serendipitous finding. Specifically, in a human probabilistic reversal learning task, we found robust MD activation upon rule updating as well as the encoding of the new rule. Because subjects learned the initial rule and detected the environmental changes to various degrees, they showed variability in their switching behavior that was well-fit by a model-based RL strategy on some reversals and model-free on others. fMRI revealed that the dorsal prefrontal cortex engaged in the former process while striatum in the latter. Remarkably, the MD engaged in both, but with its lateral subdivision showing model-based activation and the medial division model-free. Causal connectivity analysis showed that transthalamic processing was progressively recruited as the subject transitioned from deploying a stable rule to updating it utilizing a model-based strategy, with model-free updating exhibiting intermediate values. This tantalizing finding was explained by CogLinks, a class of biologically plausible mechanistic models of forebrain networks capable of solving complex tasks<sup>27</sup>. CogLinks uniquely demonstrated that under certain conditions, model-free RL is not instantiated as a distinct algorithm but instead an outcome variation of the same model-based RL algorithm. Specifically, our modeling explains the emergence of a model-free strategy in a bottom-up manner, where prefrontal-thalamic mechanisms of context inference temporarily fail, resulting in the inability to rapidly adjust prefrontal dynamics underlying new contextual learning. This results in the slow overwriting of prefrontal strategy substrates, which our fMRI decoding confirms empirically. Overall, our findings reveal an unexpected role of transthalamic processing in human cognitive flexibility and highlight the value of biologically plausible modeling in showing how complex algorithms are implemented in the human brain.

## Results

### Probabilistic rule reversal task

We leveraged a probabilistic rule reversal task in humans, in which the associations between two tactile stimuli and responses are initially learned and then reversed (Fig. 1a, b). In each trial, one out of the two tactile patterns was applied to the right index fingertip. Participants

had to find out, by trial and error, whether the applied tactile pattern was associated to a “Go”-response, in which participants should press a button with the left index finger, or to “NoGo”, in which they should refrain from pressing the button (Fig. 1a). For one tactile pattern, one response option (e.g., “Go”) had a higher reward probability ( $p = 0.7$ ) than the other (“NoGo”,  $p = 0.3$ ). For the alternative tactile pattern, probabilistic reward associations for “Go” and “NoGo” were reversed (“Go”,  $p = 0.3$ ; “NoGo”,  $p = 0.7$ ). Each block consisted of 45 trials and was divided into two phases: the initial learning phase and the reversal phase. During the initial learning phase, participants learned the stimulus-response association. At a variable time-point, ranging between the 20th to the 25th trial, the probabilistic associations between both tactile patterns and corresponding Go/NoGo responses were reversed, marking the start of the reversal phase (Fig. 1b). In the reversal phase, participants had to switch their decision strategy by reversing cue-response associations. A novel pair of tactile patterns was randomly selected from eight alternative patterns for each new block (Supplementary Fig. 1), which were presented to the participants at the beginning of each block.

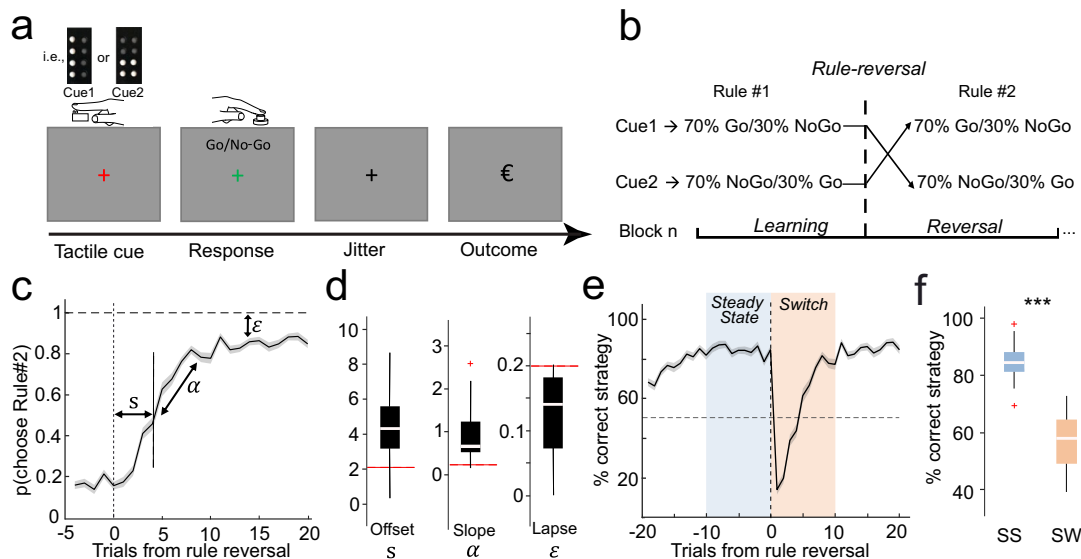
### Behavioral performance

To characterize the dynamics of behavioral adaptation in response to the rule reversals, we first calculated the average probability of selecting the correct rule (or strategy) after the reversals across 12 blocks for each participant (Fig. 1c). A logistic regression model with three parameters was then fitted to individual choices. These parameters represent latent transitions between actions: the switch offset ( $s$ ), slope ( $\alpha$ ), and lapse rate ( $\epsilon$ ). Specifically, the switch offset ( $s$ ) measures the latency of the switch, the slope ( $\alpha$ ) quantifies the sharpness of the transition, and the lapse rate ( $\epsilon$ ) reflects behavioral performance after the transition. For all participants ( $n = 32$ ), we computed the switch offset ( $s$ ), slope ( $\alpha$ ), and lapse rate ( $\epsilon$ ) (Fig. 1d). The mean switch offset ( $s$ ) was approximately 4 (mean  $\pm$  SD =  $4.4 \pm 1.6$ ), indicating that participants reached chance-level performance by approximately the fourth trial following rule reversals. In addition, the slope ( $\alpha$ ) was  $1.3 \pm 2.6$ , and the lapse rate ( $\epsilon$ ) was  $0.13 \pm 0.06$ . These analyses provide detailed insights into the dynamics of behavioral remapping following rule reversals. To benchmark participants' performance, we simulated the behavior of a basic “win-stay-lose-shift” (WSLS) agent, which updates decisions solely based on feedback (“WSLS” dashed lines, Fig. 1d). Compared to the basic WSLS agent, human participants demonstrated significantly longer switch offsets, sharper slopes, and lower lapse rates (all  $p < 0.001$ , Wilcoxon signed-rank test, two-tailed; Fig. 1d).

We then plotted the averaged proportion of correct strategy across participants aligning the reversal phase starting from the rule reversal point (Fig. 1e). During the initial learning phase, the rule about the stimulus-outcome association was quickly learned and was then held in the exploitation state to guide decisions across the next trials (i.e., Steady State, Fig. 1e). Following rule reversal, the performance dropped as participants shifted to the new strategy governed by the new rule (i.e., Switch, Fig. 1e). Based on this group performance and the individual switch offset (ranging from 1 to 9 trials across participants, Fig. 1d), we defined the 10 trials immediately following the reversals as the Switch (SW) period, as this window effectively encompasses both the exploratory phase and the successful transition to the new rule content. The ten trials before the reversals were defined as the Steady State (SS). As expected, the proportion of correct strategy in SS was significantly higher than in SW ( $t_{(31)} = 13.20$ ,  $p = 2.85 \times 10^{-14}$ , two-tailed, Fig. 1f).

### Engagement of striatum, dmPFC and MD following rule reversals

To assess neural activity underlying rule reversals, we first compared Switch to Steady State trials using a General Linear Model (GLM), time-



**Fig. 1 | Rule reversal task and behavioral performance in humans. a** Task schematic of the human rule reversal task. **b** Schematic of a learning block. In each block, 70% of trials presenting one tactile pattern were assigned to “Go,” while 70% of trials presenting the alternative tactile pattern were assigned to “NoGo.” Within each block, the stimulus-response association was reversed randomly between the 20th and 25th trials. **c** Average probability of selecting Rule#2 across all blocks, aligned to the reversal point. Shaded error bars represent SEM. A logistic regression model with three parameters—switch offset ( $s$ ), slope ( $\alpha$ ), and lapse rate ( $\epsilon$ )—was fitted to individual choice data. **d** Box plots of the fitted parameters ( $n = 32$  participants): switch offset ( $s$ ), slope ( $\alpha$ ), and lapse rate ( $\epsilon$ ). The red dashed lines represent the parameter fits for a “win-stay-lose-shift” (WSLS) agent that updates decisions based on feedback. Box plots indicate the median (middle line), 25th, and

75th percentile (box), and the maximum and minimum (whiskers), as well as the outlier (red cross). **e** Group-averaged proportion of correct strategies plotted across trials. The vertical dashed line represents the rule reversal point, while the horizontal dashed line indicates chance-level performance. Shaded error bars represent the standard error of the mean (SEM). The blue-shaded area denotes the Steady State (SS) period, where participants exploited a learned decision strategy, while the orange-shaded area indicates the Switch (SW) period, where participants adjusted their strategy. **f** The proportion of correct strategies in the Steady State was significantly higher than in the Switch period ( $p = 2.85 \times 10^{-14}$ ; paired two-sample  $t$ -tests with two tails,  $n = 32$  participants). \*\*\* $p < 0.001$ . The definition of the median, minima, and maxima for the box plot is the same as in (d). Source data are provided as a Source Data file.

locked to feedback onset. We identified enhanced neural signals ( $p < 0.05$ , whole brain family-wise error (FWE)-corrected, Fig. 2a) immediately after rule reversals in the dorsomedial prefrontal cortex (dmPFC,  $x = 6, y = 26, z = 50, t_{(31)} = 6.56$ ), the dorsal striatum (i.e., caudate nucleus, CN,  $x = 14, y = 8, z = 16, t_{(31)} = 6.99$ ), and the mediodorsal thalamus (MD,  $x = 2, y = -14, z = 10, t_{(31)} = 6.67$ ).

To investigate the causal interactions among these three brain regions (effective connectivity), we employed bilinear dynamic causal modeling (DCM). This method estimates the influence one brain region exerts on another using observed neural data and experimental inputs. The DCM analysis revealed that the model with significant reversal-related modulation of recurrent connectivity between the MD and both the dmPFC and CN best explained the observed evoked activity in these regions (posterior probability = 0.6, Fig. 2b and Supplementary Fig. 2). Furthermore, the modulatory effect of rule reversals on connectivity from the dmPFC to the MD was found to predict the transition slope (Spearman correlation:  $r = -0.484, p = 0.018$ , Fig. 2c). Given the relatively small sample size ( $n = 32$ ), these correlation results should be interpreted with caution. Nonetheless, the findings suggest that a more gradual transition—marked by increased exploratory behavior prior to reaching a stable learned state—may be associated with stronger modulation of dmPFC-to-MD connectivity.

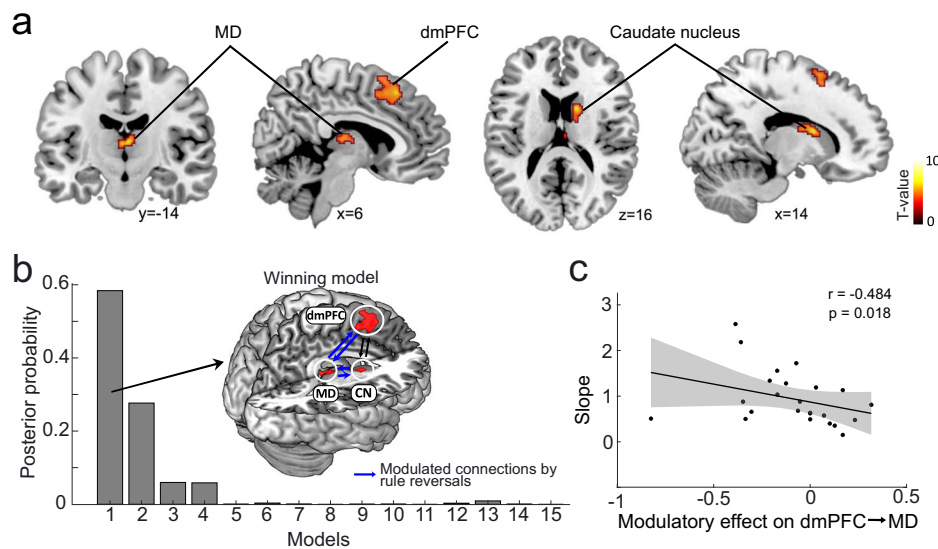
While the human fMRI data demonstrated correlations between neural responses and behavioral/experimental characterizations, they could not establish causality. To directly test the role of the MD in decision strategy updating, we employed a mouse model to examine the relationship between MD function and behavioral adaptation during a rule reversal task (Supplementary Materials and Supplementary Fig. 3a). Only sessions in which animals achieved  $>65\%$  accuracy during the Steady State were included in the analysis. Switch onset was defined as the first trial in which the cue-response mapping was

reversed. Mice successfully updated their decision strategy following rule reversals (Supplementary Fig. 3b) and performed equally well under both rule conditions during the Steady-State ( $n = 17$  sessions from two mice, Mann–Whitney  $U$  test,  $U = 137; p = 0.81$ , Supplementary Fig. 3c). Optogenetic silencing of the MD during the feedback period had no significant effect on mice’s performance during Steady State trials (Supplementary Fig. 3d, e). However, during the feedback period in the first eight trials after a rule switch, MD silencing significantly delayed switching behavior (Mann–Whitney  $U$  test,  $U = 137; p = 0.01$ , Supplementary Fig. 3f, g). These results provide causal evidence that the MD is specifically required during the feedback period to facilitate behavioral adjustments following rule reversals, and refine our understanding of the MD’s function, adding specificity to prior studies in non-human primates that utilized lesion-based methods to demonstrate similar effects<sup>28</sup>.

### The representation of the decision strategy in dmPFC and MD, but not the striatum

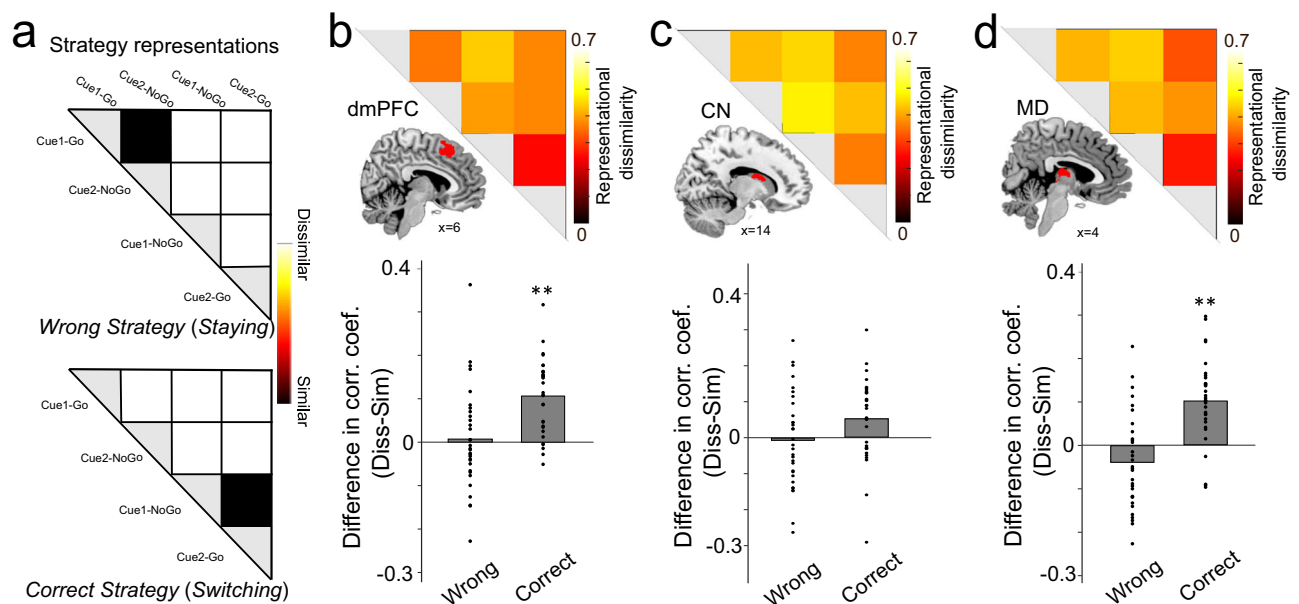
Since participants were unaware of when rule reversals would occur, they updated their strategies reactively. During the Switch period, participants are thought to initially stick to the current strategy immediately after the rule reversal (*Wrong Strategy (Staying)*, i.e., cue1  $\rightarrow$  “Go” and cue2  $\rightarrow$  “NoGo”), resulting in significant decrease of the performance. After they successfully updated their decision strategy (*Correct Strategy or Switching*, i.e., cue1  $\rightarrow$  “NoGo” and cue2  $\rightarrow$  “Go”), performance improved.

