nature portfolio

Peer Review File

Thalamic regulation of reinforcement learning strategies across prefrontal-striatal networks

Corresponding Author: Professor Michael Halassa

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In this report, Wang et al. studied reversal learning by combining (i) data from 32 healthy subjects performing a reversal learning task during fMRI scanning, (ii) data from 2 mice performing a rule reversal task while the authors applied optogenetic mediodorsal thalamus silencing and (iii) a separate data-set of more than a hundreds of subjects was also investigated to report tractography data regarding specific subregions of the mediodorsal thalamus. Using these experiments and data-sets, the authors investigate the neural substrate of reversal-based learning processes using modeling analyses to classify learning adaptations following a reversal as either being better captured by a model-free or model-based RL algorithm. They found that dissociable neural circuits underlie these processes, with a dorsal PFC-lateral MD circuit mainly activated for model-based RL strategy and a striatum- merdial MD circuit for model-free RL behavioral adaptations. They also report on (i) multi-voxel spatial analyses and representational similarity analyses (results L154-174) , (ii) an optogenetic preliminary experiment from 2 mice (L191-207; data only presented as supplemental material), (iii) they next used model-based analyses of their fMRI data and propose to split their behavioral blocks as model-based or model-free behavior to run additional analyses aiming at dissociating the neural substrate of these processes using a variety of tool (basic fMRI analyses L 209-223; followed by: 1. Voxel-wise mapping ; 2. Tractography and PPI analyses of MD subregions L224-260) and finally 3-4 network based analyses (3; DCM L 262-275 and 4. CogLinks analyses L276-344).
First of all, I would like to congratulate the authors on having carried out such an impressive number of experiments and analyses. That said, I also found the manuscript often difficult to follow, perhaps due to the paucity of information related to a critical point of the manuscript (the link between behavior, implicit cognitive processes and model-based and non-model-based learning strategies). In its present form, the manuscript is indeed characterized by a relative lack of explanation of very important aspects of task and behavior, and it could also benefit from a significant reorganization concerning the way data are presented and the way each analysis is conceptually introduced. Overall, therefore, I found this dataset to have potential but to require substantial revision in relation to a number of issues which I detail below.

Major comments.
1. Lack of behavioral support for a dissociation between model based vs. model-free learning strategy
a. The choice of the task with respect to the targeted question does not seem really justified in light of the literature and does not seem optimal compared to previous studies from other laboratories (Daw et al., 2011; Wunderlich et al. Neuron 2012; Smittenaar et al., Neuron 2013). Here, the data provided do not really allow to clearly distinguish a model-based strategy from a model-free strategy. This point being critical for the rest of the analyses, it is very important to reinforce this aspect of the work if possible. As it stands, I found that the manuscript suffers from a kind of circular reasoning since the authors use a RL-model based vs. RL model-free model-fitting procedure to distinguish the two strategies post-hoc but there is no real rationale for doing so a priori (i.e., based on the subjects' behavior, task-structure or previous litterature report on this topic to my knowledge). In addition, there is a lack of a priori justifications and references that would indicate that this type of task does indeed require an RL-based strategy (the references indicated in the manuscript use more complex tasks and in which the two types of strategies can be better dissociated). My point is that it is not because mathematically we can fit both types of models that cognitively it makes sense in the context of this experiment and that since there has not really been any significant work of simulation and there is a lack of significant behavioral data demonstrating that subjects do use a model-based strategy during this task, this point is problematic for the moment. I thus encourage the authors to strongly reinforce this important aspect which will strengthen their work and conclusions.

b. Relatedly, given the importance of this behavioral dissociation between strategies (based on amodel or not) for the

following neuroimaging analyses, I think it would be nice if the authors could more firmly address how their behavioral data clearly shows that subjects behave like a “model-free” or “model-based” in the context of their task (that is left undexplained in terms of cognitive processes at play and also not accessible to non –specialist in the method section); it would improve the readability to explain how the two models generate dissociable behavioral predictions in this tasks;

c. From figure 1c, it seems that behavior is below chance level during the 3/5 trials following a reversal before being above chance levels during the 5-10 trials post-reversals; could the authors do more on the dynamics of the behavioral adaptation ? This would imply more behavioral analyses to better delineate the dynamics of behavioral remapping occurring post-reversals;
• Similarly, I struggled with a method section dedicated to describing the method for generating Figure 1d: lines 742–756: “to plot Figure 1d, we…”: explains how the separation between the free model and the based model was done. This puzzled me because I understood the content of Figure 1d to be independent of any step in the RL model fitting procedure since I thought it showed the average learning performance during STAY trials (i.e., the average accuracy during the 10 trials occurring before a reversal) or SWITCH trials (i.e., the 10 trials occurring after the reversal)

d. I found the presentation of the behavioral data difficult to follow because the behavioral results are scattered across the four figures (in Fig. 2a; 3b and Fig. 4.e-f, using the same dataset). This relative weakness of the ms. is somewhat accentuated by the fact that the behavioral results are not always clearly presented or discussed (see below); the paper could be improved by a dedicated behavioral section describing the full behavioral properties of the task before the presentation of the neuroimaging results, also highlighting how model-free and model-based predictions differ in this task (for instance it would be nice to demontrate whether simulations of model-based vs. model free algorithms would predict different behavioral outcomes before or after rule reversal perhaps to identify relevent behavioral metrics to more convincingly sitinguish both types of strategies during the task used in the paper).

d. Fig 2a: I was unable to identify the methods behind this behavioral result. From the current text, it appears that the authors just checked how many incorrect (stay) trials they observed compared to correct choices (more rewarding after the contingency change during the reversal, i.e. switch trials). This confirms that subjects were able to perform the task above chance levels. I also found a problematic typo in the text as I assume the numbers were reversed in the main result sentence: strategy switch trials should be 5.8 on average rather than 4.2 and vice versa for strategy maintain trials (not 5.8 but rather 4.2 trials on average). This is not a striking difference and from this data point alone one can wonder how well subjects performed the task. To go further, perhaps the authors could consider alternative behavioral analyses such as assessing when performance starts to reach significance after the reversal begins? Performance seems to be very good for trials 7-10. Perhaps more statistics on this dynamic reconfiguration could also help them demonstrate on average how participants perform and perhaps how model-based learning and model-free learning predict a different dynamic reconfiguration pattern? (instead of just summing the number of correct and incorrect trials after the reversal in a block of 10 trials; I suppose this is what was done but I could not find the information in the method). Could the author improve this section of the manuscript (methods and results) by putting more effort into characterizing their subjects' behavior during the task and also more interpretation in terms of the cognitive processes at play during the two types of learning strategy (free vs. model-based)?

e. Fig. 3b shows the mean learning curves separately for model-free and model-based trial blocks. Yet, there is no clear description of (i) how model-free and model-based blocks were distinguished (the method section on this topic is rather obscure and at least a statement of the type of behavioral signature distinguishing the two types of behavior should be provided) (ii) there is no statistical quantification of model-based versus model-free learning behavior which is strange since the curves seem to diverge significantly (iii) it appears to be a post-hoc identification of learning processes with faster learning sessions after the reversal being labeled as model-based and others (less successful sessions characterized by a slower learning rate being labeled as model-free sessions). There is no discussion of alternative interpretations of these behaviors? Could the authors demonstrate that subjects rely on distinct strategies rather than alternative scenarios such as sampling bias, behavioral noise, inter-individual learning performance, other RL architectures? (iv) the comparison between the two types of learning sessions using fixed-effect across blocks is problematic (for fig 4e-f, see below). Please consider using linear mixed effect to better demonstrate that effects can be seen across subjects (and avoid fixed effects across blocks generally since reliability of such effects across subjects cannot be assessed with such a statistical approach)

2. Neuroimaging results.
a. Rule reversals section (113-44; fig1): this section is very clear
b. Representation of decision strategy (L146-89): the concept of “decision strategy” induced a lack of clarity: the authors looked at the average correct vs. incorrect number of trials following a reversal behaviorally (and they tagged this a “decision strategy” somehow artificially and post-hoc). I did not understand why they chose this terminology (for consistency across paragraph I would keep the stay vs. switch terminology as in fig 1 and I would add the layer correct vs. incorrect for this analysis that would simplify a lot the readability). The neural analyses here appear less robust and the rationale behind this analysis weaker compared to fig 1. I did not really catch the reason explain why the authors split the behavioral policy into their three categories: “overall strategy” just combines correct and incorrect post-reversal trial types, the “keep strategy trials” correspond to incorrect switch trials and “change strategy” captures correct switch trials. I found the chosen conceptual/terminology framework difficult to follow and it remains hard for me to understand for example how the neural activity of one brain regions can both represent two type of strategies (such as the dmPFC which is sensitive to both overall and change strategy);
c. Related to above: given the structure of the analysis and the RDMs shown in fig. 2b the significant effect of overall strategy could be simply driven by the strategy change (because the strategy change trials might drive the effect seen in the overall analysis)?

d. it might be informative to see the BOLD response related to the contrast between correct switch trials and STAY vs. incorrect switch trials and STAY (this might provide additional pieces of information to interpret the data provided by the RSA analysis;
e. could the author discuss/speculate on why the striatum would be active for the switch vs. stay contrast of fig 1 but not by the RSA?
f. Optogenetic preliminary data: two mice, 6 sessions. I honestly did not assess this part due to a low confidence on the replicability of such a small/preliminary data-set. The statistics reporting is also below the standards (no degrees of freedom, no specification of the kind of test done in the main result…)
g. MD regulates RI representations in PFC or striatum (L209-260):
• The authors use fig 3b to illustrate that "subjects used a mixture of model-based and free strategy"; as indicated above, I found there was a circular reasoning there because they applied a RL model fitting procedure to identify two types of learning blocks in which subjects either poorly learn (slow learning = model free) or learned more rapidly to switch their choice to the more rewarding cue (less perseverative behavior following reversal). See my comment on behavior: there is currently a lack of support based on raw behavioral data that subjects did use two strategies (or the interpretation might be very different from using a model of the environment or not). It would be nice to see any data and new analysis supporting this claim (which is central to the paper).
• To better compare the RPE vs state-base PE it might be informative to plot RPE and SPE during either model-free or model based to see how much these two signals diverge within vs. between blocks ?
• Perhaps the authors could simply do more to characterize the distinction between free vs. model-based blocks using behavioral metrics for which we would see that the data of the model-free blocks cannot be predicted by the model-based RL architecture and vice-versa: the model-based behavioral results cannot be predicted by the model-free RL architecture: this would be more solid evidence that there was something in the data that can be better captures by a model rather than the other…
That said, I found the following fMRI, tractorgraphy and PPi analyses very nice and clearly reported (Fig 3: neural data part).
h. Network analyses (262-344): I was surprised to see that part of the figure 4 (based on network based results) re-analysed RSA results (presented in figure 2; would'nt it be simpler to add the RSA layer (that simply split model free vs. model-based blocks) to figure 4 ?

DISCUSSION! I found the results were very briefly discussed and they were not even positioned relative to previous literature on the involvement of thalamic nuclei during RL processes. The dorsomesial thalamus was shown to be involved during a two-arms bandit tasks with direct electrophysiological recordings from humans (Collomb-Clerc et al. Nature comms. 2023). I guess the very short discussion might be due to an initial submission as a brief report. An expanded discussion section would be of interest for a revised manuscript.

In summary, I found the data analyses to be remarkable for the fMRI dataset and the modeling effort to be quite impressive as well. However, this came at the cost of a clear behavioral dissociation between the two learning strategies that are explored here and I felt that the impressive post-hoc effort did not outweigh the weakness of the behavioral dissociations between the two types of models: model-free and model-based predictions are indeed never directly compared, so it remains to be demonstrated whether and how RL behavior following reversal in this task can be used to assess both forms of learning strategy, unlike the initial results from 2011.

Minor comment:
Abstract: I found the term multimodal human imaging misleading (since all the relevant data come from fMRI and the dTI data-set is not central) perhaps just state fMRI and DTI techniques; please reformulate

Reviewer #2

(Remarks to the Author)
Wang and colleagues report the results of a study designed to assess the role of the mediodorsal region of the thalamus in regulating learning and flexible behavioral updating. The authors describe results detailing separate thalamocortical pathways for model-free versus model-based control of behavior and find evidence that distinct regions within MD support each process. Moreover, they found that MD-prefrontal pathways were engaged during strategy updating, which biophysical modeling suggest supports rapid updating of context representation enabling flexible behavior.

On the one hand, I am enthusiastic about the importance of the question targeted by the manuscript (which has been understudied and is important for understanding basic cognition and circuit dysfunction in disease) and am impressed with the methodological breadth of the manuscript. The combination of fMRI methods (univariate, multivariate, and dynamic causal modeling), DWI, confirmation of the importance of this pathway using optogenetic silencing in rodents, and biophysical modeling of the circuit is not only impressive but provides synergistic and converging sources of evidence about the distinct roles of medial and lateral DM in RL. This approach reflects my hope for the future of cognitive neuroscience.

On the other hand, I found this manuscript difficult to follow, even after multiple read-throughs, and lack of clarity often make it hard or impossible to evaluate the authors' claims. As a result, there were some key claims (particularly in the RSA, PPI, and biophysical modeling sections) that I could not evaluate. In other sections, I have significant concerns about the

analyses.

-On the clarity point, there are many points in the Results where claims are made without any explanation of what exactly was tested, or even what the statistical results were. This problem is especially pronounced in the subsection "Thalamocortical pathways implements the critical computations for reinforcement learning during strategy updating", which contains no statistical results in the main text. Some of these results are described in the Figure (a pattern repeated throughout the paper), but I found myself needing to switch between the Figure legends, Methods, and Supplement to try to understand a result described in the Results. I will just list a few illustrative examples rather than describing them all, but I advise a careful re-write of most of the Results and a significant portion of the Methods sections:
a. "Using bilinear dynamic causal modelling (DCM), we observed recurrent connectivity of MD with both dmPFC and CN, significantly modulated by rule reversals (posterior probability = 0.6)." What does that probability correspond to? A posterior probability of .6 does not seem very high to me to claim an effect? Additionally, readers unfamiliar with DCM may not understand what this connectivity signifies.
b. The main text says: "Consistent with this notion, functional connectivity from the MD to dmPFC predicted switching speed, ($r = -0.515$, $p = 0.008$)." It is not clear if this is "effective connectivity" from the DCM, and whether this the baseline connectivity or the modulation by rule reversals. It is also unclear what switching speed is. The figure clarifies that switching speed is the number of trials to switch, but also describes multiple comparisons corrections, which refer to additional tests described only in the Supplement. As an additional note about this analysis, I am concerned about the influence of leverage points, especially with this sample size, and the authors should conduct additional control analysis using a nonparametric or outlier-deweighting method.
c. Line 205 has no statistics associated with the result
d. The methods underlying the PPI analyses are split into three different sections in of Methods, making them extremely difficult to follow and evaluate.
e. What does it mean for the biophysical model to be capable of solving the task? How is that shown or evaluated?
f. Is figure 4e human or model data?

-SPE exhibits primarily slow fluctuations that are likely removed by the author's bandpass filter. What does the SPE regressor look like if filtered in the same way as the data, and does this correspond conceptually to SPE? Additionally, the model-based model is unclear to me: do the SPEs modulate the RPEs and learning process, or is SPE a separate process?

-In the strategy RDM analysis, overall strategy is colinear with strategy-keep and strategy-change. Additional control analyses are needed to confirm that significant overall-strategy effects in dmPFC are just driven by the strategy-change effect. The results should also make clear this was calculated during the switch period. After reading the Methods, it is unclear how the group statistical tests were conducted for these analyses.