Using multi-voxel spatial pattern analysis of neural activity (Representational Similarity Analysis, RSA), we next asked if any of the three regions represented the decision strategy (*Wrong/Correct Strategy*, or *Staying/Switching*) as subjects adapted their behavior (i.e., Switch trials). We hypothesized that brain voxels reflecting the Wrong



**Fig. 2 | Engagement of MD, dmPFC and striatum following rule reversals.** **a** The reversal-related significantly enhanced neural signals in MD, dmPFC and CN (Switch > Steady State, one-sided paired *t*-test,  $p < 0.05$ , whole-brain family-wise error (FWE)-corrected). T-maps are displayed at  $p < 0.001$ , uncorrected, for display purposes only. **b** Modulation of connections by rule reversals. Bayesian model selection revealed that rule reversals significantly modulated the recurrent

connectivity between MD and dmPFC and between MD and CN. **c** The significant correlation between the modulatory effect on the connectivity from the dmPFC to the MD (Spearman correlation:  $r = -0.484$ ,  $p = 0.018$ ) suggests that a slower transition is linked to a stronger modulatory influence on dmPFC-to-MD connectivity. The gray shaded area indicates a 95% confidence level. MD-mediadorsal thalamus, dmPFC-dorsomedial prefrontal cortex, and CN-caudate nucleus.



**Fig. 3 | The representation of the decision strategy.** **a** The RDMs of the two models (i.e., representations of the Correct Strategy (Switching) and Wrong Strategy (Staying)) based on the predicted correlation distance for the four types of trials (cue1→“Go”, cue2→“NoGo”, cue1→“NoGo” and cue2→“Go”). Black elements indicate similarity, white elements indicate dissimilarity between the response patterns of different types of trials. **b** Averaged RDMs of response pattern in dmPFC for all participants ( $n = 32$ ). The response patterns in dmPFC significantly

represented the Correct Strategy, or Switching ( $p = 0.0002$ ). The mean “similar” (black elements in model RDMs) and mean “dissimilar” (white elements in model RDMs) were compared using a one-sided permutation test (“diss”-“sim”). **c** As in (b), but for the CN. The response patterns in CN represented none of the strategies. **d** As in (b), but for the MD. The response patterns in MD significantly represented the Correct Strategy, or Switching ( $p = 0.0009$ ).  $**p < 0.01$ . MD-mediadorsal thalamus, dmPFC-dorsomedial prefrontal cortex, and CN-caudate nucleus.

Strategy would exhibit strong similarities between trials where cue1 → “Go” and cue2 → “NoGo”, whereas the Correct Strategy would show strong similarities between trials where cue1 → “NoGo” and cue2 → “Go”. Based on the predicted correlation distance between different trial types, we constructed the Representational Dissimilarity Matrices

(RDMs) of these two models (Wrong Strategy and Correct Strategy) (Fig. 3a). To assess which decision strategy is represented by each of the three brain regions (dmPFC, CN and MD), we compared the RDMs of each brain ROI with the RDM of each model separately. For Correct Strategy (Switching), we found significant representations in both

dmPFC and MD (one-sided permutation test: dmPFC: effect size = 0.99,  $p = 0.0002$ , Fig. 3b. MD: effect size = 0.82,  $p = 0.0009$ , Fig. 3c), but not in the CN (effect size = 0.44,  $p = 0.047$ , Bonferroni corrected  $p < 0.05/2$ , Fig. 3d). For Wrong Strategy (Staying), none of the three regions exhibited significant representations. The results from the anatomically-defined dmPFC, caudate and MD ROIs confirmed these findings (Supplementary Fig. 4). These multivariate analyses indicate that both the dmPFC and MD play critical roles in the cognitive processes necessary for successful strategy switching, consistent with the findings from the effective connectivity analysis (Fig. 2c).

To further investigate whether the strength of strategy representations in the dmPFC, CN, and MD influences individual adaptive behavior following rule reversals, we performed a correlation analysis. This analysis examined the relationship between the strength of the RDMs for the Correct Strategy (Switching) in dmPFC and MD and the three key behavioral parameters: switch offset ( $s$ ), slope ( $\alpha$ ), and lapse ( $\epsilon$ ). A marginally positive correlation was observed between the strength of the RDM for the Correct Strategy (Switching) in the dmPFC and the switch offset ( $s$ ) (Spearman correlation:  $r = 0.415$ ,  $p = 0.021$ , Bonferroni corrected  $p < 0.05/3$ , Supplementary Fig. 5). This implies that participants with stronger representations (lower dissimilarity) of the Correct Strategy (Switching) in the dmPFC switch more quickly after rule reversals, shedding light on the role of prefrontal signals in facilitating strategy updates by encoding the new strategies.

### The differential involvements of dmPFC, striatum and MD in reinforcement learning strategies

In our human reversal learning task, rule reversals forced the updating of the decision strategy to maximize reward. The underlying process can be understood to largely rest on RL strategies that are either MF or MB. The MF system learns the value of different behaviors solely based on reward prediction error (RPE), while the MB system builds an intrinsic model about state transitions in the decision-making process, taking state transition relationships into consideration (state prediction error, SPE)<sup>29,30</sup>. To explore this in our task, we calculated the distances between human behavioral performance and the behaviors simulated by the MF and MB models. For each participant, we assigned blocks to either the MF or MB category based on which model provided a better fit. First, we compared behavioral performance between the MF and MB blocks. During the SW period, the averaged proportion of correct strategies in MB blocks was significantly higher than in MF blocks (linear mixed-effect test,  $t_{(62)} = 4.961$ ,  $p = 5.77 \times 10^{-6}$ , Fig. 4a). Additionally, logistic regression models were separately fitted to MF and MB blocks to calculate the switch offset ( $s$ ), slope ( $\alpha$ ), and lapse ( $\epsilon$ ) for each block type. The switch offset for MF blocks was significantly longer than for MB blocks (linear mixed-effect test,  $t_{(62)} = 4.155$ ,  $p = 0.0001$ , Fig. 4b). On average, MF blocks required more than five trials (mean  $\pm$  SD =  $5.3 \pm 2.2$ ) to reach chance-level performance following a rule reversal, whereas MB blocks required fewer than four trials (mean  $\pm$  SD =  $3.7 \pm 1.5$ ). This difference suggests that the MF strategy involves more exploratory behavior after environmental changes. However, no significant differences were observed between MF and MB strategies for slope ( $t_{(62)} = -0.478$ ,  $p = 0.635$ ) or lapse ( $t_{(62)} = 1.357$ ,  $p = 0.180$ ), indicating that once the exploratory phase is completed, both strategies are equally capable of adapting to the new environment and maintaining high performance levels.

To dissociate the neural regions associated with MF and MB RL, we characterized participants' RL behavior using the reward prediction error (RPE) and state prediction error (SPE), respectively (Fig. 4c, e). These RPE and SPE signals were then modeled in the GLMs of the fMRI data as parametric modulators, allowing us to investigate how distinct brain regions contribute to the two RL strategies. We found that during Switch, the RPE was significantly correlated with activity in CN ( $x = 16$ ,  $y = 2$ ,  $z = 20$ ,  $t_{(31)} = 4.17$ ,  $p_{\text{FWE-SVC}} = 0.027$ , Fig. 4d) and MD ( $x = -2$ ,  $y = -22$ ,  $z = 10$ ,  $t_{(31)} = 4.55$ ,  $p_{\text{FWE-SVC}} = 0.011$ , Fig. 4d), but not dmPFC. In

contrast, activity in dmPFC ( $x = 14$ ,  $y = 12$ ,  $z = 64$ ,  $t_{(31)} = 6.39$ ,  $p_{\text{FWE-SVC}} < 0.001$ , Fig. 4f) and MD ( $x = 4$ ,  $y = -20$ ,  $z = 12$ ,  $t_{(31)} = 4.05$ ,  $p_{\text{FWE-SVC}} = 0.036$ , Fig. 4f), but not CN, was significantly correlated with the SPE.

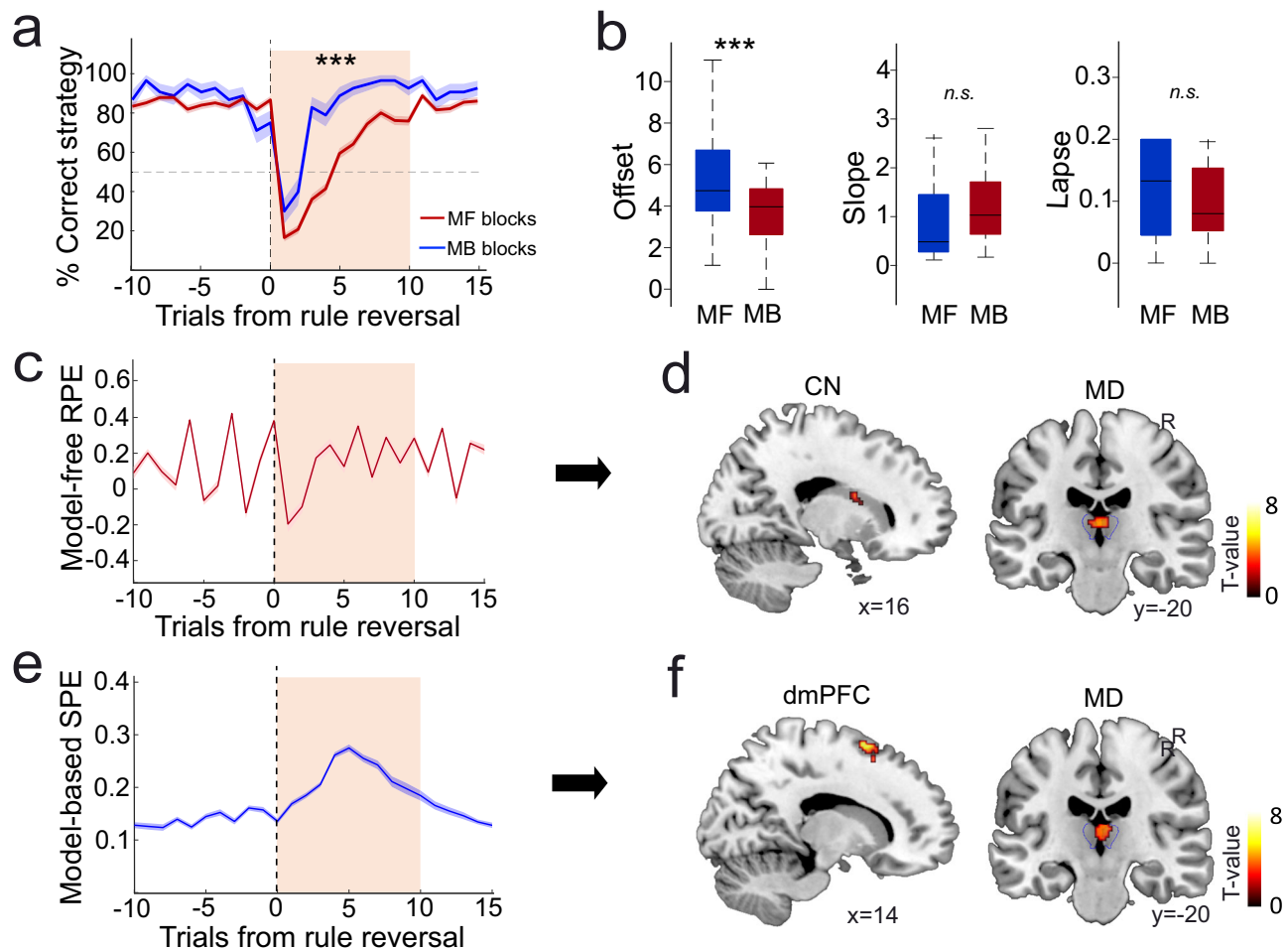
Despite the fact that the MD was involved in both RPE and SPE, we found that the model-based MD engagement was localized to its lateral subdivision, whereas the model-free MD to its medial one (Fig. 5a). Although both MD representations overlapped to a certain degree, the dissimilarity of multivariate MD response patterns encoding model-free RPE and model-based SPE indicated their mutual independence (RSA analysis, Supplementary Fig. 6). These fundamental differences in MD representations were corroborated by their structural tractography and functional connectivity (i.e., psychophysiological interaction, PPI). MD-tractography from an independent human cohort ( $n = 113$ ) indicated, that the MB-related part of the MD dominantly projected to the dorsal PFC, whereas the MF-related MD to ventral PFC (i.e., vmPFC/OFC, Fig. 5b and Supplementary Fig. 7). The PPI findings are consistent with this, showing MB-related MD connectivity with dorsal PFC ( $x = 22$ ,  $y = 22$ ,  $z = 42$ ,  $t_{(31)} = 6.01$ ,  $p_{\text{FWE-SVC}} = 0.036$ ) and MF-related MD connectivity with the OFC ( $x = 46$ ,  $y = 22$ ,  $z = -14$ ,  $t_{(31)} = 6.38$ ,  $p_{\text{FWE-SVC}} = 0.029$ , Fig. 5c and Supplementary Fig. 7).

### Thalamocortical pathways implement the critical computations for reinforcement learning during strategy updating

To explore network-level computations relevant to strategy updating, we applied another DCM that focused on the strategy-related MD subdivisions (medial, MDm and lateral, MDl), along with their cortical counterparts (OFC and dlPFC, respectively). Given the lack of recurrent connectivity in the thalamus, direct connections between the two MD nodes were discounted<sup>17</sup>. Bayesian parameter averaging (BPA) revealed significant connectivity patterns in Steady State (SS) and Switch (SW) trials that we separated into MF and MB updates ( $\text{SW}_{\text{MF}}$ ,  $\text{SW}_{\text{MB}}$ , Fig. 6a). We found a peculiar pattern of effective connectivity changes, which strikingly varied from the SS to  $\text{SW}_{\text{MB}}$ , with the  $\text{SW}_{\text{MF}}$  exhibiting intermediate values. Specifically, cortico-cortical connections (i.e., reciprocal connections between OFC and dlPFC) progressively decreased (one-way Anova,  $F_{(2,299)} = 90.62$ ,  $p = 1.9 \times 10^{-31}$ ), whereas thalamocortical connections (i.e., outputs from two MDs) increased as subjects relied on a MB strategy (one-way Anova,  $F_{(1,299)} = 216.48$ ,  $p = 3.2 \times 10^{-66}$ , Fig. 6b). These findings suggest the role of the MD in influencing multiple prefrontal areas and providing indirect transthalamic communication routes between directly connected cortical areas.

To investigate the circuit mechanisms underlying this network pattern, we employed CogLink modeling, a mechanistic framework of the forebrain network designed to solve contextual decision-making<sup>27</sup> (Fig. 7a). CogLink is distinct in its integration with normative modeling, enabling a principled approach to building a cognitive mechanistic model<sup>27</sup>. The network architecture includes recurrent circuits representing the executive dlPFC, where past actions and outcomes are encoded, and the OFC, where contextual values are stored and updated. Additionally, two MD networks represent the lateral and medial mediodorsal thalamus (MDl and MDm). MDl is responsible for inferring task context and generating update signals, while MDm processes and relays RPEs (see "Methods"; Fig. 7a).

Our model successfully solved the probability reversal task, achieving 80% accuracy before the reversal ( $82.6 \pm 1.7\%$ , Mean  $\pm$  SEM;  $n = 500$ ) and recovering to 80% accuracy before the block ended ( $79.4 \pm 1.8\%$ , Mean  $\pm$  SEM;  $n = 500$ ). MDl neurons displayed strong contextual encoding, with specific populations preferentially activating in response to the new context (Fig. 6b). Notably, this contextual representation was significantly associated with model-based behavior, as higher MDl activity corresponding to the new context correlated with an increased log posterior odd ratio of model-based vs. model-free strategy ( $r = 0.56$ ,  $p = 2.00 \times 10^{-5}$ , Fig. 7c). This relationship

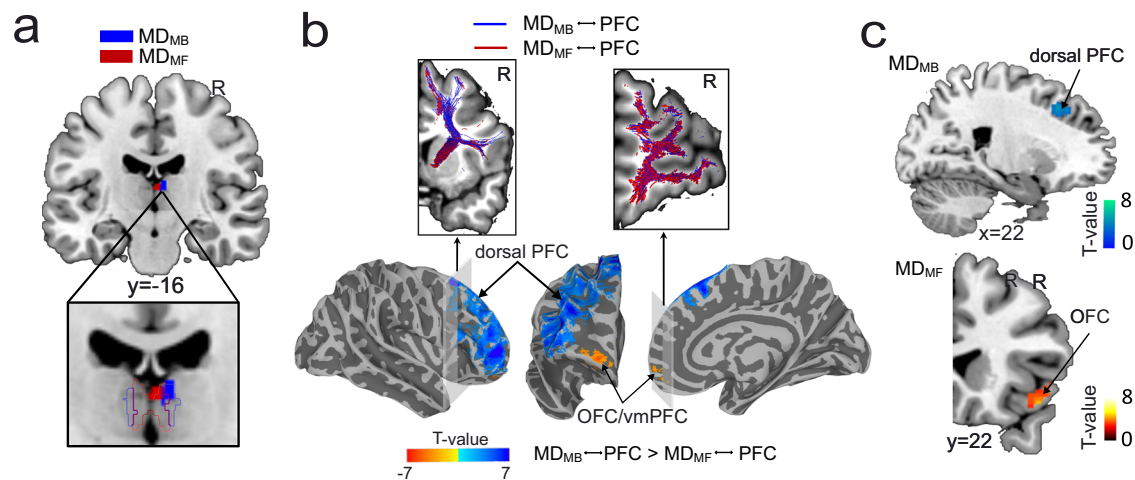


**Fig. 4 | Encoding of model-free (MF) and model-based (MB) algorithms in dmPFC, caudate and MD.** **a** The proportion of correct strategies across trials, separated by blocks better explained by MF-RL or MB-RL. The averaged proportion of correct strategies in MB blocks was significantly higher than in MF blocks (linear mixed-effect test with two tails,  $t_{(62)} = 4.961$ ,  $p = 5.77 \times 10^{-6}$ ). The vertical dashed line marks the rule reversal point, while the horizontal dashed line represents chance-level performance. The shaded error bar indicates the standard error (SEM). **b** Comparison of three parameters—switch offset ( $s$ ), slope ( $\alpha$ ), and lapse ( $\epsilon$ )—between MF and MB blocks. The switch offset for MF blocks was significantly longer than for MB blocks (linear mixed-effect test with two tails,  $t_{(62)} = 4.155$ ,  $p = 0.0001$ ,  $n = 32$  participants), whereas no significant differences were observed for slope ( $\alpha$ ) or lapse ( $\epsilon$ ). Box plots indicate the median (middle line), 25th, and 75th percentile (box), and the maximum and minimum (whiskers). \*\*\* $p < 0.001$ . **c** The group-averaged trajectory of reward prediction errors (RPE) derived from the MF-RL model. The dashed black line indicates the point of the rule reversal, and the orange

shaded area indicates the following *Switch* period. The shaded error bar indicates the standard error (SEM). **d** Both the activity of CN ( $x = 16$ ,  $y = 2$ ,  $z = 20$ , one-sided paired  $t$ -test,  $t_{(31)} = 4.17$ , FWE small-volume correction,  $p = 0.027$ ) and MD ( $x = -2$ ,  $y = -22$ ,  $z = 10$ , one-sided paired  $t$ -test,  $t_{(31)} = 4.55$ , FWE small-volume correction,  $p = 0.011$ ) were significantly correlated with RPE during *Switch*. T-maps are displayed at  $p < 0.001$ , uncorrected, for display purposes only. **e** The group-averaged trajectory of state prediction error (SPE) derived from the MB-RL model. The shaded error bar indicates the standard error (SEM). **f** The activity of both dmPFC ( $x = 14$ ,  $y = 12$ ,  $z = 64$ , one-sided paired  $t$ -test,  $t_{(31)} = 6.39$ , FWE small-volume correction,  $p < 0.001$ ) and MD ( $x = 4$ ,  $y = -20$ ,  $z = 12$ , one-sided paired  $t$ -test,  $t_{(31)} = 4.05$ , FWE small-volume correction,  $p = 0.036$ ) were significantly correlated with SPE in *Switch*. The outline in **(d)** and **(f)** indicates the locations of MD. MD mediadorsal thalamus, dmPFC dorsomedial prefrontal cortex, CN caudate nucleus. Source data are provided as a Source Data file.

aligns with human data, where the activity of MDI is significantly correlated with state prediction error in a model-based RL algorithm (Fig. 4f). To further assess this relationship, we asked whether MDI activity can be used to distinguish between MB and MF strategies that the model may exhibit. Indeed, we categorized task blocks as MB or MF using MDI activity (see “Methods”), and this approach reproduced the behavioral differences observed in human subjects: switch times were longer in MDI activity-derived MF blocks compared to MB blocks (Fig. 7d). Furthermore, and more critically, causal connectivity analysis of the CogLink network replicated findings from DCM of fMRI data. Specifically, steady-state behavior was associated with high cortico-cortical and low thalamocortical connectivity, while model-based switching showed the opposite pattern, and model-free switching exhibited intermediate values (Fig. 7e).

Our CogLink network enabled us to investigate the neural mechanisms underlying MB and MF switching, which were not accessible by fMRI alone. Analysis of MDI neural activity during MB and MF switches revealed that new contextual signals emerged only in MB blocks, whereas MF blocks failed to generate such representations (Fig. 7f). In the CogLink network, distinct contextual populations of MDI neurons selectively modulate disjoint populations in the orbitofrontal cortex (OFC), regulating both their activity and plasticity (see “Methods”; Supplementary Fig. 8a, b)<sup>17</sup>. Consequently, during MF switches, the absence of an alternative contextual representation in MD caused the same OFC population previously engaged in the initial context to remain active (Supplementary Fig. 8a, b). As a result, switching behavior relied on overwriting the existing mapping in the same OFC population. Specifically, in Fig. 7g, we observed a “crossing”



**Fig. 5 | The model-free (MF) and model-based (MB) related MD exhibit distinct anatomical and functional connectivity with PFC.** **a** The more medial MF and the more lateral MB related MD (MD<sub>MF</sub> and MD<sub>MB</sub>). The outline indicates the locations of medial and lateral MDs derived from the AAL3 atlas. **b** Fiber density contrast between white-matter connections of MD<sub>MB</sub>-PFC and MD<sub>MF</sub>-PFC. The contrast map indicates that the MD<sub>MB</sub> has preferential white-matter connection with dorsal PFC while MD<sub>MF</sub> shows preferential white-matter connection with vmPFC/OFC. Contrast map is thresholded at  $p \leq 0.005$  at voxel level and cluster size  $>40$  using a  $t$ -test with two-tails ( $n = 113$  participants). Cold color: MD<sub>MB</sub>-PFC  $>$  MD<sub>MF</sub>-PFC; Warm

color: MD<sub>MB</sub>-PFC  $<$  MD<sub>MF</sub>-PFC. Streamline visualization of the tractography between the MD ROIs and PFC in two representative brain slices is also shown. Blue: MD<sub>MB</sub>-PFC; Red: MD<sub>MF</sub>-PFC. **c** The psychophysiological interaction (PPI) shows the MB-related MD connections with the dorsal PFC and MF-related MD connections with the OFC. FWE small-volume correction was applied across the whole PFC (one-sided paired  $t$ -test, thresholded at  $p < 0.05$ ,  $n = 32$  participants). T-maps are displayed at  $p < 0.001$ , uncorrected, for display purposes only. MD mediiodorsal thalamus, dmPFC dorsomedial prefrontal cortex, CN caudate nucleus.

in the activity of two OFC populations in response to cue 1: the population tuned to “Go”, which initially exhibited high activity, decreased its activity after the switch, while the population tuned to “NoGo”, which initially exhibited low activity, increased its activity.

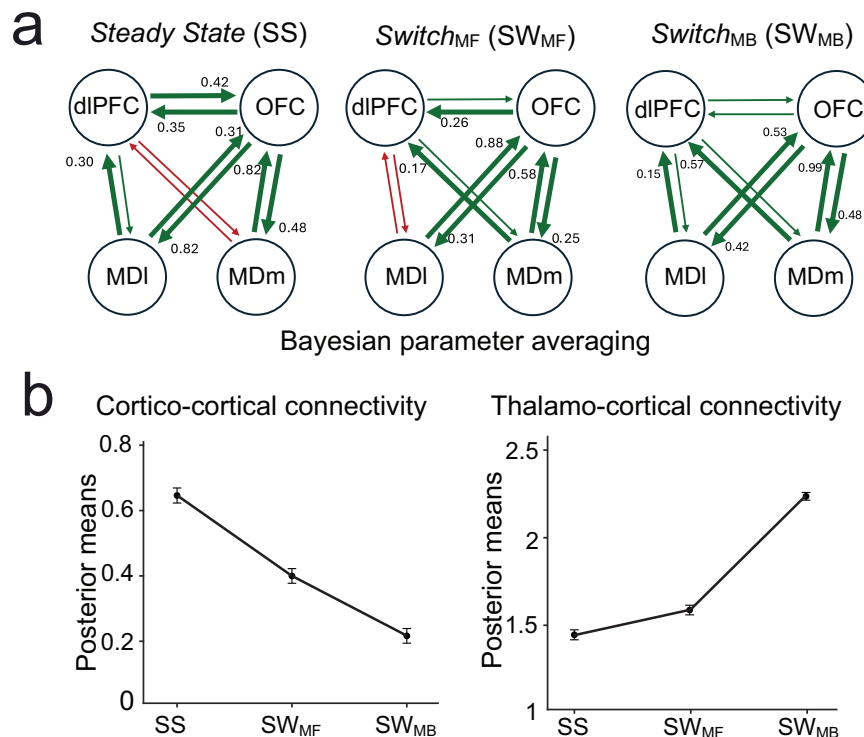
The network modeling results in a critical prediction: model-based switching involves frontal networks simultaneously encoding multiple strategies that can be rapidly toggled, while model-free switching relies on the slower overwriting of the same strategy in frontal networks. To test this prediction, we revisited our original human finding of strategy decoding from frontal networks (Fig. 3). With this insight, we separated switching behavior into model-based and model-free blocks as Fig. 4a (see “Methods”). Strategy decoding using RSA revealed a pattern consistent with the predictions of the CogLink model: MB switching exhibited representational dynamics that could reflect successful toggling between distinct strategy representations, whereas MF switching did not show this pattern. In the dmPFC, for instance, MB behavior was linked to a significantly larger difference between the two strategy representations ( $t_{(31)} = 4.15$ ,  $p = 2.5 \times 10^{-4}$ ) compared to the MF behavior, with similar effect in the dlPFC and OFC ( $t_{(31)} = 2.42$ ,  $p = 0.02$ ;  $t_{(31)} = 2.79$ ,  $p = 0.009$  respectively, Fig. 8h and Supplementary Fig. 9). These findings suggest that, during MB switching, the original (“Staying”) representation is suppressed while the network engages a new (“Switching”) representation, resulting in a pronounced difference in activation patterns. By contrast, MF switching involves overwriting the same representation, leading to less differentiation between old and new strategies. Thus, across frontal networks, MB strategy updates may involve the capacity to maintain and flexibly switch between multiple representations, a capacity that appears limited or absent in MF updates. These results provide a ‘bottom-up’ perspective on the neural substrates of these two algorithmic processes.

These findings show that model-free switching depends on slow modification of an existing strategy. However, if model-free switching relies on overwriting rather than encoding a new representation, what prevents MDI from forming an alternative contextual representation in MF blocks? To understand why MDI fails to form an alternative

contextual representation in MF blocks, we examined the strengths of dlPFC-MDI model synapses. In the CogLink model, these synapses learn the contextual generative model through Hebbian plasticity, enabling accurate contextual inference<sup>17</sup> (see “Methods”). Our analysis revealed that in MB blocks, the generative model learned by these synapses closely approximated the true generative model of the environment. In contrast, in MF blocks, the learned generative model deviated significantly (Supplementary Fig. 8c, d). Specifically, synaptic strengths during SS in MB blocks encoded larger value differences between correct and incorrect strategies compared to MF blocks (Fig. 8i). This failure in MF blocks likely results from the stochasticity of actions and rewards, which can provide conflicting evidence about the current context, resulting in insufficient learning of the environmental generative model before the rule reversal (Fig. 8i and Supplementary Fig. 8c, d). As a result, the model struggles to distinguish new contexts in MF blocks because the generative model from the previous context is not different enough from the post-reversal environment.

These CogLink results lead to a key behavioral prediction. Since larger estimated value differences between correct and incorrect strategies increase the likelihood of selecting the correct strategy, and MB blocks exhibit larger value differences compared to MF blocks (Fig. 8i), we predict that the ratio of correct strategy selection during SS should be higher in MB blocks. In other words, pre-reversal behavioral performance directly influences post-reversal strategy selection. Consistent with this prediction, we observed a significant difference in SS performance between MF and MB blocks in both the CogLink model and human data (Fig. 8j). These findings suggest that MF behaviors emerge as a consequence of incomplete learning of the pre-reversal generative model.

Taken together, these modeling results demonstrate that the model-free strategy does not arise from a distinct algorithm but instead emerges as a variation of model-based mechanisms. Specifically, our findings suggest that model-free behavior results from a temporary failure in prefrontal-thalamic mechanisms of context inference due to insufficient learning of the environmental generative model before the reversal. This bottom-up perspective provides a



**Fig. 6 | Transthalamic connectivity in Steady State (SS), model-free Switch (SW<sub>MF</sub>) and model-based Switch trials (SW<sub>MB</sub>).** **a** Dynamic causal modeling (DCM) analysis on human fMRI data. Bayesian parameter averaging was applied to infer the model parameters of a network consisting of the lateral MD, medial MD, OFC and dlPFC with reciprocal endogenous connectivity among all regions, except for the connectivity between two MDs, for Steady State trials (SS), model-free Switch trials (SW<sub>MF</sub>) and model-based Switch trials (SW<sub>MB</sub>), respectively. **b** The posterior

means of cortico-cortical (connections between OFC and dlPFC) and thalamo-cortical (outputs from two MDs) connections were summed up for SS, SW<sub>MF</sub> and SW<sub>MB</sub> ( $n = 100$  resampling). The error bar indicates the standard error of the mean (SEM) after bootstrap resampling, which was used to assess the stability of posterior means. MD mediiodorsal thalamus, MDm medial MD, MDI lateral MD, OFC orbitofrontal cortex, dlPFC dorsal PFC, MF model-free, MB model-based. Source data are provided as a Source Data file.

mechanistic explanation for how variations in neural substrates shape decision-making flexibility, highlighting that, under certain conditions, model-free RL strategies reflect a temporary limitation rather than a fundamentally separate process.