-Many of the results rely on a median splitting of so-called model-based and model-free blocks. This was surprising given the other results in the paper describing co-active model-based and model-free strategies. The authors should show that these blocks are characterized by different behavioral profiles to justify separately probing their neural covariates.

-The 10 trial definition of switch and steady-state periods needs more explanation. How did the authors' come up with this number? Additionally, if this was based on looking at the data, the analyses comparing behavioral performance in these periods are circular and should be presented in a way that makes that clear.

-The methods for the model-free and model-based RDM analysis are not clear to me. I follow the section of the Methods describing the GLM construction, but I did not find a description of the RDM analysis and therefore do not understand how the authors found that "we found that model-based MD engagement was localized to its lateral subdivision, whereas the model-free to its medial one".

-The biophysical model predictions and results are insufficiently explained. For example, I am not sure what Figure 4h depicts. The texts in the results do not describe this result in much detail, stating that switching requires overwriting of previous OFC mappings. How do the action values on the y-axis reflect OFC mappings? Where is evidence of over-writing in this plot? I used this inset as an example, but I was similarly confused for the other analyses in this figure.

-If the optogenetics results remain in the main text, the key results should be in a figure in the main text.

-Why are the plots presented as uncorrected tmaps for display purposes? In principle this is OK if justified (which is needed), if the key regions survive FWE correction, would these plots be different if only corrected data were shown?

-More detail and justification is needed for the subjects excluded for poor behavior

-How were the ROIs defined?

Reviewer #3

(Remarks to the Author)
The authors, in the current manuscript, are building on a "serendipitous finding" that the MD thalamus robustly activates upon strategy updating and encoding in a probabilistic reversal learning task. Through a number of approaches, including univariate, RSA, and connectivity-based approaches, the authors build a compelling idea for a role of
These findings suggest again that the MD is positioned between high-level prefrontal and low-level striatal representations,

which underlies the modulation of frontal cortico-striatal networks during Switch period.

High Level Comments

Have to say, I'm quite happy to see the regional analysis MD in this analysis – the idea of viewing it as monolithic (as often is an inadvertent side effect of fMRI limitations) seems quite at odds with neuroanatomy, and here it appears that a medial/lateral split is quite crucial to understanding its function. However, I find the lack of a similar treatment of striatum to be somewhat surprising – prior work has observed a split between dorsolateral striatum (associated with MF behavior) and dorsomedial striatum (associated with MB behavior). Instead, the authors position competition here between "low-level striatum" and "high-level prefrontal" regions.

In general, the study is quite methodologically broad – and this is certainly a strength. Integrating behavioral analysis, computational modeling, multiple fMRI approaches, as well as an animal activation study is no small feat, and I applaud the authors for the enormous amount of work that has clearly gone into the set of excellent results here. This, however, brings me to what (somewhat surprising to me) feels like one of the weakest points of the paper, providing a good contextualization of these results, particularly vital given the multiple approaches they span.

I find the discussion to be somewhat short, and while it presents a reasonable introduction to the ideas surrounding this literature, the authors spend very little unpacking the specific results of their analyses. As such, the paper currently reads as a number of reasonably strong and complementary analyses, but I think needs much more work in presentation to spell out for the reader how these fit into a single, consistent understanding.

Behavior and regional activation results
The authors analyzed behavior on the reversal learning task, and found that (as expected) subjects showed behavior of high accuracy prior to a switch, and a decrease followed by slow increase after a switch. This (and associated regional activations and link with behavior) all seems quite reasonable.

RSA Analysis
2a) Perhaps I'm missing the point of this figure and the associated statistical test – the expected number of strategy keep vs. change trials is obviously reflective of how quickly subjects update their policies, but I'm not sure why one would want to test for a difference in significance between them, or what that significance would imply – rather, the ratio seems like it would be entirely adjustable by changing the post-switch window size – with a shorter window size, keep would dominate, but with a larger window size, change will dominate.

2b-d) I'm struggling to understand the results in dmPFC – given that the overall strategy is a combination of the two individual strategies, how is it that we find their combination, one of the individual components, but not also the other individual component? This is less a question about interpreting their results, and more perhaps understanding how this could reasonably fall out of the analysis.

Separately, are the strengths of these relationships related between subjects? That is, do some subjects more generally have more robust rule representations, and if so, is this perhaps connected with behavior at all?

Mouse Results
The authors report an inactivation study of MD in mice – It absolutely seems like a strong argument in favor of this overall connectivity which involves MD being crucial for reversal learning, and is quite a nice addition to the story. However, unclear if it argues so much that the MD is uniquely gating this behavior, rather than involved, which appears to be much of the crucial point of this paper. I wonder if the authors could expand on this.

MF vs. MB Behavior
I don't think I entirely understand why we observe such oscillatory behavior in the model-free RPE. In particular, prior to a rule reversal, I would expect the exact timing and direction of RPEs to be relatively arbitrary, given the stochastic nature of the task. While not the part of the signal being used to identify MF RPEs, it seems like a quite striking signature of how they are deriving these model fits. Could the authors explain this? Otherwise, the findings in this section seem quite nicely tied in to the previous results.

I do have one additional question – from my understanding, the two neural correlates for MF and MB are based on assigning each block to one of the two based on performance. However, I would assume there to be a fairly substantial metalearning component – have the authors investigated whether behavior systematically changes/improves over time?

Computational Model
I have to admit I am somewhat lost with the computational model underlying a large portion of this section, which seems somewhat split between this figure, the supplement, and perhaps further detail in a cited paper which however is not yet published or otherwise available. That being said, the result with multiple concurrent representations during the MB based blocks is somewhat compelling.

I am also curious if the authors have any intuition for why results were somewhat assymetric – caudate for MF and dlPFC for MB – whereas their modeling assumes these both to be driven by distinct neocortical regions (OFC and dlPFC) and presumably both with their own BG connections.`

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I have reviewed the revised material, and I am pleased to confirm that they are satisfactory. I extend my gratitude to the authors for their meticulous consideration of all my concerns. The revised work has significantly strengthened from its previous version.

I have only one minor suggestion related to one of the new paragraphs in the discussion:
Gueguen et al. published a study (Gueguen et al. Nat. Comms 2021) in which it was shown that the ventromedial and dorsolateral prefrontal cortex encode different types of prediction errors (reward and punishment prediction errors). Could the authors modify the sentence 'reward prediction errors are associated with the ventral striatum' (L 501-502) and incorporate the results of Gueguen et al.?

Reviewer #2

(Remarks to the Author)
The authors have undertaken a substantial revision in response to my concerns and those of the other reviews. The resulting manuscript is substantially improved— the new analyses and explanations clarify the findings and impacts. I am particularly impressed with the use of the biophysical model to explain the DCM findings and to make new RSA and behavioral predictions. Overall, I am enthusiastic about the revision, and in general the synthesis of methods and approaches to interrogate the role of the MD nucleus in learning. However, I do have some additional concerns that were raised in my reading of the revision and response letter:

1) Most significantly, I am concerned about statistical circularity in performing RSA analyses on functionally-defined ROIs. The authors should ideally use anatomical or independent ROIs, or very carefully justify why the RSA analyses are statistically independent.
2) I was very confused by the description of the RSA analyses in the rebuttal letter. Figure R9 appears to be asserting a finding from a null—this should be more carefully interpreted. Figure R10 is inscrutable to me—the labels are not very informative.
3) The authors should include some caveats about interpreting correlations with their sample size (n = 32)
4) Were the average modulatory effects in the DCM positive or negative?
5) Are the SPE and RPE results collapsing across the MB and MF blocks? If so, are these result different between those blocks?
6) In the biophysical modeling, how are the MF and MB blocks defined? Is this model trained to explain the human behavior, or only to perform the task?
7) The results from the mouse data should include some more justification that mice could learn the task and that the split into switch and steady-state blocks is justified
8) The statistical reporting has been much improved, but there are still cases where only p-values are reported. Please adopt a consistent style, with all information required according to the reporting summary. Additionally, the way the multiple comparisons correction is presented is somewhat unorthodox and confusing.
9) The sentence "MDl neurons exhibited strong contextual encoding (Fig. 6b), with their activity closely linked to MB behaviors, mirroring patterns observed in human subject" needs more explanation (this seems to be explained more fully in the Figure legend than in the text
10) As mentioned, I found the RSA analyses arising from the biophysical modeling valuable. However, I think the results are too strongly interpreted as written. These RSA results are not evidence of toggling, though toggling is one possible explanation for the pattern the authors observe.

Reviewer #3

(Remarks to the Author)
I thank the authors for their lengthy revisions and accompanying analyses. I find that the paper is significantly better connected than before, though perhaps still slightly disjoint. My only significant issue at the moment is with a new analysis the authors have introduced to show that the behavior is indeed a mixture of model-based and model-free behavior. While I think that the idea of such an analysis would significantly strengthen the paper, I do not believe the current attempt at this is methodologically sound and should be in some form replaced.

Details below:

Behavior
The authors have significantly improved their characterization of switching behavior. The lapse rate, slope, and delay seem like a reasonable choice, and a much better parametrization to fit.

However, I am not at all sold on the new analysis that the authors have included in figure 1g and 1h. This analysis appears to have been added as a response to multiple reviewer comments that the model-free/model-based behavioral analysis was somewhat lacking. I believe that this new analysis introduces a number of new issues.

The authors claim that, as their task involves common/rare rewards, this maps onto the common/rare transition structure of a two-step task. Note that the two-step task already involves varying reward rates, but the common/rare analysis is explicitly at the level of transitions, not at the level of such rewards. In the two-step task, rare outcome for a model-based agent should lead to an explicit reversal of attribution, e.g. a rare rewarded trial should increase confidence in the unchosen option, and vice-versa. That is, knowledge of the structure of the task leads to opposite interpretations of the same outcomes.

However, that's clearly not the case here, principally because a model of the task doesn't allow me to differentiate common/rare unless I already quite confidently know which regime I'm in, which defeats the entire point. That is, assume option 1 to be the 70% rewarded option, I choose option 1, and it's unrewarded. The two-step logic is that, due to this being a rare transition, I should assign that lack of reward to option 2, and in fact increase my confidence in option 1. Quite frankly, that logic makes no sense on the current task. At best, a strong prior belief that option 1 is the better option means that I may be less sensitive to an unrewarded trial on option 1, but there's no plausible reason why I should take that lack of reward as evidence against option 2.

So while the authors certainly can perform this analysis on the current dataset, mapping onto common/rare reward outcomes, it's somewhat incomprehensible from my perspective. My own interpretation of figure 1h is that the reviewers convincingly show a very strong MF effect, but what they're interpreting as MB is not due to a model so much as recent trial history, and a tendency of rare or common unrewarded outcomes to be exactly that. (Lagged learning effects would likewise provide a problem in the two-step task, hence typical setups of reward probabilities that drift around 0.5).

More generally, I think the idea of a one-back model might be misguided here, and there may not be a great way to build intuition that doesn't involve an actual history of recent trials. That being said, I believe most of the analyses linked to behavior later rely on a comparison of something akin to an ideal observer changepoint model versus a model-free agent, so I hope that this should not cause too many problems?

RSA Analysis:
Figure 3 is much clearer, and the authors have removed the somewhat confusing test of overall strategy.
I find fig R13 quite interesting, though somewhat confusing. Increased RDM-related activity patterns are associated with a larger offset, e.g. slower switching behavior? Isn't this somewhat unintuitive? Also unclear on the multiple comparisons number. What does 0.05/3 refer to? However, I do appreciate that the authors have carried out these tests.

Mouse Results
The authors have elaborated on the role of the MD here, linking the mouse and human studies, quite a bit better than before.

MF vs. MB Behavior
I'm still not sure that I follow the oscillatory behavior prior to switches. The oscillation due to over/under prediction makes sense, but I find it surprising that with a pseudo-randomized sequence there would still be so much structure across blocks. I'm assuming that the pseudo-randomized sequence was relatively random between blocks – or did it simply end up with highly similar reward sequences prior to the switch even across different blocks?

I thank the authors for carrying out the metalearning analysis across blocks. It is interesting and surprising to me that the actual detection of reversal didn't improve with practice, though I suppose quite plausible.

MF vs. MB Behavior
The presentation here is much improved.

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
The revised manuscript satisfies all of my concerns.

Reviewer #3

(Remarks to the Author)
I have reviewed the updated manuscript, and I am generally satisfied with the changes the authors have made.

However, I am not sure if the authors have fully revised the methods section in-line with their response? I still see discussion of the common/rare transition logic, which probably should be removed.

If I had to summarize, it seems that the authors have essentially switched to fitting whether individual blocks are best

represented by a changepoint detection algorithm versus a more naive model-free approach. It certainly tests whether learners are picking up on what would be considered a model of task structure, but I'm not sure this falls within the classic reinforcement learning sense of model-based.

It may be worth clarifying how exactly this kind of model fits into the reinforcement learning literature somewhat in the text.

Responses to reviewers' comments

We would like to thank the editor and reviewers for giving us the opportunity to respond to the comments made on the manuscript. The encouraging remarks from the three reviewers are particularly significant to us, as they validate our research efforts and agree with the significance of our work to the field.

We also agree with the constructive criticism, which has helped to clarify the results and strengthen our conclusions. Based on reviewers' comments, the manuscript has undergone substantial revisions.

Below, please find our point-by-point responses to each reviewer's comments. Sentences in black correspond to the reviewer's original comments, followed by our responses in blue. We have highlighted all changes in the revised manuscript.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

In this report, Wang et al. studied reversal learning by combining (i) data from 32 healthy subjects performing a reversal learning task during fMRI scanning, (ii) data from 2 mice performing a rule reversal task while the authors applied optogenetic mediodorsal thalamus silencing and (iii) a separate data-set of more than a hundreds of subjects was also investigated to report tractography data regarding specific subregions of the mediodorsal thalamus. Using these experiments and data-sets, the authors investigate the neural substrate of reversal-based learning processes using modeling analyses to classify learning adaptations following a reversal as either being better captured by a model-free or model-based RL algorithm. They found that dissociable neural circuits underlie these processes, with a dorsal PFC-lateral MD circuit mainly activated for model-based RL strategy and a striatum- merdial MD circuit for model-free RL behavioral adaptations. They also report on (i) multi-voxel spatial analyses and representational similarity analyses (results L154-174) , (ii) an optogenetic preliminary experiment from 2 mice (L191-207; data only presented as supplemental material), (iii) they next used model-based analyses of their fMRI data and propose to split their behavioral blocks as model-based or model-free behavior to run additional analyses aiming at dissociating the neural substrate of these processes using a variety of tool (basic fMRI analyses L 209-223; followed by: 1. Voxel-wise mapping ; 2. Tractography and PPI analyses of MD subregions L224-260) and finally 3-4 network based analyses (3; DCM L 262-275 and 4. CogLinks analyses L276-344).

First of all, I would like to congratulate the authors on having carried out such an impressive number of experiments and analyses. That said, I also found the manuscript often difficult to follow, perhaps due to the paucity of information related to a critical point of the manuscript (the link between behavior, implicit cognitive processes and model-based and non-model-based learning strategies). In its present form, the manuscript is indeed characterized by a relative lack of explanation of very important aspects of task and behavior, and it could also benefit from a significant reorganization concerning the way data are presented and the way each analysis is conceptually introduced. Overall, therefore, I found this dataset to have potential but to require substantial revision in relation to a number of issues which I detail below.

We appreciate the reviewer's encouraging comments. In response to the reviewer's feedback, we have made substantial revisions to our manuscript, including additional analyses of the behavioral and neuroimaging data. Below, we outline how we have addressed each specific point.

Major comments.

1. Lack of behavioral support for a dissociation between model based vs. model-free learning strategy

We conducted additional analyses on the behavioral data to analyze the dynamic changes of performance across learning blocks in greater detail. To better understand participants' behavior following rule reversals, we fitted the data to a sigmoidal transition function using logistic regression with three parameters: switch offset s, switch slope α, and lapse ε.

To further differentiate between model-based and model-free RL strategies, we calculated the probability of repeating a decision (p(repeating)), which reflects the transition probability. This analysis examines how feedback from the current trial (n) influences the choice made in subsequent trials (n+1). Taking transition probability into account alongside reward characterizes model-based RL, whereas relying solely on reward feedback reflects model-free RL.

As detailed below, our additional analyses provide strong behavioral evidence supporting the distinction between exploratory, model-free strategies and structured, inference-based strategies.

a. The choice of the task with respect to the targeted question does not seem really justified in light of the literature and does not seem optimal compared to previous studies from other laboratories (Daw et al., 2011; Wunderlich et al. Neuron 2012; Smittenaar et al., Neuron 2013). Here, the data provided do not really allow to clearly distinguish a model-based strategy from a model-free strategy. This point being critical for the rest of the analyses, it is very important to reinforce this aspect of the work if possible. As it stands, I found that the manuscript suffers from a kind of circular reasoning since the authors use a RL-model based vs. RL model-free model-fitting procedure to distinguish the two strategies post-hoc but there is no real rationale for doing so a priori (i.e., based on the subjects' behavior, task-structure or previous litterature report on this topic to my knowledge). In addition, there is a lack of a priori justifications and references that would indicate that this type of task does indeed require an RL-based strategy (the references indicated in the manuscript use more complex tasks and in which the two types of strategies can be better dissociated). My point is that it is not because mathematically we can fit both types of models that cognitively it makes sense in the context of this experiment and that since there has not really been any significant work of simulation and there is a lack of significant behavioral data demonstrating that subjects do use a model-based strategy during this task, this point is problematic for the moment. I thus encourage the authors to strongly reinforce this important aspect which will strengthen their work and conclusions.

b. Relatedly, given the importance of this behavioral dissociation between strategies (based on a model or not) for the following neuroimaging analyses, I think it would be nice if the authors could more firmly address how their behavioral data clearly shows that subjects behave like a "model-free" or "model-based" in the context of their task (that is left undexplained in terms of cognitive processes at play and also not accessible to non –specialist

in the method section); it would improve the readability to explain how the two models generate dissociable behavioral predictions in this tasks;

We thank the reviewer for this valuable critique. Both comments address one critical point: the lack of prior justifications for applying different RL strategies to explain participants' learning behavior.

To further distinguish model-based and model-free strategies, we first questioned how the feedback obtained on the current trial (trial n) impacts the choice of the subsequent trials (trial n+1). To this end, we tested the probability of repeating a decision, p(repeating). We calculated and compared p(repeating) of rewarded and unrewarded trials for the *Steady State* (SS) and the *Switch* (SW) period separately, and found a significant main effect of reward ($F_{(1,31)} = 220.8$, $p = 1.2*10^{-15}$, Fig. R1), which emphasizes the reinforcing effect of reward on repeating a decision. Importantly, post-hoc analyses showed significant difference between SS and SW for unrewarded trials (paired two-sample t-test, $t_{(31)} = 4.9$, $p = 2.5*10^{-5}$, Fig. R1), but not for rewarded trials. This suggests a greater need to switch strategies when trials remain unrewarded during SW compared to SS, driven by an updated internal model of the task's transition structure following the rule reversal.

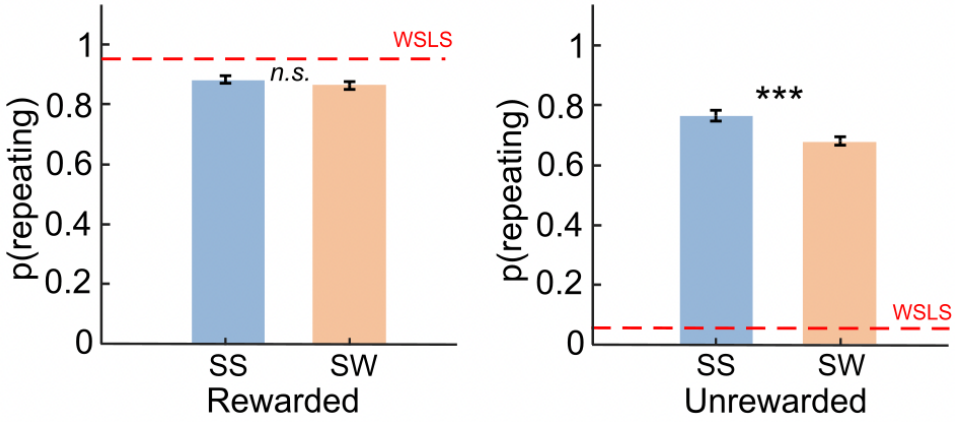


Fig. R1

In the context of a Markov decision task (Daw et. al., 2011, Neuron), model-free RL implies that a rewarded experience from a rare association directly increases the probability of repeating the previous choice (Fig. R2a). In contrast, a subject who relies on an internal model of the task's transition structure and evaluates actions prospectively exhibits a decreased tendency to choose the same option. This strategy, which incorporates the transition model into decision-making, is characterized by an interaction between reward and transition probability (Fig. R2b).

Our probabilistic reversal learning task does not involve transition probabilities but does include the probabilities of regular (common, 70%) and deviant (rare, 30%) events. Therefore, in line with the classic two-stage Markov decision task, we calculated *p*(repeating) of common (70%) and rare (30%) events for rewarded and unrewarded trials (Fig. R2c). For both SS and SW, we revealed significant results for the main effect of reward (model-free), and the interaction between reward and the probability (model-based) (SS: main effect of reward: $F_{(1,31)} = 116$, $p = 5.2*10^{-12}$, interaction: $F_{(1,31)} = 64$, $p = 4.7*10^{-9}$; SW: main effect of reward: $F_{(1,31)} =$

201, $p = 4.1*10^{-15}$, interaction: $F_{(1,31)} = 23$, $p = 3.4*10^{-5}$), suggesting that participants in our task used a mixture of exploratory, model-free strategies and deterministic, model-based behavior.

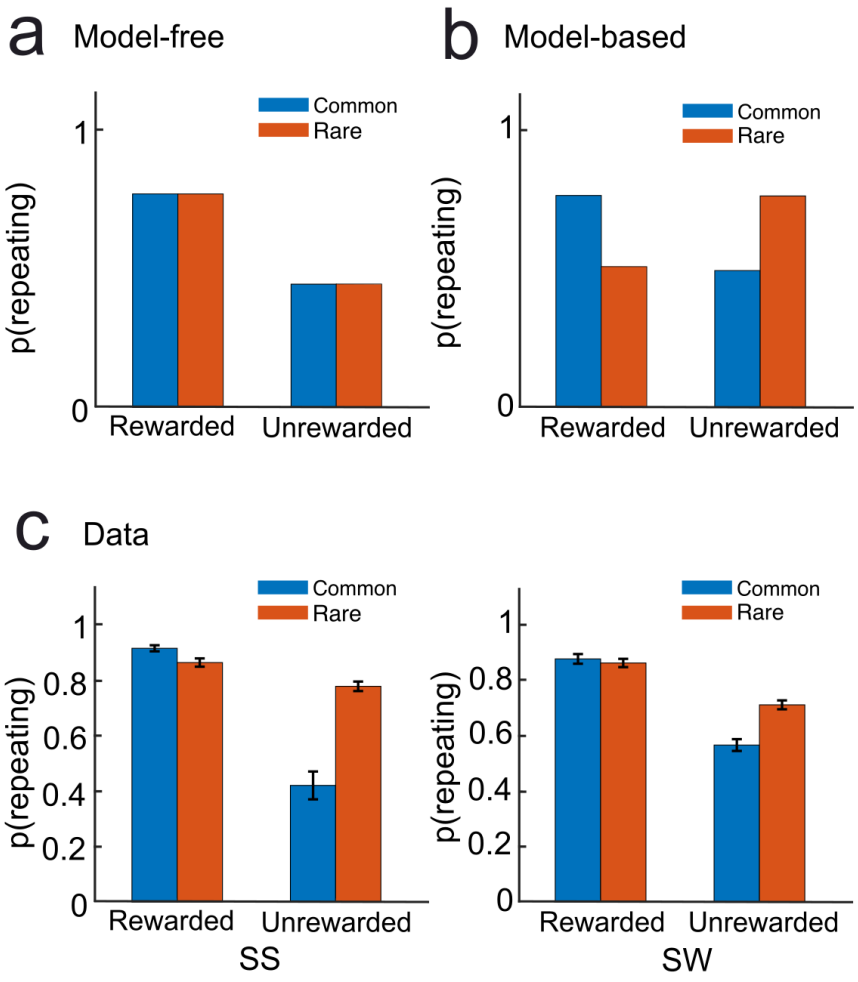


Fig. R2

These two additional analyses validate that our probabilistic reversal learning task inherently requires different RL-based strategies, leading to the observed variability in behavioral performance across blocks and participants. In the revised manuscript, we describe these additional behavioral findings, explain how different RL strategies generate dissociable behavioral predictions and that our results validate the involvement of these strategies in line with those predictions. Please refer to the **lines 141-169** in the revised manuscript.

c. From figure 1c, it seems that behavior is below chance level during the 3/5 trials following a reversal before being above chance levels during the 5-10 trials post-reversals; could the authors do more on the dynamics of the behavioral adaptation? This would imply more behavioral analyses to better delineate the dynamics of behavioral remapping occurring post-reversals;

We thank the reviewer for this important point.

To better capture the dynamics of behavioral adaptation, we calculated the average probability of choosing the correct strategy within the *Switch* period across the 12 blocks for each

participant (Fig. R3). We then fitted a logistic regression model with three parameters to individual choices:

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

where s, α, and ε are three free parameters representing the latent transition between actions: the switch offset, slope, and lapse, respectively (Fig. R3). The switch offset *s* measures the latency of the switch, the slope *α* measures the sharpness of the transition, while the lapse rate *ε* signifies the behavioral performance after the transition. We calculated the group-averaged switch offset *s*, slope *α*, and lapse *ε* for all participants (n = 32) (Fig.R3).

To compare the participants' performance and parameters to a benchmark, we simulated the behavior of a "win-stay-lose-shift" (WSLS) agent that updates the decision according to feedback. This model, as the name implies, repeats the choice if the previous action is rewarded and switches if it is unrewarded. In the noisy version of this model, it applies the win-stay-lose-shift rule with the probability of $1 - \varepsilon$, and chooses randomly with the free parameter $\varepsilon = 0.05$. In the two-choice case, the probability of choosing option k is:

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

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 for a total of 500 blocks. We calculated the average probability to choose the correct strategy within the *Switch* period the same way as for the human participants. Similarly, the logistic regression model was fitted to the WSLS agent's performance in the same way as described above. The dashed red line indicates the values of the three parameters for the WSLS agent (Fig. R3).

The analysis revealed that the median switch offset *s* was around 4 (mean ± std = 4.4 ± 1.6), indicating that participants reached above-chance levels by the 4th trial after rule reversals. The slope *α* was 1.3 ± 2.6, and lapse rate *ε* 0.13 ± 0.06. These behavioral analyses provide a more detailed delineation of the dynamics of behavioral remapping following rule reversals.

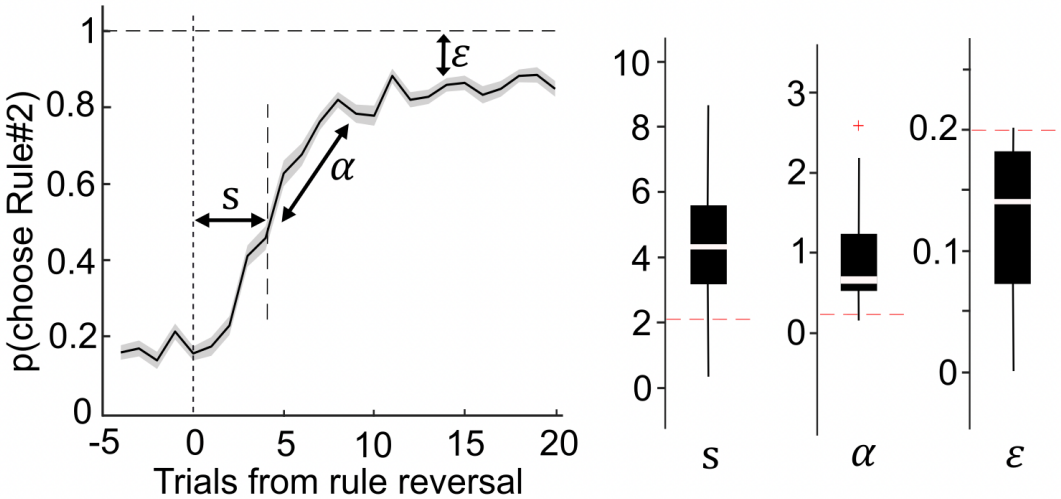


Fig. R3

We have integrated these results into the revised manuscript as follows (**Lines 115-129**):

> "To characterize the dynamics of behavioral adaptation in response to the rule reversals, we first calculated the average probability of selecting the correct 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 (α), and lapse rate (ε). Specifically, the switch offset (s) measures the latency of the switch, the slope (α) quantifies the sharpness of the transition, and the lapse rate (ε) reflects behavioral performance after the transition. For all participants (n = 32), we computed the switch offset (s), slope (α), and lapse rate (ε) (Fig. 1d). The mean switch offset (s) was approximately 4 (mean ± SD = 4.4 ± 1.6), indicating that participants reached chance-level performance by approximately the fourth trial following rule reversals. In addition, the slope (α) was 1.3 ± 2.6, and the lapse rate (ε) 0.13 ± 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)."

Similarly, I struggled with a method section dedicated to describing the method for generating Figure 1d: lines 742–756: "to plot Figure 1d, we…": explains how the separation between the free model and the based model was done. This puzzled me because I understood the content of Figure 1d to be independent of any step in the RL model fitting procedure since I thought it showed the average learning performance during STAY trials (i.e., the average accuracy during the 10 trials occurring before a reversal) or SWITCH trials (i.e., the 10 trials occurring after the reversal)

We sincerely apologize for the error. It should read "to plot Fig. 6c, we…". The reviewer is correct that Fig.1d (Fig. 1f in the revised manuscript) is totally independent of the RL modeling fitting procedure. In the Fig. 1f, we simply calculated the averaged learning performance during *Steady State* and *Switch* trials.

d. I found the presentation of the behavioral data difficult to follow because the behavioral results are scattered across the four figures (in Fig. 2a; 3b and Fig. 4.e-f, using the same dataset). This relative weakness of the ms. is somewhat accentuated by the fact that the behavioral results are not always clearly presented or discussed (see below); the paper could be improved by a dedicated behavioral section describing the full behavioral properties of the task before the presentation of the neuroimaging results, also highlighting how model-free and model-based predictions differ in this task (for instance it would be nice to demontrate whether simulations of model-based vs. model free algorithms would predict different behavioral outcomes before or after rule reversal perhaps to identify relevent behavioral metrics to more convincingly sitinguish both types of strategies during the task used in the paper).