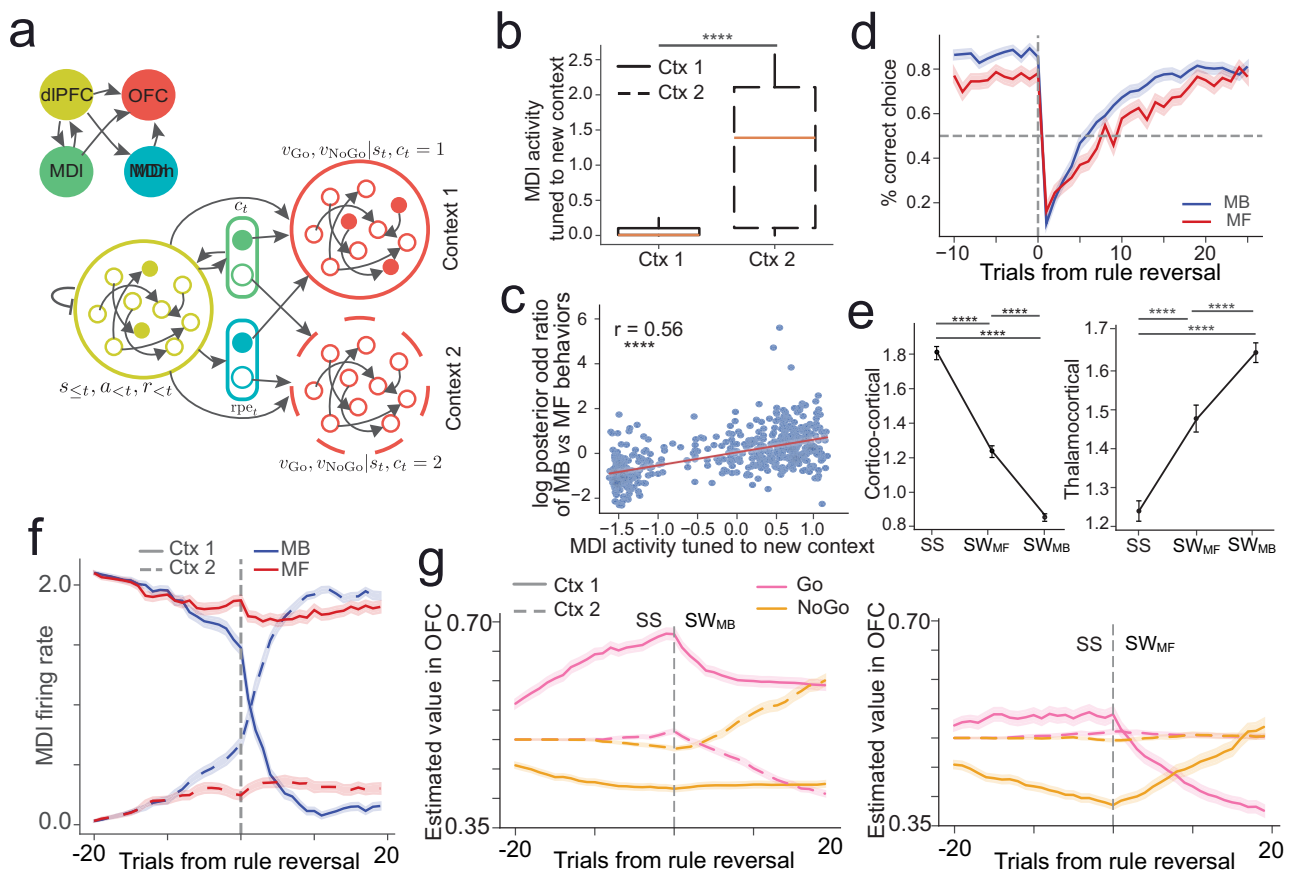
## Discussion

The reinforcement learning (RL) theory formalizes how agents select actions to optimize outcomes<sup>31</sup>. Multiple lines of evidence suggest that humans and other animals behave in a manner that conforms to this theory<sup>2,32,33</sup>, posing the question of how the relevant algorithms are implemented by the brain. Pioneering previous work indicates that the prefrontal cortical-striatal systems are key components of this implementation process<sup>8</sup>. Our study extends these findings by providing neuroimaging evidence that the mediiodorsal thalamus (MD) is a pivotal node in this implementation, particularly when strategy updating is required. Importantly, through iterative integration of biologically plausible circuit modeling and fMRI data analysis, our findings demonstrate how MD thalamic encoding of the task context<sup>34,35</sup> plays a crucial role in mediating the arbitration between different RL strategies.

Multiple studies and modeling approaches indicate that human behavior relies on a combination of model-based and model-free RL systems, with a mechanism that dynamically weights the outputs of each system, effectively modulating the influence of each in determining behavior<sup>5,36,37</sup>. Similarly, mice also utilize a combination of exploratory, model-free strategies and deterministic, inference-based behaviors during block-by-block transitions in a rule reversal learning task<sup>38</sup>, suggesting an evolutionary well-preserved behavioral mechanism. However, the neural operations that govern the arbitration and

balance between these two RL systems remain unclear. Model-free RL control is primarily associated with the dorsolateral striatum and infralimbic cortex<sup>9,11</sup>. In contrast, the model-based RL control is mainly linked to the ventromedial prefrontal cortex (vmPFC) and dorsolateral prefrontal cortex (dlPFC)<sup>12,13,39</sup>. In fact, TMS-based disruption of the right dlPFC in humans diminishes the ability to use a model-based RL strategy<sup>4</sup>. The reliability of both RL systems is encoded by the inferior lateral PFC and middle PFC, while the cingulate cortex encodes the output of a comparison between these reliability signals, implicating these regions in the arbitration process<sup>36</sup>. However, beyond reliability, other factors likely contribute to this process, such as environmental uncertainty and the availability of cognitive resources.

Both model-based and model-free RL strategies typically rely on prediction errors (PEs) to drive learning, which, despite differing in meaning and properties, represent the discrepancy between prior expectations and actual outcomes. PEs serve as the fundamental computation for making inferences about the state of the world, underpinning the formation of beliefs about potential future states and minimizing the need for energy-intensive surprise processing. Model-free RL directly utilizes PEs in the form of reward prediction errors (RPEs), whereas model-based RL indirectly employs PEs by constructing a model of state transitions and outcomes. Actions in model-based RL are evaluated through a forward search of the model, in which state prediction errors (SPEs) play a central role by signaling discrepancies between the current model and observed state transitions. Research using probabilistic Markov decision tasks has revealed distinct neural signatures for RPEs and SPEs. Notably, SPEs correlate with activity in the posterior intraparietal sulcus and lateral prefrontal cortex, implicating these regions in pure state-learning and the



**Fig. 7 | CogLink model reveals the roles of thalamocortical connections for reinforcement learning during strategy switching and underlying circuit computations.** **a** A schematic of the brain areas and circuits modeled in our CogLink model. dIPFC encodes past sensorimotor-outcome associations as well as the current stimulus; MDI encodes the inferred context; MDm encodes the contextual reward prediction errors; OFC encodes the contextual action values given the inferred context and stimulus. We evaluate our CogLink model in the probabilistic reversal task for 500 blocks. **b** The contextual encoding for MDI neurons tuned to the new context ( $p = 1.00 \times 10^{-32}$ ; two-sided rank sum test). Box plots indicate the median (middle line), 25th, and 75th percentile (box), and the maximum and minimum (whiskers), as well as the outlier (red cross). **c** A point plot ( $n = 500$  blocks) of z-score of MDI activity tuned to the new context against the z-score of the log posterior odd ratio of MB vs. MF behaviors ( $p = 2.00 \times 10^{-5}$ ; two-sided permutation test on Pearson's correlation coefficient  $r = 0.56$ ). **d** Summarized

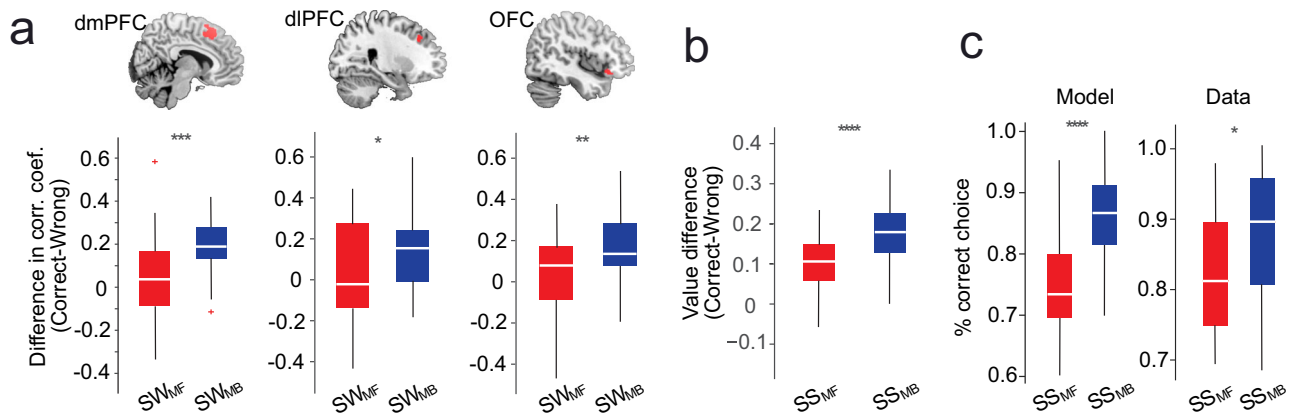
plot (Mean  $\pm$  SEM) for average accurate choice probability at MB ( $n = 308$ ) and MF blocks ( $n = 192$ ) from the model. **e** Summarized plot (mean  $\pm$  SEM,  $n = 500$  blocks) for the effective connectivity of the model at different conditions. The left column shows the effective intercortical connectivity (dIPFC to OFC and OFC to dIPFC) ( $****p < 10^{-4}$ ; Bonferroni-corrected Kruskal–Wallis test with post hoc Dunn's test). The right column shows the effective thalamocortical connectivity (from both lateral and medial parts of MD to both dIPFC and OFC) ( $****p < 10^{-4}$ ; Bonferroni-corrected Kruskal–Wallis test with post hoc Dunn's test). **f** Summarized plot (mean  $\pm$  SEM) of MDI firing rate encoding the contextual signals in MF and MB blocks. **g** Summarized plot (mean  $\pm$  SEM) of estimated action values after presenting with cue 1 in MB and MF blocks for OFC neurons. MD mediadorsal thalamus, MDm medial MD, MDI lateral MD, OFC orbitofrontal cortex, dIPFC dorsolateral PFC, SW Switch State, SS Steady State, MF model-free, MB model-based. Source data are provided as a Source Data file.

construction of a cognitive map of the environment<sup>37</sup>. Conversely, RPEs are associated with the ventral striatum, highlighting its role in reward processing and habitual learning<sup>37</sup>. While RPEs are commonly associated with the ventral striatum, recent findings by Gueguen et al. suggest that the vmPFC and IOFC contribute to encoding reward-related PEs, whereas the insula and dIPFC are more involved in punishment-related PEs, indicating a distinct cortical-subcortical organization for PE processing<sup>40</sup>. Our findings align with these results, further emphasizing the prefrontal cortex's role in model-based RL and the striatum's role in model-free RL strategies. This dichotomy underscores the brain's ability to integrate both learning mechanisms, supporting the existence of two distinct yet highly integrative computational strategies for guiding behavior. The lack of significant RSA results in the striatum can be attributed to its role in representing reward prediction errors, which are not exclusive to specific strategies. This distinction explains why the striatum contributes to the contrast between Switch and Steady State conditions but does not produce significant RSA findings. However, caution is

warranted when interpreting these results, as the categorized blocks do not represent pure model-free or model-based strategies. Instead, they align more closely with behaviors better explained by either of the two models.

Converging evidence suggests that the prefrontal cortico-striatal network is the neural architecture that implements key aspects of RL<sup>8</sup>. Within this circuit, the striatum receives inputs from both cortical and thalamic regions and is densely innervated by midbrain dopamine neurons. Information is relayed back to the cortex through the basal ganglia, which project via the thalamus. As such, the thalamus serves as a critical node within the frontostriatal circuit, facilitating the integration and regulation of RL processes.

The thalamus, centrally located in the forebrain with inputs spanning the entire nervous system and outputs to the cerebral cortex, possesses diverse anatomical connectivity motifs<sup>41–43</sup>. This extensive interconnectivity positions the thalamus to shape various aspects of brain dynamics and contribute to numerous cognitive and behavioral functions<sup>18,26</sup>. The MD's reciprocal connections with all PFC areas



**Fig. 8 | The validations of the model prediction.** **a** The RSA results separated for the MF and MB Switch trials (SW<sub>MF</sub> and SW<sub>MB</sub>) for the PFC regions (dmPFC, dlPFC and OFC) in humans. The differences of the representational dissimilarity between Wrong Strategy and Correct Strategy model (upper left element- bottom right elements in Fig. 3a) were calculated. Less differences were found for the SW<sub>MF</sub> compared to SW<sub>MB</sub> ( $p = 2.5 \times 10^{-4}$ ,  $p = 0.02$ ,  $p = 0.009$  for dmPFC, dlPFC and OFC, respectively, paired two-sample *t*-tests with two tails,  $n = 32$  participants). **b** A box plot of value difference between correct strategy and wrong strategy before reversal (Steady State, SS) for model-based and model-free blocks ( $p = 1.14 \times 10^{-26}$ ,

two-sided rank sum test, MB  $n = 328$  sessions vs. MF  $n = 172$  sessions). **c** The different behaviors between MB and MF blocks before the reversals for both the human data ( $p = 0.018$ ,  $n = 32$  participants) and the model ( $p = 3.4 \times 10^{-7}$ , two-sided rank sum test, MB  $n = [328/6]$  pseudosubjects vs. MF  $n = [172/6]$  pseudosubjects). Box plots indicate the median (middle line), 25th, and 75th percentile (box), and the maximum and minimum (whiskers), as well as the outlier (red cross). SW Switch State, SS Steady State, MF model-free, MB model-based. (\*\*\*\* $p < 10^{-4}$ , \*\*\* $p < 0.001$ , \*\* $p < 0.01$ , and \* $p < 0.05$ ). Source data are provided as a Source Data file.

enable its role in higher cognitive operations. Evidence from animals<sup>20–22,44,45</sup> and humans<sup>25,26,46,47</sup> shows that the PFC relies heavily on MD signals to guide behavioral flexibility. Additionally, the MD's connections with various PFC regions and its excitatory and inhibitory inputs from the frontal lobe, midbrain, and basal ganglia<sup>20,48,49</sup> position it well for coordinating network-wide information processing<sup>18</sup>. Specially, subcortical inputs from the basal ganglia, including ventral pallidum and globus pallidus, can be integrated at multiple levels via corticothalamic and thalamocortical pathways through the MD<sup>19</sup>, which play a critical role in implementing the MD's computational functions.