To provide a more comprehensive presentation of behavioral findings, we have re-organized the figures and now present the behavioral performance in two parts: one in Fig. 1 and the other in Fig. 4. Specifically, in the new Fig. 1, we excluded the neuroimaging results and present only the complete behavioral properties of the task, including the probability of correct

strategies, the regression parameters, and the probability of repeating the previous strategy. These behavioral properties highlight the participants' averaged performance and the dynamics of their behavioral adaptation in response to the reversals. They also serve as indicators of the involvement of a mixture of different types of RL strategies during the task.

In the new Fig.4, we present the differences in behavioral properties between model-free and model-based learning blocks. For example, we observed that the model-free strategy produced behavior with longer offsets s to reach the chance level but subsequently exhibited the same slope and lapse as the model-based strategy.

Accordingly, we expanded the behavioral section of the manuscript to describe all behavioral properties of the task before presenting the fMRI results (**Lines 114-169**).

d. Fig 2a: I was unable to identify the methods behind this behavioral result. From the current text, it appears that the authors just checked how many incorrect (stay) trials they observed compared to correct choices (more rewarding after the contingency change during the reversal, i.e. switch trials). This confirms that subjects were able to perform the task above chance levels. I also found a problematic typo in the text as I assume the numbers were reversed in the main result sentence: strategy switch trials should be 5.8 on average rather than 4.2 and vice versa for strategy maintain trials (not 5.8 but rather 4.2 trials on average). This is not a striking difference and from this data point alone one can wonder how well subjects performed the task. To go further, perhaps the authors could consider alternative behavioral analyses such as assessing when performance starts to reach significance after the reversal begins? Performance seems to be very good for trials 7-10. Perhaps more statistics on this dynamic reconfiguration could also help them demonstrate on average how participants perform and perhaps how model-based learning and model-free learning predict a different dynamic reconfiguration pattern? (instead of just summing the number of correct and incorrect trials after the reversal in a block of 10 trials; I suppose this is what was done but I could not find the information in the method). Could the author improve this section of the manuscript (methods and results) by putting more effort into characterizing their subjects' behavior during the task and also more interpretation in terms of the cognitive processes at play during the two types of learning strategy (free vs. model-based)?

We apologize for this mistake. The correct numbers are 4.2 for the keep strategy and 5.8 for the switch strategy.

Since we preformed new analyses on the behavior, we revised and reorganized corresponding results and methods sections including Fig. 3. We also removed previous Fig. 2a.

In the new analyses, as described in our reply to comment #1c, we provide a more detailed characterization of the behavioral adaptations using three parameters. Among these parameters, the switch offset *s* indicates the latency following the rule reversal before the participant reaches the chance level. As shown in the Fig. R3, the offset was 4.4 ± 1.6 (mean ± std), indicating that participants reached chance levels by the 4th trial following rule reversals. These new analyses provide a more fine-grained picture of the dynamic behavioral adaptations following rule reversals.

We also calculated the probability of repeating a decision, p(repeating), to detail how the feedback obtained on the current trial (trial n) impacts the choice in the subsequent trial (trial n+1). Based on the definition of model-free and model-based RL strategies, we now provide

new results that show that the two RL strategies jointly generate the observed different behavioral patterns, as revealed by p(repeating). For instance, the model-free free strategy could be captured by a simple reinforcement effect of rewarded outcomes: an ultimately rewarded choice is more likely to be repeated, regardless of whether that reward followed a common or rare transition. Conversely, a model-based strategy was predicted by a crossover interaction between reward and the transition probability, because a rare transition inverts the effect of the subsequent reward. These new results demonstrate that the participants used a mixture of both exploratory, model-free strategies and deterministic, model-based strategies.

We have revised the relevant sections of the paper to provide a more detailed characterization of the behavior. We thank the reviewer for this valuable comment, which has helped to strengthen the rationale for using RL strategies to describe participants' behavior.

e. Fig. 3b shows the mean learning curves separately for model-free and model-based trial blocks. Yet, there is no clear description of (i) how model-free and model-based blocks were distinguished (the method section on this topic is rather obscure and at least a statement of the type of behavioral signature distinguishing the two types of behavior should be provided) .

We apologize for the lack of clarity regarding the separation between model-free and model-based blocks. As mentioned in our response to earlier comments, we conducted additional behavioral analyses to demonstrate that participants employed a combination of both strategies.

We categorized the strategies on a block-wise basis. Given the probabilistic nature of our task, in some blocks, it may be challenging for participants to infer their current state (whether they are still in a steady state or whether they need to switch strategy). Consequently, they may tend to adopt a more model-free-like strategy to make decisions. In contrast, in other blocks, participants clearly inferred that the rule had changed and adjusted their strategy, persisting with their decisions even after unrewarded outcomes, indicating a model-based strategy. Following this logic, we separated blocks based on whether they were better explained by a model-free or model-based approach. To this end, we calculated the distance between human behavior (percentage of correct strategy) and simulations by model-free and model-based RL models. We then assigned blocks to either the model-free or model-based category for each participant by comparing the distances between them.

In the revise manuscript, we now clearly describe how we separated blocks into model-free and model-based (**Lines 274-278**):

> "Behavioral modeling indicated that our participants relied on a combination of model-free and model-based strategies to update their strategy (Fig. 1h). To explore this further, 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."

(ii) there is no statistical quantification of model-based versus model-free learning behavior which is strange since the curves seem to diverge significantly.

After separating blocks into those better explained by the "model-free" model and those better explained by the "model-based" model, we compared the differences in behavioral parameters

between them. Specifically, we compared the proportion of correct strategy, as well as the three regression parameters, switch offset s, slope α, and lapse ε.

The results show that the averaged proportion of correct strategy in model-based blocks is significantly higher than in model-free blocks (linear mixed-effect test, $t_{(62)} = 4.961$, $p = 5.77 \times 10^{-6}$) during the *Switch* period.

We additionally fitted logistic regression models to model-free and model-based blocks separately and calculated switch offset (s), slope (α), and lapse (ε) for each block type (Fig. R4). Subsequently, we compared these parameters between model-free and model-based blocks and found that the switch offset for model-free blocks was significantly longer than for model-based blocks (linear mixed-effect test, $t_{(62)} = 4.155$, $p = 0.0001$). Model-free blocks required more than 5 trials (mean ± std = 5.3 ± 2.2) to reach chance level after rule reversal, whereas model-based blocks required less than 4 trials (mean ± std = 3.7 ± 1.5). This difference in switch offset indicates that the model-free strategy exhibits more exploratory behavior following environmental changes. However, the slope and lapse did not differ significantly between model-free and model-based strategies (slope: $p = 0.65$, lapse: $p = 0.19$), suggesting that once the exploratory phase is completed, both strategies can adapt to the new environment in a timely manner and maintain a good performance level.

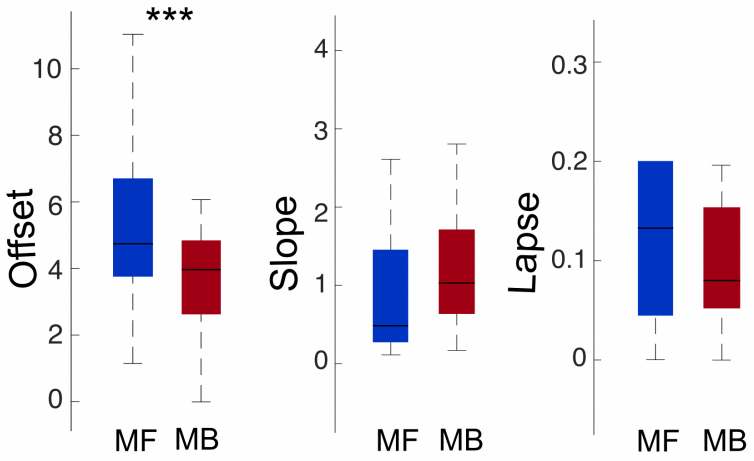


Fig. R4

These behavioral differences between model-free and model-based blocks align with our expectations and validate the rationale of separating the learning blocks, effectively capturing the essential characteristics of both strategies. We've added these new results to the revised manuscript (**Lines 280-289**) as follows:

> "Additionally, logistic regression models were separately fitted to MF and MB blocks to calculate the switch offset (*s*), slope (*α*), and lapse (*ϵ*) 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 5 trials (mean ± SD = 5.3 ± 2.2) to reach chance-level performance following a rule reversal, whereas MB blocks required fewer than 4 trials (mean ± SD = 3.7 ± 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 (p = 0.635) or lapse (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."

(iii) it appears to be a post-hoc identification of learning processes with faster learning sessions after the reversal being labeled as model-based and others (less successful sessions characterized by a slower learning rate being labeled as model-free sessions). There is no discussion of alternative interpretations of these behaviors? Could the authors demonstrate that subjects rely on distinct strategies rather than alternative scenarios such as sampling bias, behavioral noise, inter-individual learning performance, other RL architectures?

Combining our original and revised analyses, we found that participants employed a combination of different RL strategies to complete the task. We compared the empirical data from each block with simulated data, generated by model-free and model-based computational models. This approach allowed us to categorize the blocks based on their alignment with the predictions of these models, rather than relying solely on behavioral performance metrics. For instance, we did not simply designate blocks with faster learning rates as model-based and those with slower rates as model-free. Instead, blocks were classified according to the degree of correspondence between the empirical and simulated data.

Additionally, we analyzed the parameters derived from the new logistic regression model to validate the rationale for separating blocks. The behavioral differences we observed effectively capture the inherent characteristics of both model-based and model-free strategies. This finding suggests that the two block types reflect distinct strategic approaches, rather than being influenced by alternative factors such as sampling bias, behavioral noise, or inter-individual variability in learning performance.

However, we acknowledge that the blocks we categorized do not represent pure model-free or model-based strategies; instead, they are more closely aligned with the predictions of either the "model-free" or "model-based" models. In the discussion, we now explicitly highlight this point, emphasizing that it should be carefully considered when interpreting the results:

"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 the model-free or model-based models." (**Lines 509-511**)

(iv) the comparison between the two types of learning sessions using fixed-effect across blocks is problematic (for fig 4e-f, see below). Please consider using linear mixed effect to better demonstrate that effects can be seen across subjects (and avoid fixed effects across blocks generally since reliability of such effects across subjects cannot be assessed with such a statistical approach).

We appreciate the reviewer's suggestion to use the linear mixed effect to test the difference between the two types of learning strategies. We agree that this approach is particularly suitable, as it allows to model both fixed and random effects, thereby providing a more

accurate estimation of the differences between MF and MB while accounting for individual differences. Therefore, we re-analyzed the data to test the difference between MF and MB using the linear mixed-effect model (LMM).

LMM was implemented using the fitlme function in MATLAB. The model was specified using behavioral performance variables (i.e., the proportion of correct strategy, switch offset s, slope α, and lapse ε) as the response variable and block categories (MF and MB) as the fixed effect. The fixed effects coefficients indicate whether the difference between MF and MB is significant.

We found that the results remained unchanged with the fixed-effects test (Table R1), showing significant differences between MF and MB in proportion of correct strategies and the switch offset. We updated the results section accordingly.

Terms	Estimate	Standard error	t-value	p-value
Proportion of correct strategy	0.113	0.023	4.961	5.77×10^{-6}
Switch offset	1.479	0.356	4.155	0.0001
Slope	-0.451	0.944	-0.478	0.635
Lapse	0.020	0.015	1.357	0.180

Table R1

2. Neuroimaging results.

a. Rule reversals section (113-44; fig1): this section is very clear

Thanks

b. Representation of decision strategy (L146-89): the concept of “decision strategy” induced a lack of clarity: the authors looked at the average correct vs. incorrect number of trials following a reversal behaviorally (and they tagged this a “decision strategy” somehow artificially and post-hoc). I did not understand why they chose this terminology (for consistency across paragraph I would keep the stay vs. switch terminology as in fig 1 and I would add the layer correct vs. incorrect for this analysis that would simplify a lot the readability). The neural analyses here appear less robust and the rationale behind this analysis weaker compared to fig 1. I did not really catch the reason explain why the authors split the behavioral policy into their three categories: “overall strategy” just combines correct and incorrect post-reversal trial types, the “keep strategy trials” correspond to incorrect switch trials and “change strategy” captures correct switch trials. I found the chosen conceptual/terminology framework difficult to follow and it remains hard for me to understand for example how the neural activity of one brain regions can both represent two type of strategies (such as the dmPFC which is sensitive to both overall and change strategy);

c. Related to above: given the structure of the analysis and the RDMs shown in fig. 2b the significant effect of overall strategy could be simply driven by the strategy change (because the strategy change trials might drive the effect seen in the overall analysis)?

We used the term *Steady State* to refer to the period before the reversal (i.e., 10 trials prior to a reversal), while the *Switch* refers to the period after the reversal (i.e., 10 trials immediately following a reversal). In the first few trials of the *Switch* period, participants are unaware of the rule reversal, so they stick to the old strategy they learned during the preceding learning block (*Strategy_keep*). After they realize that the rule changed, they switch to the new strategy (*Strategy_change*). We agree that 'keep/change' strategy can also be defined as 'wrong/correct' strategy during *Switch*. In the revised manuscript, we replaced the terminology as suggested.

We also agree with the reviewer that one brain regions cannot represent the 'overall' strategy and the 'change' strategy at the same time. Statistically, the significant effect for the 'overall' strategy was only driven by the significant 'change' strategy. To avoid any future confusion, we removed the 'overall' strategy from RSA analyses. Our main conclusion is now that both dmPFC and MD represent the Correct Strategy (Switching), which suggests their critical roles in flexibly adapting to the new rule.

d. it might be informative to see the BOLD response related to the contrast between correct switch trials and STAY vs. incorrect switch trials and STAY (this might provide additional pieces of information to interpret the data provided by the RSA analysis;

We appreciate the reviewer's suggestion. We applied RSA to understand the representational structure of the different strategies in the three brain regions of interest. Compared to univariate GLM, multivariate RSA provides a richer, more comprehensive understanding of the spatial response patterns within brain regions capturing the complexity of neural representations.

We also followed the reviewer's suggestion by testing whether a univariate GLM comparison of correct strategy trials during the *Switch* period with those during the *Steady-State* period provides additional meaningful insights. We indeed found that both the MD and dmPFC were significantly involved when participants switched their strategy (correct strategy during *Switch*), with stronger activity within the MD (FWE small-volume correction, $p < 0.05$, Fig. R5). These results confirm the RSA findings showing that both dmPFC and MD represent the 'switch' strategy after the rule reversal.

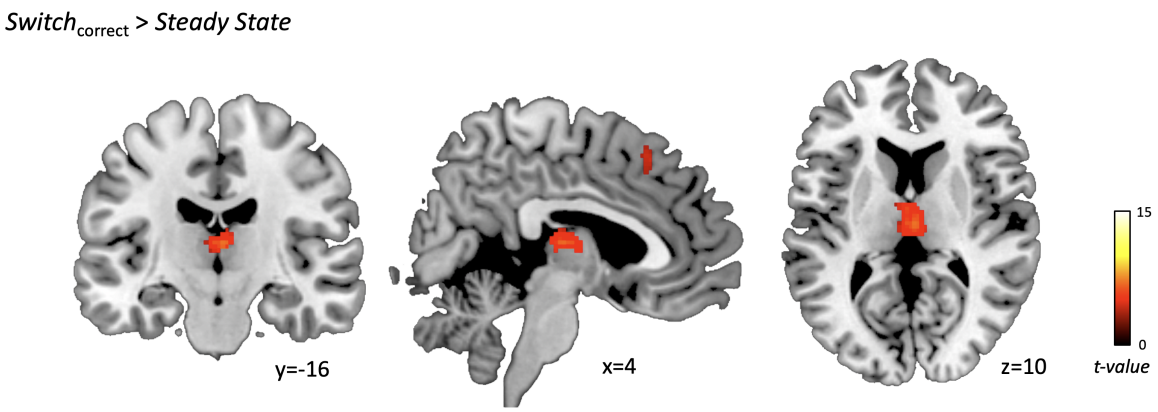


Fig. R5

e. could the author discuss/speculate on why the striatum would be active for the switch vs. stay contrast of fig 1 but not by the RSA?