Collomb-Clerc et al. recently reported that low-frequency oscillations in the MD correlate with outcomes during reward- and punishment-based learning, demonstrating that both components of reward prediction errors were signaled in the human thalamus during an RL task<sup>50</sup>. These findings underscore the role of the thalamus in reinforcement-based decision-making, particularly in processing expected values and outcomes during learning<sup>50</sup>. While our inactivation studies in mice demonstrate that the MD is essential for reversal learning, they do not directly elucidate the mechanism by which the MD governs the switching between strategies. However, when combined with human neuroimaging and computational modeling results, a more comprehensive understanding of the MD's role in mediating RL strategies emerges. First, our neuroimaging results revealed a functional dissociation between the lateral and medial MD within the RL framework. This finding prompted further investigation into whether the model-free and model-based components of the MD exhibit distinct anatomical connectivity patterns with cortical regions, particularly within the PFC. Consistent with findings in primates and rodents<sup>19</sup>, we observed that the medial MD primarily connects with the ventral PFC, while the lateral MD connects with the dorsal PFC.

Emerging studies suggest that the MD can mediate cortico-cortical communication by providing an indirect route, thereby enhancing the flexibility and efficiency of intercortical processing<sup>19,51,52</sup>. High-order thalamic nuclei, such as the MD, act as hierarchical pathways in perceptual decision-making, delivering performance-relevant information from lower-order cortical areas to higher-order cortical areas<sup>53</sup>. These nuclei selectively gate cortico-cortical information

transfer, positioning the thalamus as a critical player in the selection of cognitive processing streams<sup>54</sup>. In alignment with this perspective, the increased reliance on transthalamic processing during strategy updating highlights the essential role of the MD in switching between alternative behaviors<sup>23</sup>. This agrees with our human results and underscores the MD's importance in dynamic decision-making. Importantly, our CogLink modeling validated this transthalamic finding and offered a compelling explanation for the strongest signal being observed during reversals aligned with a model-based RL strategy. In contrast, model-free reversals are interpreted as failures to infer the generative structure of the environment, leading to slower learning of action values and their relay to prefrontal targets. This gave rise to a prediction regarding the overwriting of strategy representations in prefrontal circuits, which was corroborated by fMRI decoding analysis.

Together, our findings extend the framework that the MD sustains and coordinates task-relevant cortico-subcortical representations, enhancing performance in complex cognitive tasks<sup>16,18,20,55</sup>. Although computational modeling and neuroimaging techniques offer valuable insights into thalamic function in humans, further advancements require the application of high-resolution neuroimaging modalities such as 7T MRI and fast BOLD imaging. These technologies facilitate a more precise investigation of thalamic activity and its complex connectivity, including, for instance, its distinct interactions with the dorsomedial and dorsolateral striatum, which have been shown to be differentially involved in RL, respectively<sup>8</sup>. The exceptional sensitivity and resolution of these approaches will enable a deeper exploration of the functional roles of thalamic circuits in RL. This is particularly crucial given that dysfunction in thalamic regulations of brain-wide information processing contributes to behavioral abnormalities observed in conditions such as schizophrenia, autism, and obsessive-compulsive disorder<sup>56–59</sup>.

## Methods

### Dataset for human rule reversal task

**Human participants.** Forty human participants (22 females, mean age  $\pm$  SD:  $24.5 \pm 3.3$  years) were recruited for the human rule reversal task. All participants were right-handed and had normal or corrected-to-normal vision. The participants with a history of psychiatric or

neurological disorders or who were taking regular medication were excluded. The study was approved by the local ethics committee of the Ruhr-University Bochum. All participants provided written informed consent prior to participation. Four participants were excluded due to technical issues with the fMRI scans. Thirty-six participants successfully completed the task during fMRI scanning. Participants who achieved less than 60% correct responses and/or lacked a clear learning and reversal effect in their performance during the fMRI experiment were excluded from further analysis. Based on these criteria, data from four participants were excluded. Consequently, the final analysis included data from 32 participants (16 females; mean age  $\pm$  SD: 24.5  $\pm$  3.5 years).

### Human rule reversal task

**Tactile stimuli.** The tactile stimuli were generated and delivered to the index fingertip of the right (dominant) hand using an MRI-compatible Braille device (Metec, Stuttgart, Germany). The Braille device was controlled using the Presentation software (version 20.1, Neurobehavioral Systems, Berkeley, CA, USA) by TCP-IP commands. The device consisted of eight plastic pins, aligned in two series of four pins (pin diameter 1.2 mm, rounded top, inter-pin spacing 2.45 mm) (Fig. 1a). Eight alternative tactile stimulation patterns were used, which always consisted of four raised and four lowered pins. Participants underwent a tactile detection test prior to task training and fMRI to ensure that all tactile stimulation patterns were correctly perceived. They were asked to report the pattern received until they accurately distinguished all patterns 100%.

**Experimental design.** We employed a probabilistic reversal learning Go/NoGo task as described recently<sup>60</sup>. In each block, two tactile patterns were randomly selected from the eight alternative patterns. A total of 70% of trials with one tactile pattern were assigned to “Go”, and 70% of trials with the alternative tactile pattern were assigned to “NoGo” (Fig. 1b). By trial and error, participants had to learn which of the two available responses (“Go” and “NoGo”) had the higher reward probability for each of the two tactile patterns. In each individual block, the association between tactile stimuli and responses was reversed at a random trial between trial 20 to trial 25, requiring participants to adjust their behavior to gain reward (Fig. 1b). Participants were informed of the probabilistic nature of the association and the existence of a rule switch in each block, but they were not given information about the levels of probability or the specific timing of the reversal.

In each trial, participants first received one out of two tactile stimulation patterns for 500 ms on the index fingertip of the right (dominant) hand. A red fixation cross was simultaneously presented via fMRI-compatible LCD-goggles (Visuastim Digital, Resonance Technology Inc., Northridge, CA, USA). After the tactile cue, the red fixation cross turned green, instructing the participants to press the button (LumiTouch keypads, Photon Control Inc., Burnaby, BC, Canada) with the index finger of the left hand (“Go”), or refrain from pressing the button (“NoGo”). Participants had to press the button within 1000 ms if action was needed. After an interval of 500–1500 ms, the outcomes were presented for 500 ms to indicate whether the action was correct or wrong. Trials were presented with a randomized intertrial interval (ITI) ranging between 1500 and 3000 ms in 100 ms steps.

The task was organized in blocks of 45 trials, and consisted of 3 runs, each included four blocks. A novel pair of tactile patterns were used on each new block, which were presented to the participants at the beginning of each block. Before the fMRI scanning, each participant completed a short practice block with 90% probability to ensure she/he was able to follow the instructions. In total, the fMRI experiment consisted of 540 trials, which we split into three runs, each lasting about 16 min, resulting in a total scanning time of about 50 min.

Within each block of the rule reversal task, the associations between stimuli and responses were reversed at a random trial dividing the block into two phases: (1) the initial learning phase, in which the participants learned the stimulus-response association for each stimulus, and (2) the reversal phase, in which they had to reverse their choice preference to gain reward. To investigate the dynamic changes along the learning process, we aligned the reversal phase using the reversal point and averaged the proportion of correct strategy across blocks. One-sample *t*-test was leveraged to compare the difference in the proportion of correct strategy between Switch and Steady State with a significance threshold of  $p < 0.05$ .

### Analysis of human performance

To investigate behavioral adaptation, we first aligned the reversal phase starting from the rule reversal point and calculated the proportion of correct strategies across blocks for each trial. The averaged proportions of correct strategies were then compared between the Steady State and Switch periods.

To more precisely characterize the dynamics of behavioral adaptation, we calculated the average probability of selecting the correct strategy during the Switch period across all blocks for each participant. A logistic regression model was then fitted to the individual observed choices:

$$a_n = \varepsilon + \frac{1 - 2\varepsilon}{1 + \exp(-\alpha(n - s))} \quad (1)$$

Where  $s$ ,  $\alpha$ , and  $\varepsilon$  are the three free parameters representing the latent transition between actions: The switch offset ( $s$ ) measures the latency of the switch, the slope ( $\alpha$ ) quantifies the sharpness of the transition, and the lapse rate ( $\varepsilon$ ) indicates the behavioral performance after the transition. We calculated the group-averaged values of  $s$ ,  $\alpha$ , and  $\varepsilon$  for all participants.

To benchmark participants' performance and parameters, we simulated the behavior of a “win-stay-lose-shift” (WSLS) agent, which updates decisions based on feedback. As the name suggests, this model repeats the choice if the previous action is rewarded and switches if it is unrewarded. In the two-choice scenario, the probability of choosing option  $k$  is expressed as:

$$p_t^k = \begin{cases} 1 - \varepsilon/2 \text{ if } (c_{t-1} = k \text{ and } r_{t-1} = 1) \text{ or } (c_{t-1} \neq k \text{ and } r_{t-1} = 0) \\ \varepsilon/2 \text{ if } (c_{t-1} \neq k \text{ and } r_{t-1} = 1) \text{ or } (c_{t-1} = k \text{ and } r_{t-1} = 0) \end{cases} \quad (2)$$

where  $c_{t-1}$  is the choice in the previous trial, and  $r_{t-1}$  the outcome (wrong or correct) in the previous trial.

The WSLS agent was simulated across a total of 500 blocks. The average probability of choosing the correct strategy during the Switch period was calculated in the same manner as for human participants. Similarly, the logistic regression model was fitted to the WSLS agent's performance as described previously. The dashed red line indicates the parameter values ( $s$ ,  $\alpha$ , and  $\varepsilon$ ) derived for the WSLS agent.

Decision-making behaviors can be governed by two RL computational models: “model-free” and “model-based” RL strategies. These models share the same core algorithm but differ in complexity and the extent of task knowledge incorporated. Computationally, the “model-free” model assumes minimal information of the task design and maintains an expected probability of reward for each pairing of cue and action. This expected probability is updated per trial based on the reward prediction error (RPE), which is the difference between actual reward and expected reward. Specifically, the model considers a matrix of the value/preference of pairings between cues and actions (Q-matrix, with  $q_{ij}$  as value of action  $j$  given cue  $i$ , with  $i, j = 1$  or  $2$ ). On each trial, given cue  $i$ , the probability of action  $j$  is:  $P_j = e^{\tau q_{ij}} / (e^{\tau q_{i1}} + e^{\tau q_{i2}})$ . Given the chosen action  $j$ , the Q-matrix is updated with a Q-learning rule:  $q_{ij} = q_{ij} + \alpha \text{ RPE}$ , where  $\alpha$  is the learning rate.

A basic assumption is held as task knowledge: the Q-matrix is assumed to have anti-correlated entries (i.e., subjects are assumed to be aware that, for instance,  $q_{11} + q_{12} = 1$ ). In other words,  $q_{kl} = q_{kl} + (-1)^{\text{mod}(k-i+l-j, 2)} \rho \alpha$  RPE for the other three entries  $k, l$ , with  $\rho$  as the degree of anticorrelation. In this model,  $\tau$ ,  $\alpha$ , and  $\rho$  are the fit parameters. The “model-based” model extends from the “model-free” model with the additional knowledge that there are two (loosely defined) sets of stimulus-response association, which the task jumps to and from (rule reversal)—one where cue 1 prefers Go (and cue 2 prefers NoGo) and one where cue 1 prefers NoGo (and cue 2 prefers Go). As such, the model has a confidence-based module that estimates the probability of rule reversal based on errors in recent trials. The method is similar to that detailed in the previous study<sup>61</sup>. Briefly, a belief of rule reversal is computed as the posterior probability of rule reversal given the trial history, divided by that of no rule reversal. The belief is reformulated as a Gaussian distribution, and outputs a state prediction error signal (SPE) as the portion of the distribution above a threshold (arbitrarily set as 1). This module consists of 2 fit parameters (standard deviation of the Gaussian, and a perseverance factor augmenting the Gaussian mean), resulting in 5 fit parameters of the “model-based” model (together with  $\tau$ ,  $\alpha$ , and  $\rho$ ). Therefore, our “model-based” categorization operationalizes sensitivity to task structure and rule reversals. While it may not correspond exactly to canonical model-based planning over an internal transition-reward model, it captures the essential distinction between structure-sensitive (model-based) and reinforcement-driven (model-free) strategies in the task.

Based on the computational “model-free” (MF) and “model-based” (MB) models, we divided the blocks into two categories: those better explained by the MF model and those better explained by the MB model. This categorization was performed by calculating the behavioral distance (% correct strategy) between the observed data and the fitted MF model, then subtracting the distance between the observed data and the fitted MB model. Blocks with lower values were classified as MF, while the remaining blocks were classified as MB. Additionally, the RPEs were computed from the MF model applied to the MF blocks, and the SPEs were derived from the MB model applied to the MB blocks. RPEs and SPEs play distinct roles in the learning processes of MF and MB systems, respectively. SPEs are critical for MB learning, refining the internal model of the environment by identifying discrepancies between predicted and actual state transitions. This process enables iterative updates and improvements to the model. RPEs are essential for MF learning, where action values are updated based on the difference between expected and actual rewards. Unlike SPEs, this process relies on direct experience and trial-and-error without constructing an explicit model of the environment<sup>37</sup>. To dissociate the neural regions associated with these prediction error signals, the RPEs and SPEs were incorporated into generalized linear models (GLMs) of fMRI data as parametric modulators.

To evaluate behavioral performance differences between the MF and MB blocks, we applied a linear mixed-effect model (LMM) to account for the repeated measures within the same participants. This approach accommodates both fixed and random effects, enabling more precise estimation of task differences while accounting for individual variability. The LMM was implemented using the fitlme function in MATLAB, with behavioral performance variables as the response variable and block categories (MF and MB) as the fixed effect. This method ensures robust analysis of the complex data structure and provides reliable inferences regarding differences in behavioral performance between MF and MB strategies.

### Human fMRI data acquisition

The fMRI data were collected on a Philips Achieva 3.0T X-series scanner using a 32-channel head coil. A multi-band echo-planar imaging (EPI) sequence with a multi-band acceleration factor of 2 was used for functional scans. Thirty-eight transaxial slices parallel to the

anterior-posterior commissure (AC-PC) covering the whole brain were acquired with a voxel size of  $2 \times 2 \times 3 \text{ mm}^3$ , TR = 2200 ms, TE = 24 ms, flip angle = 90, field of view = 224 mm, and no interslice gap. For each participant, high-resolution T1-weighted structural images were also acquired, with 176 transversally oriented slices covering the whole brain, to correct for geometric distortions and perform co-registration with the EPIs (isotropic T1 TFE sequence: voxel size:  $1 \times 1 \times 1 \text{ mm}^3$ , field of view  $240 \times 176 \text{ mm}^2$ ).

### FMRI data preprocessing and GLMs

**Preprocessing.** For each run, we acquired a total of 453 EPI volumes. To allow for T1-equilibration, five dummy scans preceded data acquisition in each run, which were removed before further processing. Each participant’s EPI volumes were preprocessed and analyzed with the Statistical Parametric Mapping software SPM12 (Wellcome Department of Imaging Neuroscience, University College London, UK; <http://www.fil.ion.ucl.ac.uk/spm>) implemented in MATLAB R2017b (MathWorks Inc). For preprocessing, images were first applied to slice time correction using sinc interpolation to the middle slice. Then, the T1w image was normalized to the Montreal Neurological Institute (MNI) reference space using the unified segmentation approach. Subsequently, the resulting transformation was applied to the individual EPI volumes to transform the images into standard MNI space and resample them into  $2 \times 2 \times 2 \text{ mm}^3$  voxel space. Spatial smoothing with a 6-mm FWHM Gaussian kernel was applied to the fMRI images only for univariate general linear modeling but not for RSA analyses. Data were high-pass filtered at 1/128 Hz to remove low-frequency signal drifts. For each participant, the preprocessed fMRI data were analyzed in an event-related manner in the following GLMs.

**General linear models (GLMs).** The first univariate GLM, which was used to analyze the univariate BOLD effect elicited by the rule reversals, included two main regressors of interest per block, which accounted for the ten trials in the Steady State before the reversals and the ten trials of the *Switch* after the reversals. The onset was time-locked to the outcome presentation of each trial. For each of these two main regressors, the values of the trial-by-trial variables derived from the model-free (i.e., reward prediction errors, RPEs) and model-based (i.e., state prediction error, SPE) RL were independently defined as the parametric modulator. In the univariate GLMs, four additional regressors of no interest accounting for the unsigned trials before the reversal (the trials not belong to Steady State, time-locked to outcome), the unsigned trials after the reversal (the trials not belong to *Switch*, time-locked to outcome), presentation of the stimuli (all trials collapsed to a single regressor, time-locked to the onset of cue presentation) and invalid trials (i.e., late responses time-locked to outcome) were included. To account for motion-related artifacts during the task, six head-motion parameters estimated during the realignment procedure were also defined as regressors of no interest in the univariate GLM. All regressors were convolved with the canonical hemodynamic response function in an event-related fashion.

Another univariate GLM was used to assess the functional connectivity of model-free and model-based MD with the PFC using PPI. Different with the first univariate GLM, the *Switch* regressor was divided into two regressors: one represented the model-free, and the other the model-based. The Steady State regressor and additional regressors of no interest were identical to the first univariate GLM.

Two multivariate GLMs were also included to assess the representational similarity between different types of trials during *Switch* using RSA. The multivariate GLMs were consisted of the unsmoothed fMRI data. In the first multivariate GLM, different with the first univariate GLM, the *Switch* regressor was divided into four regressors accounted for the four types of trials (cue1-“Go”, cue2-“NoGo”, cue1-“NoGo”, cue2-“Go”) during the *Switch* phase of the task. The onset of events within these regressors was locked to the onset of the outcome

in each trial. The Steady State regressor and additional regressors of no interest were identical with the first univariate GLM. To assess the difference of neural representation between the model-free and model-based trials during *Switch*, another multivariate GLM was created from the first multivariate GLM. In this multivariate GLM, the *Switch* trials were categorized into model-free or model-based regressors based on whether they are from model-free or model-based blocks, with each further divided into four regressors accounted for the four types of trials (cue1-“Go”, cue2-“NoGo”, cue1-“NoGo”, cue2-“Go”). This resulted in eight regressors for the *Switch* trials in total in the second multivariate GLM. The Steady State regressor and additional regressors of no interest were identical with the first multivariate GLM.

**Statistical analyses.** Using the GLM for univariate analysis, we compared Steady State to *Switch* trials (*Switch* > Steady State) to assess whole-brain responses to reversals. The contrast images were next applied to the group-level one-sample *t*-test and thresholded at  $p = 0.05$ , whole-brain family-wise error (FWE)-corrected. Using the univariate GLM, we also analyzed the modulatory effect of model-free and model-based RL variables during *Switch* trials. The respective *t*-contrast images of variable-related responses during *Switch* trials for each participant were applied to the group-level one-sample *t*-test. We hypothesized that the modulatory effects of the model-free and model-based RL parameters correspond to brain responses in the *Switch* trials (contrast: *Switch* > Steady State). Therefore, we performed small volume correction (SVC) by restricting the search volume to a mask encompassing MD, dmPFC and CN obtained from the contrast of “*Switch* > Steady State” at  $p < 0.001$ , uncorrected. The statistical significance at the group level was thresholded at  $p < 0.05$  with a voxel-level FWE small-volume correction.

### Representational similarity analysis (RSA)

To test how the multi-voxel response pattern in MD, dmPFC and CN represents the decision strategy after the reversals, we performed a representational similarity analysis (RSA) by assessing the representational similarity between different types of trials during *Switch* based on the first multivariate GLM. Multi-voxel measures of neural activity are quantitatively related to each other and to computational theory by comparing representational dissimilarity matrices (RDMs).

**Construction of model RDMs.** During the *Switch*, the trials can be categorized into four types (cue1-“Go”, cue2-“NoGo”, cue1-“NoGo”, cue2-“Go”). The participants stuck to the old strategy used before the reversals (Wrong Strategy, or Staying) in trials with the associations of cue1-“Go” and cue2-“NoGo”, while they updated their strategy to the new rule (Correct Strategy, or Switching) in trials with the alternative two associations (cue1-“NoGo”, cue2-“Go”). Based on the predicted correlation distance for these four types of trials, two strategy model RDMs were constructed to investigate whether the multi-voxel spatial patterns of activity in MD, dmPFC and CN at the time of outcome presentation are sensitive either for the Wrong Strategy or Correct Strategy. The Wrong Strategy representation model assumes that the activity in the respective brain region shows a greater representational similarity between trials where cue1-“Go” and cue2-“NoGo”, whereas the Correct Strategy would show stronger similarities between trials where cue1-“NoGo” and cue2-“Go”.

**Construction of ROI RDMs.** Based on the contrast of *Switch* > Steady State applied to the univariate fMRI analysis, we defined three ROIs with the threshold of  $p < 0.001$ , uncorrected: MD, dmPFC and CN, respectively. Using the unsmoothed *t*-statistic maps, activity patterns of the three regions/ROIs were extracted. The relative similarity between the response patterns for the four types of trials was assessed using Pearson correlations and expressed as correlation coefficients.

For each participant, the response patterns were compared among the four types of trials during *Switch*, resulting in a symmetric  $4 \times 4$  matrix.

Another RSA analysis was performed based on the second multivariate GLM to assess the difference of neural representations between the model-free and model-based trials during *Switch* for the PFC regions (dmPFC, dlPFC and OFC). The ROI of dmPFC is consistent with the ROI defined in the first RSA, which was derived from the contrast of *Switch* > Steady State, while the ROIs of dlPFC and OFC were identified through the results of the PPI analysis (see below). These analyses revealed the functional connectivity with model-free and model-based MDs, respectively, using the threshold of  $p < 0.001$ , uncorrected. Different with the first RSA analysis based on all the *Switch* trials, the second RSA analysis was performed based on the model-free and model-based *Switch* trials separately.

**Statistical analyses.** The response patterns in MD, dmPFC and CN during the *Switch* period were analyzed using RSA to investigate whether these regions encoded the decision strategy. To this end, we first constructed RDMs for each ROI. For each model, we compared the mean of the “similar” elements (black elements in the model RDMs) to the mean of the “dissimilar” elements (white elements in the model RDMs) within each ROI separately. Group-level analyses were conducted using a one-sided permutation test across participants. To account for multiple comparisons, a Bonferroni correction was applied, adjusting the significance threshold to  $p < 0.05/2$  to account for the two alternative comparisons (Wrong Strategy and Correct Strategy models). For the second RSA analysis, the differences in correlation coefficients between Wrong and Correct conditions (Staying-Switching) were compared between the model-free and model-based RDMs. A paired two-sample *t*-test was conducted with a significance threshold of  $p < 0.05$ .

### Dynamic causal modeling (DCM)

**Time series extraction and Specification of DCMs.** Based on the first univariate GLM results (*Switch* > Steady State), we performed the first DCM analysis by selecting those brain regions that significantly responded to the reversals (i.e., MD, dmPFC, and CN in the right hemisphere) to investigate the effective connectivity under different model assumptions. Subject-specific time series were extracted from the nearest local maximum within a sphere with a radius of 12 mm centered on the group maximum. The first Eigenvariate was then extracted across all voxels surviving  $p = 0.05$  uncorrected within a 6 mm sphere centered on the individual peak voxel. The resulting BOLD time series were adjusted for effects of no interest (e.g., invalid trials and movement parameters). Five participants had to be excluded from DCM analyses because we could not extract valid activity from one or more of the three ROIs.

The DCMs are specified in terms of fixed (endogenous) connections between brain areas and condition-dependent changes in their connection strength (i.e., modulatory or bilinear effects). In the first DCM, we focused on the connectivity between MD, dmPFC and CN in the right hemisphere according to our GLM results (*Switch* > Steady State). We assumed reciprocal endogenous connectivity among the three regions; that is, all forward connections were accompanied by respective backward connections. We specified three model families to determine whether the input drives: (1) MD; (2) dmPFC; or (3) right CN. For each of these three families, we specified models with different modulatory (bilinear) effects on four connections. This resulted in 15 models for each driving input family (Supplementary Fig. 2). Finally, 45 models were evaluated.

We applied another DCM, based on the second univariate GLM, for the inference of model parameters, namely, whether a specific connection is more likely to exert an excitatory or an inhibitory effect on its target region in a given model. In this DCM, we focused on the strategy-related MD subdivisions (medial and lateral), along with their

cortical counterparts (OFC and dIPFC, respectively). The constructed DCM networks consisted of reciprocal endogenous connections among all regions, except the two MDs, because of the lack of recurrent connectivity in the thalamus, with the inputs given to both MDs. To better separate the signals between lateral and medial MD, we here used a mask from the AAL3 atlas (<https://www.gin.cnrs.fr/en/tools/aal/>) to extract time series from the right lateral and medial MD ROI. The masks of OFC and dorsal PFC were from the PPI analysis (Fig. 5a).

**DCM model comparison.** For the first DCM, we used a two-step fixed-effects Bayesian model selection (BMS)<sup>62</sup> to infer the best fitting model for our observed responses in MD, dmPFC, and CN. First, we used family-level BMS to determine whether models with tactile input to either MD, dmPFC or CN best explained the observed data. Second, the models of the winning family were compared to identify the most plausible model explaining condition-related effects.

**Bayesian parameter averaging.** The second DCM, encompassing lateral MD, medial MD, dorsal PFC and OFC, was applied for the inference on the model parameters during the steady-state period, predominately model-free RL switch period, and predominately model-based RL switch period. The parameters of the given model were then summarized by Bayesian parameter averaging (BPA), which computes a joint posterior density for the entire group by combining the individual posterior densities. A posterior probability criterion of 90% was considered to reflect significant effective connectivity. The posterior means of cortico-cortical (connections between OFC and dPFC) and thalamo-cortical (outputs from two MDs) connections in the network were summed up. The bootstrap resampling (1000 iterations) was applied to assess the reliability of the posterior means of BPA.

### Psychophysiological interaction (PPI) analysis

To assess the functional connectivity of model-free and model-based MDs with PFC, we performed two PPI analyses, one using the model-free MD ( $x = -2$ ,  $y = -22$ ,  $z = 10$ ) and the other using model-based MD ( $x = 4$ ,  $y = -20$ ,  $z = 12$ ) as the seed region. The first Eigenvariate was calculated across all voxels within a 6-mm sphere centered on the peak MNI coordinates of MDs derived from the modulatory effect of model-free and model-based variables. The resulting BOLD time series were adjusted for effects of no interest (e.g., invalid trials and movement parameters) and deconvolved to generate time series required for constructing first-level GLMs for the PPIs. The PPI GLM at the single-subject level contained five regressors: two regressors representing model-free and model-based Switch, and two PPI regressors representing the interactions between the physiological variable (i.e., time series of the seed region) with model-free and model-based Switch. The last regressor represented the physiological variable.

We examined the connectivity of MD with PFC in the model-free and model-based conditions separately. To this end, first-level contrast images were created using the PPI regressor of the interaction between the physiological variable and trials of model-free Switch, or the interaction between the physiological variable and trials of model-based Switch. Next, the first-level contrast images were applied to the group-level one-sample *t*-test. We performed a small volume correction (SVC) by restricting the search volume to the PFC mask with a threshold at FWE-corrected peak-level of  $p < 0.05$ . The PFC mask was defined with the Automated Anatomical Labelling (AAL) atlas.

### Dataset for mouse rule reversal task

**Mice.** Two male C57Bl/6 mice (aged 1.5–5 months) obtained from The Jackson Laboratory were used in this study. Mice were maintained on a 12-h light–dark cycle in a temperature- and humidity-controlled environment (20–22 °C; 40–60% humidity). They were group housed with ad libitum access to food. Prior to behavioral training onset, mice

were single-housed, placed on food regulation and maintained at 85–90% of ad libitum body weight. The animal experiment was carried out according to the guidelines of the US National Institutes of Health and the Institutional Animal Care and Use Committee at the Tufts University School of Medicine.

### Setup and training

Animals were trained in custom-built behavioral boxes. Behavioral boxes were sitting on custom grid floors and were equipped with a floor-mounted initiation port as well as 4 response ports lined up at the front wall, 5 cm apart from each other and ~6 cm away from the initiation port. Access to the response ports was only allowed during the choice period of the task by lifting vertically sliding gates using a servo motor (Tower Hobbies, Champaign, IL). IR barriers (Digikay, Thief River Falls, MN) in each port allowed to detect a nose poke. Auditory cues (high-pass and low-pass filtered broadband noise) were presented through speakers mounted to the left and right of the box setup and controlled through an RX8 I/O processing system (Tucker-Davis Technologies, Alachua, FL). Correct responses were rewarded with 10  $\mu$ l of evaporated milk (Nestle, NY) delivered directly at the response port via a syringe pump (New Era Pump Systems, Farmingdale, NY). Overall trial logic was controlled via an Arduino Mega (Ivrea, Italy) interfacing with custom written MATLAB (MathWorks, Natick, MA) software.

In the first step of training, mice were habituated to the reward and the movement of the sliding doors through a “free reward” schedule, where a reward was delivered in any one of the four reward ports, which remained available until the animal collected the reward. After reward collection, the ports were closed. Following an intertrial interval (ITI) of 15 s, a new reward was made available and access to all ports was granted. A poke into a non-rewarded port had no consequences. This stage was completed once an animal collected 30 rewards. During the next step, white noise signaled the availability of a trial. Animals had to briefly poke into the initiation port to get access to the response ports. Following a poke into the initiation port, a 100 ms high-pass or low-pass filtered sound was played, signaling reward availability in the left or right response port, respectively. At this stage animals were only trained on a set of two response ports (inner or outer) and the cue-side mapping was kept constant. A poke in the wrong response port resulted in immediate closure of all ports and an additional 10 s timeout was added to the ITI. The time that animals were required to keep their snout in the initiation port for a successful trial initiation was gradually increased over training sessions to include the cue presentation. This ensured that the animal’s head position during cue presentation was comparable across trials. This stage of training was completed when animals could perform >70% correct. At this point, animals were trained on the second set of response ports only and with inverted cue-side mapping. Once performance under these conditions also reached >70% correct animals were moved to the final task. At the beginning of a new session, the correct cue-side mapping was instructed during the first 20 trials by only giving access to the corresponding two response ports. After 20 trials, all four ports were opened simultaneously. Cue-side mapping (and with that the rewarded set of response ports, i.e., inner vs. outer) kept reversing over the course of a session after blocks of 40–80 trials.

### Optogenetic manipulations

Mice underwent virus injection and optical fiber implant surgery as previously described<sup>22</sup>. Briefly, following anesthesia with 1–2% isoflurane, the skull was exposed, and small burr holes were drilled bilaterally above the mediodorsal thalamus (A/P: 1.4, M/L:  $\pm 0.6$ ). A total of 300 nl of AAV2-CamkII-eNPHR3.0-eYFP (UNC vector core) were injected 3.0 mm below the brain surface, followed by implantation of 200 nm core optical fibers (Doric Lenses, Quebec, Canada) lowered to 2.8 mm below the brain surface. For behavioral testing, optic fibers on

the head of the mouse were connected via a commutator (Doric Lenses, Quebec) to a 561 nm laser (Duding, Germany) with power output adjusted to 5–7 mW. For all manipulations, MD was inactivated during the first 2 s following feedback (reward or timeout). For Steady State manipulations, the laser was activated randomly on ~ a third of the trials. For Switch manipulations, MD was inactivated during the first 8 trials following a rule switch.