The contrast between *Switch* and *Steady State* revealed neural responses related to rule reversals, which encompasses various aspects of strategy switching, including prediction, error-related processing, exploration, and, finally, the strategy updating/shifting. Our findings suggest that the three key regions—dmPFC, striatum, and MD—are involved in all these processes, either individually or through interactions.

Based on the correlation between the connectivity of the dmPFC and MD and the speed of strategy switching, we concluded that both regions may reflect the representation of different strategies. To test this, we applied RSA. Our results indeed revealed significant strategy representations within dmPFC and MD, but not striatum, which suggests that the striatum does not represent strategies during the *Switch* period as the other two regions.

In subsequent computational analyses, we found that the striatum is associated with processing reward prediction errors that occur when applying both, correct or wrong strategies. Taken together, the striatum seems involved in prediction-error related processes during *Switch*, but not in in representing correct/wrong strategies.

To address this point, we have included the following lines in the discussion section (**Lines 506-509):**

> "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."

f. Optogenetic preliminary data: two mice, 6 sessions. I honestly did not assess this part due to a low confidence on the replicability of such a small/preliminary data-set. The statistics reporting is also below the standards (no degrees of freedom, no specification of the kind of test done in the main result…).

We thank the reviewer for pointing out this concern.

In our study, the mouse data only serves as supplementary evidence to our main human results. That is why we decided to move the corresponding results, methods and figures to the supplement, extended by more information to improve clarity and replicability.

g. MD regulates RI representations in PFC or striatum (L209-260):

• The authors use fig 3b to illustrate that “subjects used a mixture of model-based and free strategy”; as indicated above, I found there was a circular reasoning there because they applied a RL model fitting procedure to identify two types of learning blocks in which subjects either poorly learn (slow learning = model free) or learned more rapidly to switch their choice to the more rewarding cue (less perseverative behavior following reversal). See my comment on behavior: there is currently a lack of support based on raw behavioral data that subjects did use two strategies (or the interpretation might be very different from using a model of the environment or not). It would be nice to see any data and new analysis supporting this claim (which is central to the paper).

In our revision, we present a range of new analyses that offer deeper insights into how participants developed and employed different strategies.

As detailed in our answer to comment #1a and b, we have conducted new analyses that incorporate transition probability alongside reward. In brief, participants applying the model-free RL relied solely on rewarded experiences from uncommon associations, increasing the likelihood of repeating their previous choice. In contrast, when using the model-based strategy, participants considered an internal model of the task's transition structure, evaluating actions prospectively, which leads to a decreased tendency to choose that same option. These new findings show that participants applied a mixture of both exploratory, model-free strategies and deterministic, inference-based model-based strategies, explaining their overall choice sequences (Fig. R2). These additional analyses offer behavioral evidence supporting the distinction between both RL strategies in our task.

• To better compare the RPE vs state-base PE it might be informative to plot RPE and SPE during either model-free or model based to see how much these two signals diverge within vs. between blocks?

We derived reward prediction errors (RPE) from the "model-free" blocks and state prediction errors (SPE) from the "model-based" blocks. These were subsequently utilized as parametric regressors in GLM analyses of fMRI data to identify associated neural signals.

In response to the reviewer's comment, we also plotted the RPE from the "model-based" blocks and the SPE from the "model-free" blocks to compare the differences in these signals across the two block types. Our analysis revealed no significant difference in RPE suggesting that RPE remains consistent regardless of whether learning is driven by a simple, direct reinforcement signal (model-free) or a more complex, goal-directed strategy incorporating task structure knowledge (model-based). In contrast, the SPE was significantly higher in the "model-based" blocks compared to the "model-free" blocks (linear mixed-effects test, $t(62) = 3.10$, $p = 0.003$, see Fig. R6), indicating that discrepancies between predicted and actual environmental states are more pronounced after rule reversals when participants engage in model-based learning.

This may reflect the increased complexity and cognitive demands of model-based learning, which requires individuals to actively maintain and update a mental representation of the task structure to guide informed decision-making.

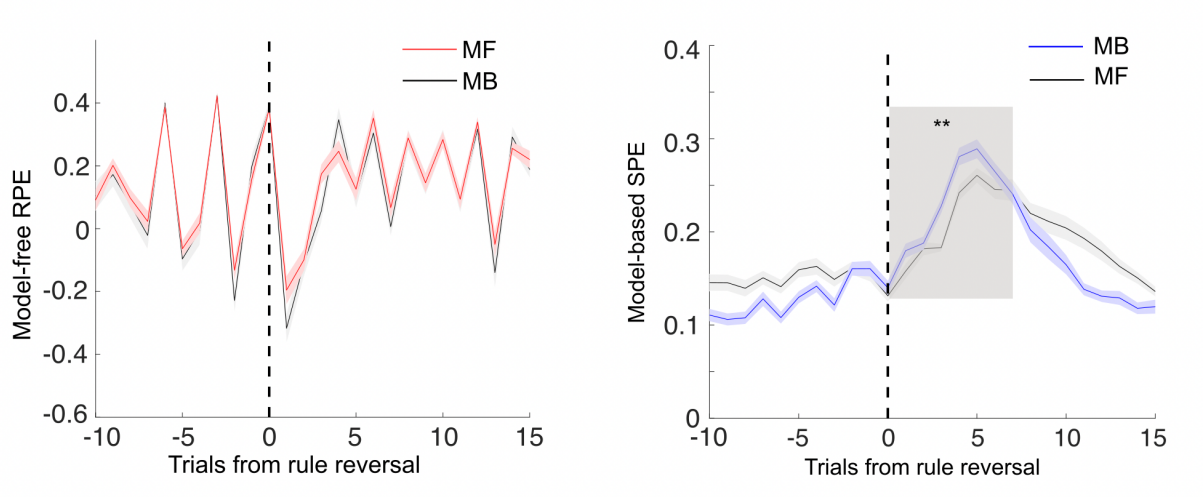


Fig. R6

In the revised manuscript, we included these results in the supplementary materials to provide additional evidence supporting the RPE and SPE across different block types.

• Perhaps the authors could simply do more to characterize the distinction between free vs. model-based blocks using behavioral metrics for which we would see that the data of the model-free blocks cannot be predicted by the model-based RL architecture and vice-versa: the model-based behavioral results cannot be predicted by the model-free RL architecture: this would be more solid evidence that there was something in the data that can be better captures by a model rather than the other…

In response to the earlier comments, we conducted additional analyses on behavioral performance, examining the differences between model-free and model-based blocks in terms of the probability of repeating the previous choice and the switch offset. Please also refer to our replies to the reviewer's comments 1a-d.

That said, I found the following fMRI, tractorgraphy and PPi analyses very nice and clearly reported (Fig 3: neural data part).

Thanks.

h. Network analyses (262-344): I was surprised to see that part of the figure 4 (based on network based results) re-analysed RSA results (presented in figure 2; would'nt it be simpler to add the RSA layer (that simply split model free vs. model-based blocks) to figure 4 ?

Indeed, we adopted this approach by separating the model-free and model-based blocks and conducting the RSA analysis independently for each. However, for Figure 4, our additional work focused primarily on three PFC regions—the dmPFC, dlPFC, and OFC—rather than the original three regions of interest (dmPFC, caudate, and MD). This shift was guided by the findings from the DCM analysis and neural network modeling, aimed at demonstrating the distinct overwriting patterns of PFC representations.

DISCUSSION! I found the results were very briefly discussed and they were not even positioned relative to previous literature on the involvement of thalamic nuclei during RL

processes. The dorsomesial thalamus was shown to be involved during a two-arms bandit tasks with direct electrophysiological recordings from humans (Collomb-Clerc et al. Nature comms. 2023). I guess the very short discussion might be due to an initial submission as a brief report. An expanded discussion section would be of interest for a revised manuscript.

We have expanded the discussion section to include evidence highlighting the involvement of the MD in reinforcement learning processes. Thank you for suggesting the reference.

In summary, I found the data analyses to be remarkable for the fMRI dataset and the modeling effort to be quite impressive as well. However, this came at the cost of a clear behavioral dissociation between the two learning strategies that are explored here and I felt that the impressive post-hoc effort did not outweigh the weakness of the behavioral dissociations between the two types of models: model-free and model-based predictions are indeed never directly compared, so it remains to be demonstrated whether and how RL behavior following reversal in this task can be used to assess both forms of learning strategy, unlike the initial results from 2011.

We sincerely appreciate the reviewer's valuable feedback on our manuscript, which has been instrumental in refining and validating our findings. In response, we conducted additional analyses to explore the behavioral dissociations between the two types of RL models. These analyses suggest that individual participants likely employed a combination of exploratory, model-free strategies and deterministic, inference-based behaviors, accounting for the variability observed in their choice sequences. Additionally, we compared behavioral performance measures, such as the probability of repeating a choice and the switch offset between the two strategies, highlighting the key differences in their characteristics.

We hope that the revisions to the manuscript will align with the reviewer's expectations and further strengthen their confidence in our work.

Minor comment:

Abstract: I found the term multimodal human imaging misleading (since all the relevant data come from fMRI and the dTI data-set is not central) perhaps just state fMRI and DTI techniques; please reformulate

This has been addressed in the revised manuscript.

Reviewer #2 (Remarks to the Author):

Wang and colleagues report the results of a study designed to assess the role of the mediodorsal region of the thalamus in regulating learning and flexible behavioral updating. The authors describe results detailing separate thalamocortical pathways for model-free versus model-based control of behavior and find evidence that distinct regions within MD support each process. Moreover, they found that MD-prefrontal pathways were engaged during strategy updating, which biophysical modeling suggest supports rapid updating of context representation enabling flexible behavior.

On the one hand, I am enthusiastic about the importance of the question targeted by the manuscript (which has been understudied and is important for understanding basic cognition and circuit dysfunction in disease) and am impressed with the methodological breadth of the manuscript. The combination of fMRI methods (univariate, multivariate, and dynamic causal modeling), DWI, confirmation of the importance of this pathway using optogenetic silencing in rodents, and biophysical modeling of the circuit is not only impressive but provides synergistic and converging sources of evidence about the distinct roles of medial and lateral MD in RL. This approach reflects my hope for the future of cognitive neuroscience.

On the other hand, I found this manuscript difficult to follow, even after multiple read-throughs, and lack of clarity often make it hard or impossible to evaluate the authors' claims. As a result, there were some key claims (particularly in the RSA, PPI, and biophysical modeling sections) that I could not evaluate. In other sections, I have significant concerns about the analyses.

-On the clarity point, there are many points in the Results where claims are made without any explanation of what exactly was tested, or even what the statistical results were. This problem is especially pronounced in the subsection "Thalamocortical pathways implements the critical computations for reinforcement learning during strategy updating", which contains no statistical results in the main text. Some of these results are described in the Figure (a pattern repeated throughout the paper), but I found myself needing to switch between the Figure legends, Methods, and Supplement to try to understand a result described in the Results. I will just list a few illustrative examples rather than describing them all, but I advise a careful re-write of most of the Results and a significant portion of the Methods sections:

We sincerely appreciate the reviewer's recognition of our work. The reviewer's critiques have been invaluable in improving the manuscript, particularly in addressing challenges related to the narrative flow and the lack of detail in the results description. In response, we have included additional statistical results and reorganized the figures to enhance the clarity of our findings, with particular focus on the computational model section. Additionally, we thoroughly reviewed the manuscript to ensure that the presentation of results is both comprehensive and transparent. We have also revised the Methods section to improve its readability, making it more accessible and straightforward for readers.

a. "Using bilinear dynamic causal modelling (DCM), we observed recurrent connectivity of MD with both dmPFC and CN, significantly modulated by rule reversals (posterior probability = 0.6)." What does that probability correspond to? A posterior probability of .6 does not seem very high to me to claim an effect? Additionally, readers unfamiliar with DCM may not understand what this connectivity signifies.

In DCM analysis, posterior probability represents the estimated likelihood of a model's parameters or structure given the observed data and the model itself. As a cornerstone of

Bayesian inference, it reflects the degree of belief in the model parameters or structure after incorporating the observed data. Posterior probability is commonly used to evaluate the relative support for different models. For example, when comparing two models, the one with a higher posterior probability is considered more likely to align with the observed data. Thus, the posterior probability indicates the extent to which a model is supported by the data in the context of the models being compared. However, it is important to emphasize that the highest posterior probability does not guarantee that the model is the true one. Rather, it signifies that this model is the most plausible among the options considered, given the observed data.

In our analysis, the posterior probability of the winning model is 0.6, exceeding the combined probabilities of all other models. The second-best model has a posterior probability of only 0.27. Therefore, we can confidently conclude that the model incorporating modulation of the recurrent connectivity of the MD with both the dmPFC and CN during reversals provides a superior explanation of the observed data compared to the other 14 models.

In the revised manuscript, we rephrased the corresponding section and included additional information to clarify the concept of effective connectivity in DCM, as detailed in **Lines 195-200**:

> "To investigate the causal interactions among these three key 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 modulation of recurrent connectivity between the MD and both the dmPFC and CN during rule reversals provided the best explanation for the observed evoked activity in these regions (posterior probability = 0.6, Fig. 2b and Supplementary Fig. 2)."

b. The main text says: "Consistent with this notion, functional connectivity from the MD to dmPFC predicted switching speed, ($r = -0.515$, $p = 0.008$)." It is not clear if this is "effective connectivity" from the DCM, and whether this the baseline connectivity or the modulation by rule reversals. It is also unclear what switching speed is. The figure clarifies that switching speed is the number of trials to switch, but also describes multiple comparisons corrections, which refer to additional tests described only in the Supplement. As an additional note about this analysis, I am concerned about the influence of leverage points, especially with this sample size, and the authors should conduct additional control analysis using a nonparametric or outlier-deweighting method.

In this analysis, the effective connectivity, correlated with the behavioral parameters, corresponds to the modulatory effect of reversals (matrix-B in DCM). To assess this relationship, we employed Spearman correlation, a non-parametric measure, instead of Pearson correlation, which assumes linear relationship. Spearman correlation does not rely on data distribution and is less sensitive to the influence of outliers, making it a robust choice for this analysis.

Initially, we defined switch speed as the time (in trials) it took for the average proportion of correct strategies to reach its minimum point following a reversal. However, we acknowledge that this measure is arbitrary and inadequate for accurately capturing behavioral adaptation to new rules following reversals, as this process undoubtedly involves more than a single point or trial. To better characterize the dynamics of behavioral adaptation, we conducted additional analyses focused on post-reversal adjustments. Specifically, we calculated three parameters

by fitting the data to a logistic regression model: the switch offset s, slope α, and lapse rate ϵ (Fig. R7). The switch offset s quantifies the latency of the switch, the slope α reflects the sharpness of the transition, and the lapse rate ϵ represents the level of behavioral performance after the transition.