### CogLink modeling

**CogLink Model.** This section details the CogLink model, a slightly modified version of the original CogLink model<sup>27</sup> tailored to the study's scope. The model consists of collections of rate neurons governed by differential equations. The core of the model includes the dlPFC-MDI circuit for contextual inferences, two OFC circuit copies to encode contextual action values, and the MDm circuit to update these values in the OFC. The dlPFC-MDI connections learn the environmental generative model through Hebbian learning, and the dlPFC-MDI circuit accumulates the learned likelihood to infer the current context. MDI then activates different OFC ensembles through interneuron pathways. MDm encodes the reward prediction error from the last trial and projects to the OFC to update PFC-OFC connections for learning contextual action values. For notation, let  $S=2$  be the number of stimuli,  $A=2$  be the number of actions and  $C=2$  be the number of contexts.

At trial  $t$ , the dlPFC consists of two populations, the first population,  $x_s^{pfc_1} \in \mathbb{R}^S$ , encodes the current cue,

$$x_s^{pfc_1} = \begin{cases} 1, & \text{if } s = s_t \\ 0, & \text{otherwise} \end{cases}, \quad (3)$$

while the second population,  $x_{s,a,r}^{pfc_2} \in \mathbb{R}^{S \times A \times 2}$ , encodes the sensorimotor-outcome associations at the last trial:

$$x_{s,a,r}^{pfc_2} = \begin{cases} 1, & \text{if } s = s_{t-1}, a = a_{t-1}, r = r_{t-1} \\ 0, & \text{otherwise} \end{cases}. \quad (4)$$

The MDI activities, denoted as  $x^{mdl} \in \mathbb{R}^C$ , evolve according to the equation:

$$\tau^{eff} \frac{dx^{mdl}}{dt} = -x^{mdl} + f_{mdl}(W^{mdl}x^{mdl}) + I^{pfc/mdl}. \quad (5)$$

Here,  $\tau^{eff} = 5$  represents the effective time constant for accumulation dynamics,  $W^{mdl} \in \mathbb{R}^{C \times C}$  denotes recurrent synaptic weights, given by:

$$W^{mdl} = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \quad (6)$$

The nonlinearity function,  $f_{mdl}: \mathbb{R} \rightarrow \mathbb{R}$ , is defined as:

$$f_{mdl}(x) = \begin{cases} x - 1 & \text{if } -1 \leq x \leq 1 \\ 0 & \text{if } x > 1 \\ -2 & \text{otherwise} \end{cases}, \quad (7)$$

and the PFC-MDI inputs, denoted as  $I^{pfc/mdl} \in \mathbb{R}^C$ , are given by:

$$\forall c \in [C], I_c^{pfc/mdl} = f_{pfc}(W_c^{pfc/mdl} \cdot x_s^{pfc_1}). \quad (8)$$

Here,  $W_c^{pfc/mdl} \in \mathbb{R}^{S \times A \times 2}$  represents PFC-MD projecting to MD neurons tuned to context  $c$ , and  $\cdot$  signifies the tensor inner product. Additionally,  $f_{pfc}(x) = [2.7 + \log(x)]_+$ .

We update the PFC-MDI connections through Hebbian learning as follows:

$$\forall c \in [C], s \in [S], a \in [A], r \in [2], \Delta W_{c,s,a,r}^{pfc/mdl} = \eta_c \left( f_{hebb}(x_c^{mdl}) x_{s,a,r}^{pfc_1} - f_{hebb}(x_c^{mdl}) W_{s,c,a,r}^{pfc/mdl} \right) \quad (9)$$

Here,  $W_{c,s,a,r}^{pfc/mdl}$  represents the synapse  $f_{hebb}: \mathbb{R} \rightarrow \mathbb{R}$  denotes the sigmoidal nonlinearity function,

$$f_{hebb}(x) = \frac{[1 - e^{-4x}]_+}{[1 + e^{-4x}]_+} \quad (10)$$

The learning rate is determined by  $\forall c \in [C], a \in [A], \eta_c \in \mathbb{R}^2$ , given by  $\eta_c = \max\left\{0.1, \frac{f_{hebb}(x_c^{mdl})}{6 + N_c}\right\}$ , where  $N_c$  represents a rolling sum of  $f_{hebb}(x_c^{mdl})$ , and is updated as  $N_c \leftarrow N_c + f_{hebb}(x_c^{mdl})$ .

MD neurons then modulate the downstream OFC ensembles via interneuron-mediated pathways. Specifically, the interneuron activities are defined as:

$$\forall c \in [C], \tau_{vip} x_c^{vip} = -x_c^{vip} + x_c^{mdl} \quad (11)$$

and

$$\forall c \in [C], \tau_{pv} x_c^{pv} = -x_c^{pv} + x_c^{mdl} \quad (12)$$

Here, the interneuron membrane time constant is  $\tau_{vip} = \tau_{pv} = 0.1$ , and  $\bar{c}$  represents the context different from  $c$ . These activities modulate the OFC ensembles,  $x^{ofc} \in \mathbb{R}^{C \times A}$ , as follows:

$$\forall c \in [C], s \in [S], a \in [A], \tau^{ofc} \frac{dx_{c,a}^{ofc}}{dt} = -x_{c,a}^{ofc} + f(x_s^{pfc_2}) \circ f_{in}(x_c^{vip} - x_c^{pv}) W_{c,s,a}^{pfc/ofc}. \quad (13)$$

where

$$f_{in}(x) = \frac{[1 - e^{-2x}]_+}{[1 + e^{-2x}]_+} \quad (14)$$

and  $W^{pfc/ofc} \in \mathbb{R}^{C \times S \times A}$  denotes PFC-OFC synapses.

These interneuron-mediated pathways also modulate plasticity. Specifically, given the activity of MDm,  $x^{mdm} \in \mathbb{R}^C$ ,

$$\forall c \in [C], x_c^{mdm} = r_{t-1} - W_{c,s_{t-1},a_{t-1}}^{pfc/ofc} \quad (15)$$

the interneurons-mediate pathways change the PFC-OFC plasticity as follows:

$$\forall c \in [C], \Delta W_{c,s_{t-1},a_{t-1}}^{pfc/ofc} = \eta_{c,s_{t-1},a_{t-1}} f_{in}(x_c^{vip} - x_c^{pv}) x^{mdm}. \quad (16)$$

where  $\eta_{c,s,a} = \left\{0.2, \frac{1}{4 + N_{c,s,a}}\right\}$  and  $N_{c,s,a}$  is the rolling sum of  $1_{s=s_t, a=a_t} f_{in}(x_c^{vip} - x_c^{pv})$ , with  $N_{a,c} \leftarrow N_{a,c} + 1_{s=s_t, a=a_t} f_{in}(x_c^{vip} - x_c^{pv})$ .

To do action selection, we aggregate OFC action values onto motor neurons,  $x^{mct} \in \mathbb{R}^A$ , as follows:

$$\forall a \in [A], I_a^{mct} = \frac{1}{3} \sum_{c \in [C]} x_{c,a}^{ofc} \quad (17)$$

and

$$\forall a \in [A], \tau^{mct} \frac{dx_a^{mct}}{dt} = -x_a^{mct} + g \left( \sum_{i=1}^A W_{at}^{mct} x_i^{mct} \right) + I_a^{mct}. \quad (18)$$

Here, the membrane time constant  $\tau^{mct} = 1$ . The recurrent synaptic weights,  $W^{mct} \in \mathbb{R}^{A \times A}$ , are defined as:

$$W^{mct} = \begin{bmatrix} 1 & -1 & \dots & -1 \\ -1 & 1 & \dots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \dots & 1 \end{bmatrix}. \quad (19)$$

The nonlinearity function  $g: \mathbb{R} \rightarrow \mathbb{R}$  is defined as:

$$g(x) = \begin{cases} 1 & \text{if } x > 1 \\ x & \text{if } 1 \geq x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (20)$$

The action  $a$  is chosen as  $a_t$  if either  $x_a^{mct}$  reaches the threshold = 0.8 within 5 s after the trial starts or chosen stochastically from a softmax distribution,  $a_t \sim \text{softmax}(25x^{mct})$ .

The simulation is conducted by discretizing the differential equation using  $dt = 0.005$ .

### Connection to normative models

One of the unique aspects of CogLink modeling is its tight connection to normative modeling. By optimizing parameters in a normative model and then translating such parameters into a mechanistic model, one is able to effectively build a mechanistic model capable of solving complex cognitive tasks<sup>27</sup>. In this section, we will state such connections from the original CogLink paper<sup>27</sup>:

Let  $dt = \tau^{eff}$ . Let  $X = x_2^{mdl} - x_1^{mdl}$ . If we set  $X_0 = -2$ ,  $S_t = X_t + D$  and assume  $|p_1^{pfc/mdl} - p_2^{pfc/mdl}| \ll 1$ , the evolution of  $X$  approximates to

$$S_t = \min\left(4, \max\left(0, S_{t-1} + p_1^{pfc/mdl} - p_2^{pfc/mdl}\right)\right). \quad (21)$$

Notably, when  $p_1^{pfc/mdl}(t) = \log P(s_t, a_t, r_t | c=1) + \alpha$ ,  $p_2^{pfc/mdl}(t) = \log P(s_t, a_t, r_t | c=2) + \alpha$  for any  $\alpha > 0$  and  $S_n < 4$ , this corresponds exactly to the CUSUM algorithm:

$$S_t = \max\left(0, S_{t-1} + \log P(s_t, a_t, r_t | c=1) - \log P(s_t, a_t, r_t | c=2)\right). \quad (22)$$

In particular, CUSUM algorithm is a theoretically optimal algorithm to detect environmental changes<sup>63,64</sup>. This provides a mathematical basis on model's flexibility on adaptation.

### Model-based versus model-free splitting

To plot Fig. 6c, we construct a model-free and a model-based algorithm and calculate the log posterior odd ratio of the two models given the actions of CogLink model. For the model-free model, we have the action value estimates  $V \in \mathbb{R}^{S \times A}$  and we update the estimates as follows:

$$\Delta V_{s_t, a_t} = \eta_{s_t, a_t} (r_t - V_{s_t, a_t}). \quad (23)$$

Here,  $\eta_{s, a} = \max\left\{0.2, \frac{1}{4 + N_{s, a}}\right\}$ , where  $N_{s, a}$  represents a rolling sum of  $\mathbf{1}_{s=s_t, a=a_t}$ , and is updated as  $N_{s, a} \leftarrow N_{s, a} + \mathbf{1}_{s=s_t, a=a_t}$ . The action distribution at trial  $t$  is parametrized by  $\text{softmax}(25V_{s_t, -})$ .

For the model-based model, we consider a fixed hidden Markov model (HMM) with emit probability specified by the task parameter and the transition probability specified by the mean probability of switching, that is 1/25. One can then infer the most likely context state  $c_t$  at trial  $t$  in such HMM and we specify the action distribution at trial  $t$  to be  $\text{softmax}\left(25 \sum_{c \in [C]} c_t, c V_{c, s_t, -}\right)$ . Here,  $V_{c, s, a}$  denote the probability of receiving a reward at context 1 given cue  $s$  and action  $a$ .

After observing in Fig. 6c that the MDI activity at the alternative context  $x_2^{mdl}$  is highly correlated with the log posterior odds ratio comparing model-based to model-free strategies, we categorized the blocks as follows: blocks with a positive z-score of  $x_2^{mdl}$  were classified as model-based (MB) blocks ( $n = 308$ ), while blocks with a negative z-score of  $x_{mdl} 2$  were categorized as model-free (MF) blocks ( $n = 192$ ).

### Causal effective connectivity analysis

Similar to the DCM analysis in human data, we choose the data 10 trials before the switch as the Steady State condition ( $n = 500$ ), 10 trials after the switch in the model-based blocks as the model-based condition ( $n = 308$ ) and 10 trials after the switch in the model-free blocks as the model-free condition ( $n = 192$ ). Assume the underlying effective dynamic is  $\dot{x} = Wx$  and the data is  $\hat{x}$ . Then we find the effective connectivity as  $W_{eff} = \text{argmin} \int_0^T |Wx(t) - \hat{x}(t)| dt + |W|_2^2$ . By discretizing the integral, this can be approximated as ridge regression. Since each brain area has multiple neurons with various activity patterns, we cannot sum up all possible connections to represent the effective connectivity. To calculate the effective connectivity from one region,  $x_{pre}$ , to another,  $x_{post}$ , we calculate the projection of effective inputs onto postsynaptic activity's direction as follows:

$$W_{eff} x_{pre} \cdot x_{post} / |x_{post}|. \quad (24)$$

For the cortical connectivity, we sum up effective connectivity of both OFC to dIPFC and dIPFC to OFC and for the thalamocortical connectivity, we sum up effective connectivity from both MDI and MDm to OFC and dIPFC.

### Dataset for tractography

**Human participants.** A separate dataset of 113 healthy human participants (mean  $\pm$  SD = 24.5  $\pm$  4.33 years; 65 females) was used for the tractography analyses. All participants were right-handed and had normal or corrected-to-normal vision. No participant had a history of major medical, neurologic, or psychiatric disorders. The study protocol was approved by the Ethics Committee of the Basque Center on Cognition, Brain and Language and was conducted in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki) for experiments involving human participants. Before their inclusion in the study, all participants provided informed written consent. Participants received monetary compensation for their participation.

### Data acquisition

Whole-brain MRI data acquisition was conducted on a 3 T whole-body MRI scanner (Prisma Fit, Siemens Medical Solutions) using a 64-channel whole-head coil. The MRI acquisitions included one T1-weighted (T1w) structural image and diffusion-weighted imaging (DWI) sequences. High-resolution MPRAGE T1-weighted structural images were collected with the following parameters: TR = 2530 ms; TE = 2.36 ms; flip angle = 7°; field of view = 256 mm; voxel size = 1 mm isotropic; 176 slices. In total, 100 diffusion-weighted images were acquired with an anterior-to-posterior phase-encoding direction and 50 isotropically distributed diffusion-encoding gradient directions. The 100 diffusion-weighted images included 50 images with a  $b$ -value of 1000 s/mm<sup>2</sup> and 50 images with a  $b$ -value of 2000 s/mm<sup>2</sup>. Twelve images with no diffusion weighting ( $b$ -value = 0 s/mm<sup>2</sup>) were obtained for motion correction and geometrical distortion correction, which comprised five images with the same phase-encoding direction as the DWI images and seven images with a reversed-phase encoding direction (posterior to anterior). Both DWI and b0 images shared the following parameters: TR = 3600 ms; TE = 73 ms; flip angle = 78°; voxel size = 2 isotropic; 72 slices with no gap and a multiband acceleration factor of 3.

## Tractography

The white-matter pathways between MD seeds and PFC were estimated using the Reproducible Tract Profiles 2 (RTP2) pipeline<sup>65</sup>.

The T1w image was used to acquire ROIs and to register the DWI to individual space. The seeds MD<sub>MB</sub> and MD<sub>MF</sub> were transformed from MNI space to individual space using a nonlinear transformation in Advanced Normalization Tools (ANTs; <http://stnava.github.io/ANTs>). The PFC was acquired by running recon-all from freesurfer (<http://surfer.nmr.mgh.harvard.edu/>).

The DWI data were preprocessed using MRtrix<sup>66</sup> functions in the following steps: (1) data denoising based on random matrix theory, which exploits data redundancy in the patch-level principal component analysis domain<sup>67</sup> using dwidenoise; (2) Gibbs Ringing correction<sup>68</sup> using mrdegibbs; (3) susceptibility-induced distortions and motion correction with the FSL topup and eddy tools<sup>69</sup> called by dwifslpreproc; (4) B1 field inhomogeneity correction with dwibiascorrect and Rician background noise removal with mrcalc and lastly (5) a rigid transformation matrix to align the DWI images to the corresponding T1w image using ANTs.