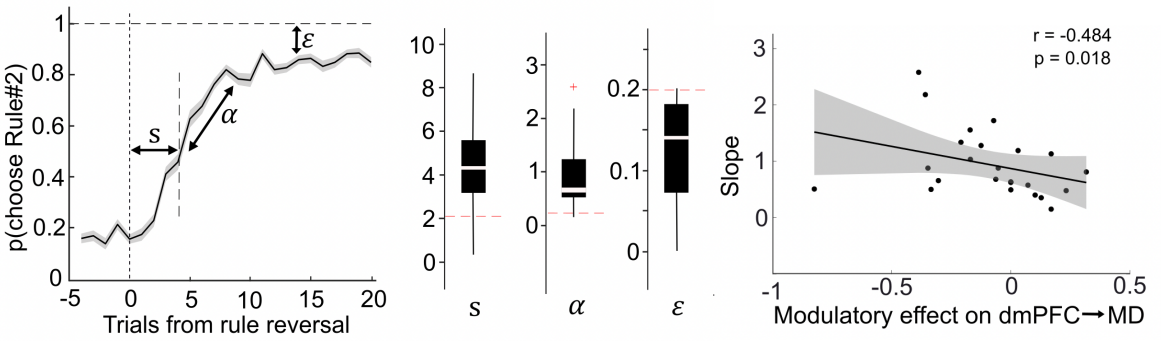


Fig. R7

Next, we examined the correlations between these three parameters as well as their modulatory effects on the recurrent connections between the MD and dmPFC, and between the MD and CN. Our analysis revealed a significant Spearman correlation between the slope α and the modulatory effect of reversals on the connectivity from the dmPFC to the MD (r = -0.484, p = 0.018).

In the revised manuscript, we have clearly detailed the specifics of this correlation analysis as follows (**Lines 200-204**):

> "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). This indicates that a more gradual transition—characterized by increased exploratory behavior before reaching a learned state—requires stronger modulation of dmPFC-to-MD connectivity."

c. Line 205 has no statistics associated with the result

> We have now included the statistics for this comparison, as detailed below:
>
> "…However, during the feedback period in the first eight trials after a rule switch, MD silencing significantly delayed switching behavior ($p<0.05$, Supplementary Fig. 3)."

d. The methods underlying the PPI analyses are split into three different sections in of Methods, making them extremely difficult to follow and evaluate.

We apologize for any inconvenience caused by the separation of the methods, such as for the PPI analysis, which have made the manuscript more challenging to follow.

We have consolidated the various methodological sections, including DCM, PPI, and others, into a single, unified section. This reorganization is intended to provide a more cohesive and streamlined presentation of our methods, enhancing the overall clarity and readability of the manuscript. Please refer to the updated manuscript on pages 32–38 for these revisions.

e. What does it mean for the biophysical model to be capable of solving the task? How is that shown or evaluated?

The CogLink networks used in this study represent a novel class of biologically plausible mechanistic models of forebrain networks designed to solve complex tasks. In this context, the model's ability to solve complex tasks refers to its capacity to effectively simulate and replicate decision-making processes and behaviors observed in real-world scenarios. The model's performance is quantitatively evaluated using metrics like accuracy. Specifically, for our probabilistic reversal learning task, we mean that the model must achieve high accuracy before the reversal, indicating that it has successfully learned within a context, and recover to high accuracy before the block ends, demonstrating its ability to adapt after the reversal. We define high accuracy as reaching 80%. As shown in Fig. R8, our model successfully solved the probability reversal task with a 80% accuracy before the reversal and before the block ended.

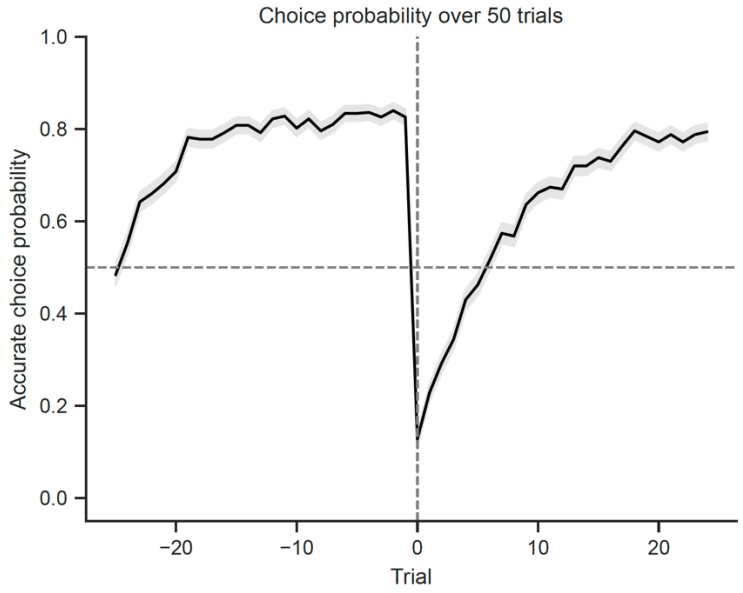


Fig. R8

We made the relevant revision below in the revised manuscript (**Lines 370–372**):

> "Our model successfully solved the probability reversal task, achieving at 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; $n = 500$)."

f. Is figure 4e human or model data?

We apologize for any confusion regarding Figure 4e. Figure 4e presents model data, which is the Fig. 6d in the revised manuscript. Its purpose is to demonstrate that switch times in model-free (MF) blocks are longer than those in model-based (MB) blocks, mirroring behavioral patterns observed in humans.

To prevent any future misunderstandings, we have explicitly clarified this information in the revised manuscript as below (**Lines 374-377**):

> "To further assess this relationship, we asked whether MDl activity can be used to distinguish between MB and MF strategies that the model may exhibit. Indeed, this approach reproduced the behavioral differences observed in human subjects: switch times were longer in MDl activity-derived MF blocks compared to MB blocks (Fig. 6d)."

-SPE exhibits primarily slow fluctuations that are likely removed by the author's bandpass filter. What does the SPE regressor look like if filtered in the same way as the data, and does this correspond conceptually to SPE? Additionally, the model-based model is unclear to me: do the SPEs modulate the RPEs and learning process, or is SPE a separate process?

We did not apply a bandpass filter to the SPE; the plot is presented in its raw form.

In the reinforcement learning framework, the model-based (MB) system constructs an internal model of state transitions during the decision-making process, incorporating state transition relationships (state prediction error, SPE) to guide behavioral choices. In contrast, the model-free (MF) system relies on reward prediction error (RPE) to learn the value of different actions, drawing exclusively from prior experience with specific stimulus–response associations.

SPE and RPE play distinct roles in the learning processes of model-based and model-free systems, respectively. SPE is integral to model-based learning, as it detects discrepancies between predicted and actual state transitions, enabling the refinement of the internal model of the environment. This process is essential for planning and making decisions based on the updated model. In contrast, RPE is central to model-free learning, where it updates the value of actions by capturing the difference between expected and actual rewards. Model-free learning relies on direct experience and trial-and-error, without constructing an explicit representation of the environment.

Although SPE and RPE are distinct signals, they can interact in complex ways. For instance, SPE can enhance the learning process by providing additional information about the dynamics of the environment, which may indirectly influence reward estimation and the calculation of RPE. Despite this interplay, the two signals are processed in different neural regions: SPE is primarily associated with the intraparietal sulcus and lateral prefrontal cortex, while RPE is linked to the ventral striatum (Gläscher et. al., 2010, Neuron). Thus, while SPE and RPE can influence one another, they remain fundamentally distinct processes that contribute to separate aspects of learning within reinforcement learning systems.

In the revised manuscript, we have added more information and clearly outlined the distinct roles of SPE and RPE (**Lines 687–695**).

> "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. 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."

-In the strategy RDM analysis, overall strategy is colinear with strategy-keep and strategy-change. Additional control analyses are needed to confirm that significant overall-strategy effects in dmPFC are just driven by the strategy-change effect. The results should also make clear this was calculated during the switch period. After reading the Methods, it is unclear how the group statistical tests were conducted for these analyses.

We appreciate the reviewer for highlighting this point. We agree that the overall strategy could be influenced by strategy changes. In light of this, and considering the comment raised by the first reviewer, we have decided to exclude the overall strategy from the RSA analysis to prevent any confusion regarding the interpretation of the results.

Additionally, in the revised manuscript, we have explicitly stated that this analysis was conducted on the *Switch* period. The statistical procedures for the RSA have also been clearly detailed. Please refer to the revised text in the manuscript.

> "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." (**Lines 771–774**).

> "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 Wilcoxon signed-rank 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)." (**Lines 801-808**)

-Many of the results rely on a median splitting of so-called model-based and model-free blocks. This was surprising given the other results in the paper describing co-active model-based and model-free strategies. The authors should show that these blocks are characterized by different behavioral profiles to justify separately probing their neural covariates.

We appreciate the reviewer for highlighting this important point. To better understand behavioral profiles, we conducted additional analyses to provide evidence and a-priori justification that the probabilistic reversal learning task inherently requires different RL-based strategies, resulting in variations in behavioral performance across blocks and participants. Below, we briefly summarize the steps we took and the corresponding results.

First, we analyzed the probability of repeating decisions to distinguish between model-based and model-free reinforcement learning (RL) strategies based on how feedback from the current trial (trial n) influenced choices in subsequent trials (trial n+1). In a model-free RL strategy, a rewarded experience from an uncommon association would simply increase the probability of repeating the previous choice. In contrast, employing a model-based RL strategy, which incorporates the task's transition structure to evaluate actions prospectively,

would exhibit a decreased tendency to choose the same option following such feedback. This model-based strategy, which accounts for the transition model in decision-making, is revealed through an interaction between reward and transition probability.

For both *Steady-State* (SS) and *Switch* (SW) transitions, the probability of repeating decisions (p(repeating)) revealed significant effects indicative of both strategies. Specifically, we observed a significant main effect of reward (i.e., model-free strategy) and a significant interaction between reward and transition probability (i.e., model-based strategy) (SS: main effect of reward: $F_{(1,31)}=116$, $p=5.2\times10^{-12}$, interaction: $F_{(1,31)}=64$, $p=4.7\times10^{-9}$; SW: main effect of reward: $F_{(1,31)}=201$, $p=4.1\times10^{-15}$, interaction: $F_{(1,31)}=23$, $p=3.4\times10^{-5}$. These findings suggest that participants employed a mixture of exploratory, model-free strategies and deterministic, model-based behaviors, accounting for the observed patterns in their choice sequences.

As part of a post-hoc analysis to dissociate model-free and model-based blocks, we compared behavioral parameters between the two. Specifically, we examined (1) the proportion of correct strategy, (2) the probability of repeating previous decisions, and (3) regression parameters including switch offset (s), slope (α), and lapse (ϵ).

The results revealed that the average proportion of correct strategy was significantly higher in model-based blocks compared to model-free blocks ($t_{(31)}=4.88$,$p=3.0\times10^{-5}$). The probability of repeating a decision was also higher in model-based blocks, regardless of whether the previous trial was rewarded or unrewarded (rewarded: $t_{(31)}=8.21$,$p=2.86\times10^{-9}$; unrewarded: $t_{(31)}=3.14$,$p=0.0037$; Fig. R4). This finding supports the concept that model-based strategies incorporate the transition model into decision-making rather than being solely influenced by past outcomes.

Additionally, we observed that the switch offset (s) was significantly longer for model-free blocks compared to model-based blocks ($t_{(31)}=4.09$,$p=2.85\times10^{-4}$). Model-free blocks required an average of more than 5 trials (5.3±2.2) to reach chance performance after a rule reversal, whereas model-based blocks required fewer than 4 trials (3.8±1.5).

These behavioral differences between model-free and model-based blocks align with our expectations and validate the rationale for separating the blocks, effectively capturing the core characteristics of both strategies. These results have been added to the revised manuscript; please refer to revised Fig. 1 and Fig. 4a/4b.

-The 10 trial definition of switch and steady-state periods needs more explanation. How did the authors' come up with this number? Additionally, if this was based on looking at the data, the analyses comparing behavioral performance in these periods are circular and should be presented in a way that makes that clear.

To objectively define the *Switch* period, we determined the switch offset (time from the reversal point to the point with behavior reaching the chance-level) for each participant. To this end, we calculated the average probability of choosing the correct strategy after the reversals across the 12 blocks (see Fig. R7 for detailed results and our response to comment #b). We then fitted a logistic regression model to individual choice data, estimating three parameters. Among these, the switch offset (s) measures the latency of the switch, indicating the number of trials required for participants to reach the chance performance levels.

As shown in Fig. R7, the mean switch offset was approximately 4 trials (mean±std=4.4±1.6), suggesting that participants typically reached chance performance by the 4th trial after rule reversals. Across participants, the offset ranged from 1 to 9 trials, which validates the appropriateness of our 10-trial window. This window effectively captures both the exploratory phase and the successful transition to the new rule content within the defined period.

To further validate the definition of the *Switch* period following rule reversals, we calculated the participants' behavioral performance. As illustrated in Figure 1e, the proportion of correct strategy choices stabilizes after the 10th trial post-reversal, providing a rationale for selecting this 10-trial window. Moreover, we compared behavioral performance between the *Steady State* and *Switch* periods. This comparison confirmed that the 10-trial window effectively distinguishes these two distinct phases of the task.

We acknowledge the reviewer's concern regarding the potential for circularity in our analyses. To address this, we have reorganized the results presentation and ensured that the definition of the *Switch* period is based on independent criteria and is not derived from the same data used for performance comparisons. Specifically, the logistic regression analysis provides an objective measure of switch latency, which serves to validate the 10-trial window rather than directly defining it from the performance data. This approach ensures that our analyses are free from circular reasoning and grounded in robust, independent evidence. Then we validate our definition by the comparison of the two periods.

We have included more detailed information and clearly outlined the evidence supporting the definition of the *Switch* period in the revised manuscript:

> "To characterize the dynamics of behavioral adaptation in response to the rule reversals, we first calculated the average probability of selecting the correct strategy (rule 2) 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 (α), and lapse rate (ε). Specifically, the switch offset (s) measures the latency of the switch, the slope (α) quantifies the sharpness of the transition, and the lapse rate (ε) reflects behavioral performance after the transition. For all participants (n = 32), we computed the switch offset (s), slope (α), and lapse rate (ε) (Fig. 1f). The mean switch offset (s) was approximately 4 (mean ± SD = 4.4 ± 1.6), indicating that participants reached chance-level performance by approximately the fourth trial following rule reversals. In addition, the slope (α) was 1.3 ± 2.6, and the lapse rate (ε) 0.13 ± 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 strategy 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 s (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 10 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 ($t(31)=13.20$, $p < 0.0001$, two-tailed, Fig. 1f). ”

-The methods for the model-free and model-based RDM analysis are not clear to me. I follow the section of the Methods describing the GLM construction, but I did not find a description of the RDM analysis and therefore do not understand how the authors found that “we found that model-based MD engagement was localized to its lateral subdivision, whereas the model-free to its medial one”.

We apologize for the unclear description of the model-free and model-based RDMs.

The differential involvement of the lateral and medial MD in model-free and model-based blocks was not derived from the model-free and model-based RSA analysis. Instead, this conclusion was based on overlaying model-free (RPE) and model-based (SPE) MD activations obtained from the parametric GLM analysis.

To further validate this functional separation, we performed an RSA analysis to test whether the spatial response patterns of the MD during model-free and model-based blocks are distinct. Specifically, we created an MD mask based on the contrast of ‘*Switch*’ vs. ‘*Steady State*’ ($p<0.001$). Using unsmoothed t-statistic maps, we extracted the activity patterns from the MD separately for model-free and model-based blocks.

The relative similarity between the response patterns in model-free and model-based blocks was assessed using Pearson correlations, expressed as correlation coefficients. To quantify the strength of this similarity, we compared it to similarity values derived from random data generated by shuffling voxel sequences. As random data represent entirely unrelated and highly dissimilar patterns, a finding that the MD response patterns exhibit similar levels of dissimilarity to the random data would suggest that the response patterns for model-free and model-based MD are entirely unrelated.

Our results showed no significant difference in representational dissimilarity between the MD response patterns in model-free and model-based blocks (Fig. R9b and c), confirming that the response patterns in the MD for these two conditions are distinct.

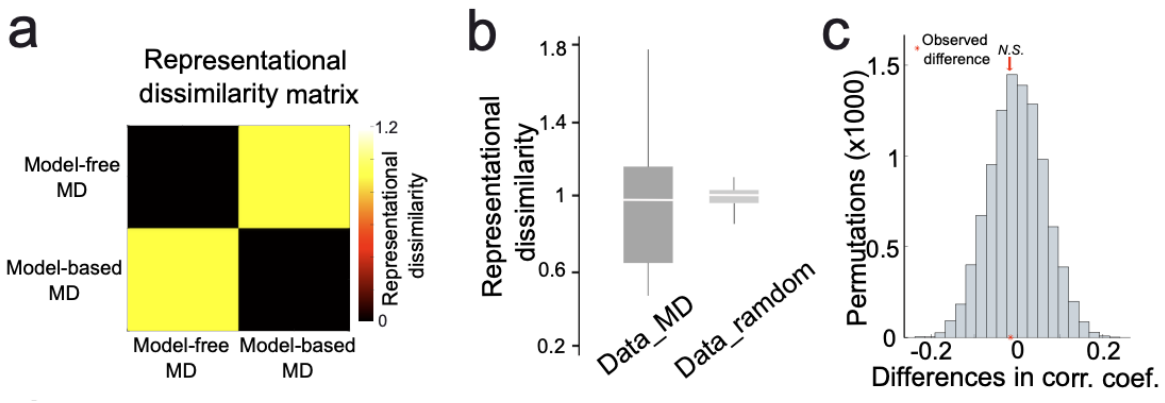


Fig. R9

Additionally, we hypothesized that the response patterns within the model-free (MF) or model-based (MB) MD would exhibit low dissimilarity, whereas the dissimilarity between MF-related and MB-related MD would be high if their response patterns are indeed distinct. To test this hypothesis, we constructed RDMs based on the predicted correlation distances between MF and MB blocks (see Fig. R10a).

For this analysis, we divided the MF and MB blocks into two halves (MD_MF1/MD_MF2 and MD_MB1/MD_MB2) and extracted the response patterns of the MD from these four contexts. The resulting data RDM, derived from the MD masks, is shown in Fig. R10b.

A permutation test revealed that the dissimilarity between the MD activity patterns corresponding to different RL strategies was significantly higher (Fig. R10c). This result provides robust evidence supporting the distinct response patterns of the MF and MB MD.

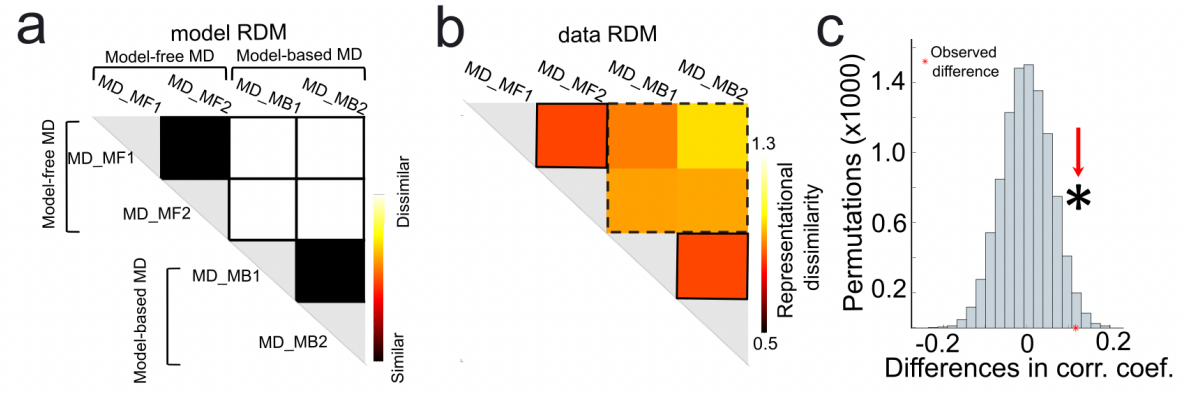


Fig. R10

We have included additional details about these two supplementary RSA analyses to improve clarity and prevent any potential confusion.

-The biophysical model predictions and results are insufficiently explained. For example, I am not sure what Figure 4h depicts. The texts in the results do not describe this result in much detail, stating that switching requires overwriting of previous OFC mappings. How do the

action values on the y-axis reflect OFC mappings? Where is evidence of over-writing in this plot? I used this inset as an example, but I was similarly confused for the other analyses in this figure.

We thank the reviewer for raising this important point. We have expanded the explanation to clarify how the model results lead to neural and behavioral predictions and explicitly address the evidence of overwriting in OFC mappings.

Regarding Figure 4h (Fig. 6g in the revised manuscript), the action values on the y-axis represent the learned value associated with each action in a given context, which is computed based on OFC activity. Overwriting occurs in MF blocks because, in the absence of an alternative contextual representation in MDl, the same OFC population that was engaged before the switch remains active. This results in a shift in action values, as the Go-tuned and NoGo-tuned populations switch their relative activity levels.

The following revision provides additional details on how these mechanisms are reflected in our model results (**Lines 382-393**):

> "Our CogLink network enabled us to investigate the neural mechanisms underlying MB and MF switching which were not accessible by fMRI alone. Analysis of MDl 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. 6f). In the CogLink network, distinct contextual populations of MDl neurons selectively modulate disjoint populations in the orbitofrontal cortex (OFC), regulating both their activity and plasticity (see Methods; Supplementary Fig. 7a, b)17. 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. 7a, b). As a result, switching behavior relied on overwriting the existing mapping in the same OFC population. Specifically, in Fig. 6g, we observed a "crossing" 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."

Similarly, we have expanded our explanation of why MDl fails to form an alternative representation in MF blocks and included key behavioral predictions that validate this hypothesis. These revisions clarify how the model provides a mechanistic explanation for the neural basis of different strategies, highlighting that model-free strategies reflect a temporary limitation rather than a fundamentally separate process.

Additionally, we have carefully revised the text throughout and Figure 6 to ensure that all analyses are clearly explained. We hope these clarifications fully address the reviewer's concerns.

-If the optogenetics results remain in the main text, the key results should be in a figure in the main text.

In our study, the mouse data serves as supplementary evidence supporting our human main results. As such, we have chosen to present the mouse data, including the results, methods,

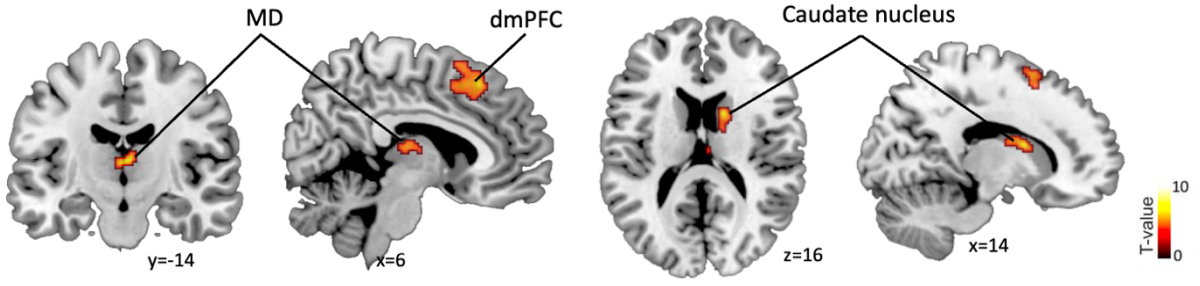
and figures, as supplementary material to provide additional context rather than including it in the main text. However, we have retained a brief introduction to the mouse results in the main text, following the fMRI findings from humans, as outlined in **lines 206–216.**

> "While the human fMRI data demonstrated correlations between neural responses and behavioral/experimental characterizations, they could not establish causality. To directly test the MD's role in decision strategy updating, we used a mouse model to examine the relationship between MD function and behavioral adaptation during a rule reversal task (Supplementary Fig. 3). Optogenetic silencing of the MD during the feedback period had no significant effect on mice's performance during *Steady State* trials (n=6 sessions from two mice, $p>0.05$). However, during the feedback period in the first eight trials after a rule switch, MD silencing significantly delayed switching behavior ($p<0.05$). These results provide causal evidence that the MD is specifically required during the feedback period to facilitate behavioral adjustments following rule reversals. These findings 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."

-Why are the plots presented as uncorrected tmaps for display purposes? In principle this is OK if justified (which is needed), if the key regions survive FWE correction, would these plots be different if only corrected data were shown?

In the main text, we reported significant outcomes following FWE correction, including details such as peak voxels, p-values, and t-values. However, for display purposes, we chose to present figures using uncorrected t-maps with a threshold of $p<0.001$. This approach provides a clearer and more visually pronounced representation, as the lower threshold highlights the broader patterns in the data. In contrast, FWE-corrected results, while statistically rigorous, often involve only a small number of voxels, making them less visually prominent. Using uncorrected t-maps for visualization is a common practice, especially when corrected results involve only a limited number of voxels.

Below, we provide a comparison of uncorrected (upper) and corrected (lower) figures (Fig. R11), using the results of the contrast "*Switch* > *Steady State*" as an example:



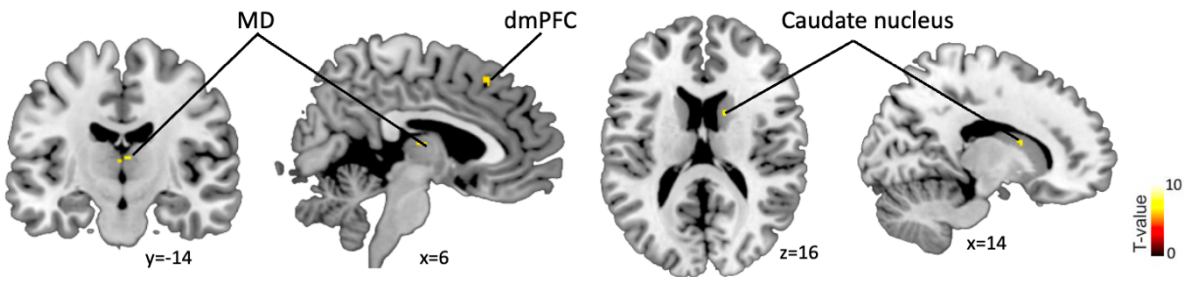


Fig. R11

-More detail and justification is needed for the subjects excluded for poor behavior

The exclusion criteria were based on an average proportion of correct responses below 60% and/or the absence of clear learning and reversal effects in their performance. These criteria indicate that the participant failed to learn the association between cues and responses and did not detect the reversal points. Fig. R12 below illustrates an example of a participant who was excluded due to poor performance:

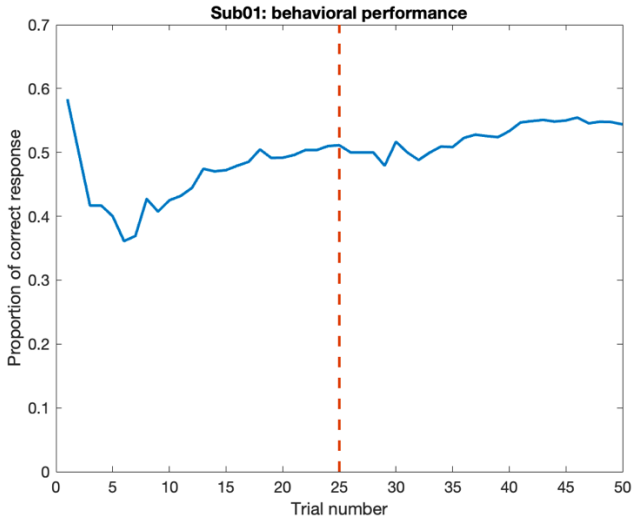


Fig. R12

We have now included the criteria in the updated manuscript as follows (**Lines 574–580**):

> "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 ± SD: 24.5 ± 3.5 years)."

-How were the ROIs defined?

We defined regions of interest (ROIs) for multiple analyses, including DCM analyses, RSA analyses, parametric modulation GLM analysis, and PPI analysis.

In our study, two DCM analyses (Fig. 2b and Fig. 5a in the revised manuscript) were conducted. The first DCM (Fig. 2b) used the dmPFC, striatum, and MD as ROIs. These ROIs were derived from the group peak coordinates identified by the contrast of *Switch* > *Steady State* in the univariate fMRI analysis. The subject-specific ROIs were defined by the overlap of two images: (1) a 12 mm radius sphere centered on the group peak coordinates and, (2) all voxels surviving p=0.05, uncorrected, within a 6 mm sphere centered on the individual peak voxel.

In the second DCM (Fig. 5a), the ROIs included the lateral and medial MDs, dlPFC, and OFC. Unlike the first DCM, the MD ROIs were defined using the AAL3 atlas to better separate signals from the two MD subregions. The dlPFC and OFC ROIs were derived from the results of the PPI analysis, which showed significant functional connectivity with the MDs. Time series were extracted for these ROIs using the same method as in the first DCM.

Similarly, two RSA analyses (Fig. 3 and Fig. 6h) utilized five ROIs in total to extract time series and construct RDMs. In the first RSA (Fig. 3), the dmPFC, striatum, and MD were used as ROIs, derived from the *Switch* > *Steady State* contrast in the univariate fMRI analysis, with a threshold of $p<0.001$, uncorrected. In the second RSA (Fig. 6h), three ROIs were used: the dmPFC (the same as in the first RSA), and the dlPFC and OFC, derived from PPI results with the same threshold ($p<0.001$, uncorrected).

To examine the modulatory effects of model-free and model-based RL parameters (RPE and SPE) on brain regions active during the *Switch* period (*Switch* > *Steady State*), we created an ROI mask consisting of the dmPFC, caudate, and MD. These regions were defined using the *Switch* > *Steady State* contrast from the univariate fMRI analysis with a threshold of $p<0.001$, uncorrected. Results were restricted to this ROI, and small-volume correction was performed within it.

Lastly, PPI analysis was conducted to identify PFC regions connected with the MDs. For this, a PFC ROI was defined again using the AAL3 atlas. Results were restricted to this PFC ROI, and small-volume correction was applied within the ROI using an FWE-corrected peak-level threshold of $p<0.05$.

In summary, several ROIs were defined based on our results and hypotheses. We have now added the missing details about ROI definitions to ensure transparency.

Reviewer #3 (Remarks to the Author):

The authors, in the current manuscript, are building on a “serendipitous finding” that the MD thalamus robustly activates upon strategy updating and encoding in a probabilistic reversal learning task. Through a number of approaches, including univariate, RSA, and connectivity-based approaches, the authors build a compelling idea for a role of. These findings suggest again that the MD is positioned between high-level prefrontal and low-level striatal representations, which underlies the modulation of frontal cortico-striatal networks during Switch period.