White-matter pathways were reconstructed from the preprocessed DWI data. We first modeled the diffusion information at the voxel level to obtain a map of preferred directions with fiber orientation distributions (FODs). For this modeling, we used the MRtrix3 CSD algorithm<sup>70</sup>, as it can discern crossing fibers and provide more than one direction in each voxel. Next, streamline tracking was performed separately for the two pathways: MD<sub>MB</sub>-PFC pathways and MD<sub>MF</sub>-PFC pathways. The tracking is performed on the estimated FODs using the MRtrix iFOD2 algorithm<sup>71</sup> with the following parameters: step size, 1 mm; FODs amplitude threshold, 0.05; angle threshold, 45°; maximum length, 200 mm; minimum length, 20 mm.

## Statistical analysis

To compare the terminal distributions of the two pathways on the PFC, we examined the difference in streamline density between them. Streamline density maps were generated for each individual using tckmap from MRtrix and transformed to MNI space for group analysis. A paired *t*-test was conducted to compare the streamline density between the MD<sub>MB</sub>-PFC and MD<sub>MF</sub>-PFC pathways using 3dttest++ in AFNI<sup>72</sup>.

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The raw behaviors, processed fMRI data and group-level fMRI data have been deposited on the Open Science Framework (OSF) (<https://osf.io/6n7db/>). The raw imaging data are not publicly available due to restrictions related to the individual information that could compromise the privacy of research participants. The data used to generate the CogLink results were deposited in Mendeley with <https://doi.org/10.17632/2jbbvcdwy7.1>. Source data are provided with this paper.

## Code availability

The codes for the human data analysis have been deposited on the GitHub ([https://github.com/Bin-A-Wang2/ReversalLearning\\_Thalamic\\_regulations](https://github.com/Bin-A-Wang2/ReversalLearning_Thalamic_regulations)) and the Zenodo (<https://doi.org/10.5281/zenodo.16942458>)<sup>73</sup> repositories. The source code for CogLink Modeling have been also deposited on the GitHub (<https://github.com/brabeeba/thalamic-regulation-of-RL-strategies>).

## References

- Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
- Dayan, P. & Niv, Y. Reinforcement learning: the good, the bad and the ugly. *Curr. Opin. Neurobiol.* **18**, 185–196 (2008).
- Dayan, P. Goal-directed control and its antipodes. *Neural Netw.* **22**, 213–219 (2009).
- Smittenaar, P., FitzGerald, T. H. B., Romei, V., Wright, N. D. & Dolan, R. J. Disruption of dorsolateral prefrontal cortex decreases model-based in favor of model-free control in humans. *Neuron* **80**, 914–919 (2013).
- Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
- Gershman, S. J., Markman, A. B. & Otto, A. R. Retrospective reevaluation in sequential decision making: a tale of two systems. *J. Exp. Psychol. Gen.* **143**, 182–194 (2014).
- Otto, A. R., Gershman, S. J., Markman, A. B. & Daw, N. D. The curse of planning: dissecting multiple reinforcement-learning systems by taxing the central executive. *Psychol. Sci.* **24**, 751–761 (2013).
- Averbeck, B. & O'Doherty, J. P. Reinforcement-learning in fronto-striatal circuits. *Neuropsychopharmacology* **47**, 147–162 (2022).
- Yin, H. H., Knowlton, B. J. & Balleine, B. W. Lesions of dorsolateral striatum preserve outcome expectancy but disrupt habit formation in instrumental learning. *Eur. J. Neurosci.* **19**, 181–189 (2004).
- Wunderlich, K., Smittenaar, P. & Dolan, R. J. Dopamine enhances model-based over model-free choice behavior. *Neuron* **75**, 418–424 (2012).
- Balleine, B. W. & O'Doherty, J. P. Human and rodent homologues in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* **35**, 48–69 (2010).
- Alexander, W. H. & Brown, J. W. Medial prefrontal cortex as an action-outcome predictor. *Nat. Neurosci.* **14**, 1338–1344 (2011).
- de Wit, S., Corlett, P. R., Aitken, M. R., Dickinson, A. & Fletcher, P. C. Differential engagement of the ventromedial prefrontal cortex by goal-directed and habitual behavior toward food pictures in humans. *J. Neurosci.* **29**, 11330–11338 (2009).
- Akkermans, S. E. A. et al. Frontostriatal functional connectivity correlates with repetitive behaviour across autism spectrum disorder and obsessive-compulsive disorder. *Psychol. Med.* **49**, 2247–2255 (2019).
- Waltz, J. A. The neural underpinnings of cognitive flexibility and their disruption in psychotic illness. *Neuroscience* **345**, 203–217 (2017).
- Rikhye, R. V., Wimmer, R. D. & Halassa, M. M. Toward an integrative theory of thalamic function. *Annu. Rev. Neurosci.* **41**, 163–183 (2018).
- Halassa, M. M. & Sherman, S. M. Thalamocortical circuit motifs: a general framework. *Neuron* **103**, 762–770 (2019).
- Shine, J. M., Lewis, L. D., Garrett, D. D. & Hwang, K. The impact of the human thalamus on brain-wide information processing. *Nat. Rev. Neurosci.* **24**, 416–430 (2023).
- Wolff, M. & Halassa, M. M. The mediodorsal thalamus in executive control. *Neuron* **112**, 893–908 (2024).
- Halassa, M. M. & Kastner, S. Thalamic functions in distributed cognitive control. *Nat. Neurosci.* **20**, 1669–1679 (2017).
- Rikhye, R. V., Gilra, A. & Halassa, M. M. Thalamic regulation of switching between cortical representations enables cognitive flexibility. *Nat. Neurosci.* **21**, 1753–1763 (2018).
- Schmitt, L. I. et al. Thalamic amplification of cortical connectivity sustains attentional control. *Nature* **545**, 219–223 (2017).
- Lam, N. H. et al. Prefrontal transthalamic uncertainty processing drives flexible switching. *Nature* <https://doi.org/10.1038/s41586-024-08180-8> (2024).
- Kosciessa, J. Q., Lindenberger, U. & Garrett, D. D. Thalamocortical excitability modulation guides human perception under uncertainty. *Nat. Commun.* **12**, 2430 (2021).

25. Hummos, A., Wang, B. A., Drammis, S., Halassa, M. M. & Pleger, B. Thalamic regulation of frontal interactions in human cognitive flexibility. *PLoS Comput. Biol.* **18**, e1010500 (2022).
26. Hwang, K., Bertolero, M. A., Liu, W. B. & D'Esposito, M. The human thalamus is an integrative hub for functional brain networks. *J. Neurosci.* **37**, 5594–5607 (2017).
27. Wang, M. B., Lynch, N. & Halassa, M. M. The neural basis for uncertainty processing in hierarchical decision making. *Nat. Commun.* <https://doi.org/10.1038/s41467-025-63994-y> (2025).
28. Chakraborty, S., Kolling, N., Walton, M. E. & Mitchell, A. S. Critical role for the mediodorsal thalamus in permitting rapid reward-guided updating in stochastic reward environments. *Elife* **5**, e13588 (2016).
29. Daw, N. D. Are we of two minds? *Nat. Neurosci.* **21**, 1497–1499 (2018).
30. Dolan, R. J. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).
31. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, 1998).
32. Gershman, S. J. & Daw, N. D. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annu. Rev. Psychol.* **68**, 101–128 (2017).
33. Muller, T. H. et al. Distributional reinforcement learning in prefrontal cortex. *Nat. Neurosci.* **27**, 403–408 (2024).
34. Scott, D. N., Mukherjee, A., Nassar, M. R. & Halassa, M. M. Thalamocortical architectures for flexible cognition and efficient learning. *Trends Cogn. Sci.* **28**, 739–756 (2024).
35. Zheng, W.-L., Wu, Z., Hummos, A., Yang, G. R. & Halassa, M. M. Rapid context inference in a thalamocortical model using recurrent neural networks. *Nat. Commun.* **15**, 8275 (2024).
36. Lee, S. W., Shimojo, S. & O'Doherty, J. P. Neural computations underlying arbitration between model-based and model-free learning. *Neuron* **81**, 687–699 (2014).
37. Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).
38. Le, N. M. et al. Mixtures of strategies underlie rodent behavior during reversal learning. *PLoS Comput. Biol.* **19**, e1011430 (2023).
39. Xue, G., Juan, C.-H., Chang, C.-F., Lu, Z.-L. & Dong, Q. Lateral prefrontal cortex contributes to maladaptive decisions. *Proc. Natl. Acad. Sci. USA* **109**, 4401–4406 (2012).
40. Gueguen, M. C. M. et al. Anatomical dissociation of intracerebral signals for reward and punishment prediction errors in humans. *Nat. Commun.* **12**, 3344 (2021).
41. Jones, E. G. The thalamic matrix and thalamocortical synchrony. *Trends Neurosci.* **24**, 595–601 (2001).
42. Halassa, M. M. *The Thalamus* (Cambridge University Press, 2022).
43. Mukherjee, A. et al. Variation of connectivity across exemplar sensory and associative thalamocortical loops in the mouse. *Elife* **9**, e25554 (2020).
44. Mukherjee, A., Lam, N. H., Wimmer, R. D. & Halassa, M. M. Thalamic circuits for independent control of prefrontal signal and noise. *Nature* **600**, 100–104 (2021).
45. Wimmer, R. D. et al. Thalamic control of sensory selection in divided attention. *Nature* **526**, 705–709 (2015).
46. Wen, X. et al. Exploring communication between the thalamus and cognitive control-related functional networks in the cerebral cortex. *Cogn. Affect. Behav. Neurosci.* **21**, 656–677 (2021).
47. Wang, B. A. & Pleger, B. Confidence in decision-making during probabilistic tactile learning related to distinct thalamo-prefrontal pathways. *Cereb. Cortex* **30**, 4677–4688 (2020).
48. Wolff, M., Morceau, S., Folkard, R., Martin-Cortecero, J. & Groh, A. A thalamic bridge from sensory perception to cognition. *Neurosci. Biobehav. Rev.* **120**, 222–235 (2021).
49. Leung, B. K. & Balleine, B. W. Ventral pallidal projections to mediodorsal thalamus and ventral tegmental area play distinct roles in outcome-specific Pavlovian-instrumental transfer. *J. Neurosci.* **35**, 4953–4964 (2015).
50. Collomb-Clerc, A. et al. Human thalamic low-frequency oscillations correlate with expected value and outcomes during reinforcement learning. *Nat. Commun.* **14**, 6534 (2023).
51. Sherman, S. M. Thalamus plays a central role in ongoing cortical functioning. *Nat. Neurosci.* **19**, 533–541 (2016).
52. Saalmann, Y. B., Pinsk, M. A., Wang, L., Li, X. & Kastner, S. The pulvinar regulates information transmission between cortical areas based on attention demands. *Science* **337**, 753–756 (2012).
53. Mo, C., McKinnon, C. & Murray Sherman, S. A transthalamic pathway crucial for perception. *Nat. Commun.* **15**, 6300 (2024).
54. Sherman, S. M. & Guillery, R. W. Distinct functions for direct and transthalamic corticocortical connections. *J. Neurophysiol.* **106**, 1068–1077 (2011).
55. Nakajima, M. & Halassa, M. M. Thalamic control of functional cortical connectivity. *Curr. Opin. Neurobiol.* **44**, 127–131 (2017).
56. Halassa, M. M. et al. Developing algorithmic psychiatry via multi-level spanning computational models. *Cell Rep. Med.* **6**, 102094 (2025).
57. Huang, A. S. et al. A prefrontal thalamocortical readout for conflict-related executive dysfunction in schizophrenia. *Cell Rep. Med.* **5**, 101802 (2024).
58. Mukherjee, A. & Halassa, M. M. The associative thalamus: a switchboard for cortical operations and a promising target for schizophrenia. *Neuroscientist* **30**, 132–147 (2024).
59. Anticevic, A. & Halassa, M. M. The thalamus in psychosis spectrum disorder. *Front Neurosci.* **17**, 1163600 (2023).
60. Wang, B. A., Veismann, M., Banerjee, A. & Pleger, B. Human orbitofrontal cortex signals decision outcomes to sensory cortex during behavioral adaptations. *Nat. Commun.* **14**, 3552 (2023).
61. Sarafyazd, M. & Jazayeri, M. Hierarchical reasoning by neural circuits in the frontal cortex. *Science* **364**, 6441 (2019).
62. Penny, W. D. et al. Comparing families of dynamic causal models. *PLoS Comput. Biol.* **6**, e1000709 (2010).
63. Moustakides, G. V. Optimal stopping times for detecting changes in distributions. *Ann. Stat.* **14**, 1379–1387 (1986).
64. Lorden, G. Procedures for reacting to a change in distribution. *Ann. Math. Stat.* **42**, 1897–1908 (1971).
65. Lerma-Usabiaga, G., Liu, M., Paz-Alonso, P. M. & Wandell, B. A. Reproducible Tract Profiles 2 (RTP2) suite, from diffusion MRI acquisition to clinical practice and research. *Sci. Rep.* **13**, 6010 (2023).
66. Tournier, J.-D. et al. MRtrix3: a fast, flexible and open software framework for medical image processing and visualisation. *Neuroimage* **202**, 116137 (2019).
67. Veraart, J. et al. Denoising of diffusion MRI using random matrix theory. *Neuroimage* **142**, 394–406 (2016).
68. Kellner, E., Dhital, B., Kiselev, V. G. & Reiser, M. Gibbs-ringing artifact removal based on local subvoxel-shifts. *Magn. Reson. Med.* **76**, 1574–1581 (2016).
69. Smith, S. M. et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage* **23**, S208–S219 (2004).
70. Jeurissen, B., Tournier, J.-D., Dhollander, T., Connelly, A. & Sijbers, J. Multi-tissue constrained spherical deconvolution for improved analysis of multi-shell diffusion MRI data. *Neuroimage* **103**, 411–426 (2014).
71. Tournier, J. D., Calamante, F. & Connelly, A. Improved probabilistic streamlines tractography by 2nd order integration over fibre orientation distributions. *Proc. Intl. Soc. Mag. Reson. Med. (ISMRM)*. Vol. 18 (2010).

72. Cox, R. W. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162–173 (1996).
73. Bin-A-Wang2. Bin-A-Wang2/ReversalLearning\_Thalamic\_regulations: Thalamic\_regulations\_for\_RL\_strategies\_v1.O.O. Zenodo <https://doi.org/10.5281/zenodo.16942458> (2025).

## Acknowledgements

This work was supported by the following funding sources: National Natural Science Foundation of China (Project number 32200867) to B.A.W.; Research Center for Brain Cognition and Human Development, Guangdong, China (Project number 2024B0303390003) to B.A.W.; Guangdong Basic and Applied Basic Research Foundation to B.A.W. (Project number 2025A1515010766); Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) to B.P. (Project number 122679504 and PL602/6-1), and the FoRUM grant F971-2019 and F1081N-2023, medical faculty of the Ruhr-University Bochum to B.P.

## Author contributions

B.A.W., B.P. and M.M.H. conceived the project. B.A.W. performed the human data collection and analyzed the human data with the inputs from S.L. on methodology. M.B.W. performed the CogLink modeling analyses. N.H.L. and R.D.W. collected and analyzed the mouse data. L.M. and P.M.P. collected and analyzed the diffusion-weighted human data. B.A.W., M.M.H. and B.P. wrote and edited the manuscript. All authors read the final version of the manuscript. M.M.H. and B.P. supervised the work.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-63995-x>.

**Correspondence** and requests for materials should be addressed to Michael M. Halassa.

**Peer review information** *Nature Communications* thanks Julien Bastin and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025