High Level Comments

Have to say, I’m quite happy to see the regional analysis MD in this analysis – the idea of viewing it as monolithic (as often is an inadvertent side effect of fMRI limitations) seems quite at odds with neuroanatomy, and here it appears that a medial/lateral split is quite crucial to understanding its function. However, I find the lack of a similar treatment of striatum to be somewhat surprising – prior work has observed a split between dorsolateral striatum (associated with MF behavior) and dorsomedial striatum (associated with MB behavior). Instead, the authors position competition here between “low-level striatum” and “high-level prefrontal” regions.

We are pleased that the reviewer values our analysis of the MD and appreciates the importance of considering its medial/lateral division, which aligns with established neuroanatomical findings. We acknowledge the reviewer's insightful point regarding the lack of a similar distinction in our treatment of the striatum, particularly given the well-documented differences between the dorsolateral and dorsomedial striatum and their respective associations with model-free (MF) and model-based (MB) behaviors.

Our univariate fMRI parametric analysis, designed to identify the neural correlates of MF versus MB RL, revealed that neural responses in the dorsal striatum were exclusively linked to reward prediction error (RPE), a hallmark of MF RL. In contrast, state prediction error (SPE), which represents MB RL, was detected in the PFC, but not in the striatum. These results are in line with prior study by Gläscher et al. (2010, Neuron), which used a probabilistic Markov decision task and found the neural signature of an SPE in the lateral prefrontal cortex, while RPE was detected in the striatum. We agree that prior findings have also shown the dorsomedial striatum to be involved in MB behavior. However, due to the lack of dissociation of the dorsal striatum in our results, we could not perform a similar analysis as we did for the MD. Additionally, we confined the search for RPE- and SPE-related regions to brain areas identified through the contrast of *Switch* vs. *Steady State* conditions (i.e., dmPFC, MD, caudate), which could limit the possibility of detecting involvement of a wider range of brain areas.

However, the functional and anatomical architecture of neural circuits involving the dorsomedial/dorsolateral striatum, in mediating different reinforcement learning strategies during flexible decision-making is indeed a critical area of research. We agree that further investigation in both human and animal studies is needed to provide a more comprehensive understanding of the neural mechanisms underlying these complex cognitive processes. We incorporated this important point in our discussion as a perspective for future research.

In general, the study is quite methodologically broad – and this is certainly a strength. Integrating behavioral analysis, computational modeling, multiple fMRI approaches, as well as an animal activation study is no small feat, and I applaud the authors for the enormous amount of work that has clearly gone into the set of excellent results here. This, however, brings me to what (somewhat surprising to me) feels like one of the weakest points of the paper, providing a good contextualization of these results, particularly vital given the multiple approaches they span.

We sincerely appreciate the reviewer's encouraging comments on our manuscript. We agree that providing a comprehensive contextualization of our results, particularly given the multiple approaches employed, is essential for enhancing the reader's understanding. In response to the reviewer's valuable feedback, we have substantially revised the manuscript. These revisions include conducting additional analyses, reorganizing the figures, results, and methods sections, and expanding the discussion section.

In particular, we focused on clarifying the methodological integration to better explain the rationale behind the choice of methods and their combined contributions to our conclusions. We believe these revisions significantly enhance the contextualization of our results and offer a clearer perspective on how our findings advance understanding in the field.

I find the discussion to be somewhat short, and while it presents a reasonable introduction to the ideas surrounding this literature, the authors spend very little unpacking the specific results of their analyses. As such, the paper currently reads as a number of reasonably strong and complementary analyses, but I think needs much more work in presentation to spell out for the reader how these fit into a single, consistent understanding.

We agree that a thorough interpretation of our findings is essential for providing a comprehensive understanding of the study. In the revised manuscript, we have significantly expanded the discussion section to include a more detailed interpretation of each analysis, incorporating additional references and comparisons to relevant studies.

Behavior and regional activation results

The authors analyzed behavior on the reversal learning task, and found that (as expected) subjects showed behavior of high accuracy prior to a switch, and a decrease followed by slow increase after a switch. This (and associated regional activations and link with behavior) all seems quite reasonable.

Thanks

RSA Analysis

2a) Perhaps I'm missing the point of this figure and the associated statistical test – the expected number of strategy keep vs. change trials is obviously reflective of how quickly subjects update their policies, but I'm not sure why one would want to test for a difference in significance between them, or what that significance would imply – rather, the ratio seems like it would be entirely adjustable by changing the post-switch window size – with a shorter window size, keep would dominate, but with a larger window size, change will dominate.

We agree with the reviewer that the statistical test in question is not appropriate in this context. Since we preformed some new analyses on the behavior, we removed Fig. 2a from the revised manuscript.

In response to this comment, as well as to the comment#1c from the Reviewer#1, we performed new behavioral analyses to characterize the adaptations more precisely using three parameters. Among these, the switch offset (s) measures the latency for participants to reach chance performance levels. As shown in Fig. R3, the switch offset was 4.4±1.6 (mean ± std), indicating that participants typically reached chance performance by the 4th trial after rule reversals. This new analysis provides additional statistics on the dynamic reconfiguration of behavior, offering a clearer view of how participants adapted to the task on average. Based on the range of trials from the reversal point to the switch offset (from 1 to 9 trials across participants), we defined the 10 trials immediately following the reversals as the *Switch* period, as this window effectively encompasses both the exploratory phase and the successful transition to the new rule content.

Accordingly, we have revised this section to provide a more precise characterization of the behavioral representations.

2b-d) I'm struggling to understand the results in dmPFC – given that the overall strategy is a combination of the two individual strategies, how is it that we find their combination, one of the individual components, but not also the other individual component? This is less a question about interpreting their results, and more perhaps understanding how this could reasonably fall out of the analysis.

We realized that the overall strategy in the RSA analysis is only driven by strategy changes. To avoid potential confusion in the interpretation of the results, we removed the overall strategy from the RSA analysis. In addition, based on the previous comment, we restructured the figure and reorganized the results section for the RSA analysis to improve clarity and coherence.

Separately, are the strengths of these relationships related between subjects? That is, do some subjects more generally have more robust rule representations, and if so, is this perhaps connected with behavior at all?

We appreciate the reviewer's suggestion. For each participant, we generated one RDM, which captures the relationship between two different types of trials (i.e., correct strategy (Switching representation): cue1 → 'NoGo' and cue2 → 'Go'). This RDM serves as an indicator of behavioral performance. We conducted a correlation analysis between the strength of the RDM for the Correct Strategy (Switching) in dmPFC and MD and the three behavioral parameters: switch offset (s), slope (α), and lapse (ϵ).

We identified a marginally significant correlation between the strength of the RDM for the Correct Strategy (*Switching*) in the dmPFC and the switch offset ($p = 0.021$, 0.05/3 for the multiple corrections, Fig. R13). This indicates that stronger similarity (or weaker dissimilarity) between the cue1 → 'NoGo' and cue2 → 'Go' trials, reflecting a stronger representation of the correct strategy, is associated with a shorter switch offset. In other words, participants with a

stronger representation of the correct strategy in the dmPFC transitioned more quickly after rule reversals. This finding suggests the critical roles for the dmPFC in facilitating the switching process. Importantly, these results align well with subsequent analyses, which revealed that the dmPFC is involved in the MB strategy and MB learning blocks are associated with a shorter switch offset compared to MF blocks (Fig. 4 in the revised manuscript).

We have incorporated these findings into the manuscript (Supplementary Fig. 4), significantly strengthening the narrative and providing further evidence that different strategies, mediated by distinct neural regions, underlie behavior in our task.

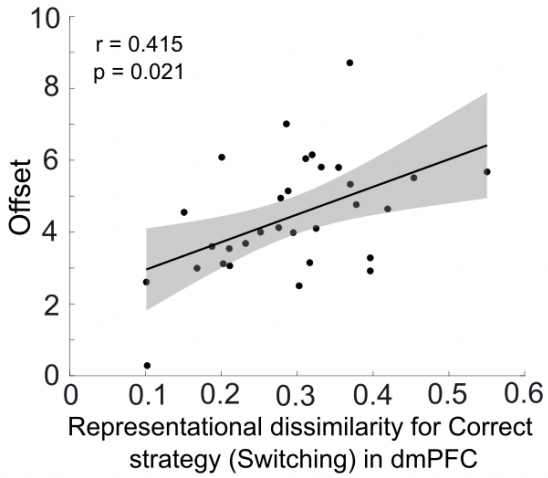


Fig. R13

Mouse Results

The authors report an inactivation study of MD in mice – It absolutely seems like a strong argument in favor of this overall connectivity which involves MD being crucial for reversal learning, and is quite a nice addition to the story. However, unclear if it argues so much that the MD is uniquely gating this behavior, rather than involved, which appears to be much of the crucial point of this paper. I wonder if the authors could expand on this.

The reviewer is correct that the mouse study alone cannot provide definitive evidence for MD's gating role. While the findings demonstrate that MD is necessary for reversal learning, they do not directly reveal the mechanism by which MD controls the switching between strategies. However, combined with neuroimaging and computational modeling results, a more comprehensive picture emerges. Neuroimaging studies show that MD activity correlates with the dynamic changes in task rules and cognitive demands. Computational modeling further supports this by demonstrating that MD can gate the neural representations of prefrontal cortex responses through prefrontal transthalamic pathways. Together, these findings suggest that MD is not merely involved in the task but actively gates the implementation of different strategies during reversal learning.

We have revised the manuscript to better highlight this distinction and to incorporate supporting evidence for the methodologies in the discussion. (**Lines 532-557**)

> "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, we observed that the medial MD primarily connects with the ventral PFC, while the lateral MD connects with the dorsal PFC.
>
> Emerging evidence suggests that the MD can mediate cortico-cortical communication by providing an indirect route, thereby enhancing the flexibility and efficiency of intercortical processing. 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[50]. These nuclei selectively gate cortico-cortical information transfer, positioning the thalamus as a critical player in the selection of cognitive processing streams.…
>
> Together, our findings extend the framework that the MD sustains and coordinates task-relevant cortico-subcortical representations, enhancing performance in complex cognitive tasks."

MF vs. MB Behavior

I don't think I entirely understand why we observe such oscillatory behavior in the model-free RPE. In particular, prior to a rule reversal, I would expect the exact timing and direction of RPEs to be relatively arbitrary, given the stochastic nature of the task. While not the part of the signal being used to identify MF RPEs, it seems like a quite striking signature of how they are deriving these model fits. Could the authors explain this? Otherwise, the findings in this section seem quite nicely tied in to the previous results.

As the reviewer correctly pointed out, the stochastic nature of the task introduces variability in the timing and direction of RPEs. However, the observed oscillatory behavior in the model-free RPE is not entirely random. This behavior arises from the interaction between the learning dynamics and the task structure.

In the model-free learning framework, participants update their expectations based on the difference between predicted and actual rewards (RPE). The oscillatory pattern reflects these trial-by-trial adjustments. Specifically, when a reward is received, the RPE is positive, prompting an upward update in the expected value. Conversely, if the reward is absent in the subsequent trial, the RPE becomes negative, leading to a downward adjustment in the expected value. This alternating pattern of positive and negative RPEs produces the observed oscillatory behavior.

The task structure further influences the shape of the RPE plot. The reversal learning task incorporates probabilistic rewards, with a 30% deviation sequence pseudorandomized across blocks and fixed across participants to ensure intersubject comparability in the learning process. The RPE plot represents the average across both blocks and participants and shows how participants continually adjust their expectations based on recent outcomes. This combination of individual learning dynamics and task design explains the observed behavior of model-free RPE.

I do have one additional question – from my understanding, the two neural correlates for MF and MB are based on assigning each block to one of the two based on performance. However, I would assume there to be a fairly substantial metalearning component – have the authors investigated whether behavior systematically changes/improves over time?

We appreciate the reviewer for raising this question. Based on comments from other reviewers, we conducted additional analyses on the behavioral data to more precisely characterize the dynamics of behavioral adaptation within each block. We fitted a logistic regression model with three parameters to the observed individual choices: switch offset s, slope α, and lapse ε (for more details, please refer to our reply to the comment #1c from the Reviewer #1).

In response to the reviewer's comment on behavioral changes over time, we conducted additional analyses. Specifically, we divided the blocks into four groups (first, second, third, and fourth block) based on their presentation order within each run. Since our experiment consisted of three runs, each comprising four blocks, we assumed that any time effects would primarily occur within each run rather than across runs. Therefore, we averaged the behavioral performance of blocks presented at the same time across the three runs (e.g., the first block of each run). We then fitted a logistic regression model with three parameters to the averaged behavioral performance for each block separately. A one-way ANOVA did not reveal a significant time effect on either the switch offset ($F=1.4764$, $p=0.2243$) or slope ($F=2.0929$, $p=0.1046$; Fig. R14). However, a significant group effect was found for the lapse rate ($F=4.7092$, $p=0.004$). Post-hoc analyses revealed that the lapse rate in the second block was significantly higher than in the other three blocks ($p<0.001$, multiple comparison correction). While this finding is somewhat unexpected, it does not provide sufficient evidence to conclude that behavior systematically changes or improves over time, especially given the non-significant results for switch offset and slope.

This finding aligns with our expectation that there is no time effect on behavioral performance, as each block required participants to relearn a stimulus-response association from scratch with a novel pair of tactile stimuli. Additionally, the rule reversal point was unpredictable from block to block, preventing participants from anticipating or adapting based on prior experiences. As a result, any experience gained in one block is unlikely to influence reversal behavior in the next, as the three regression parameters should primarily be driven by experiences and outcomes within the current block.

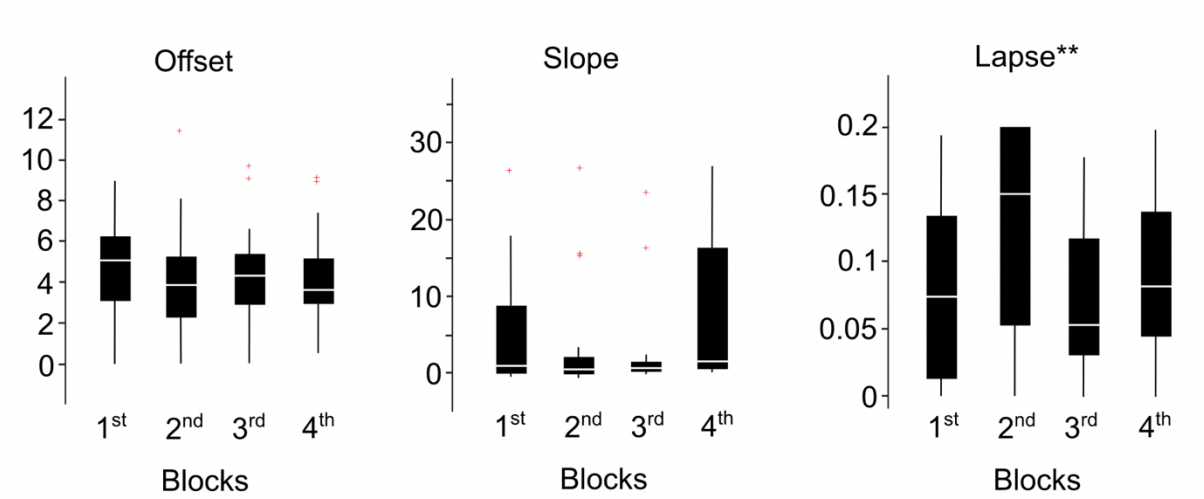


Fig. R14

Computational Model

I have to admit I am somewhat lost with the computational model underlying a large portion of this section, which seems somewhat split between this figure, the supplement, and perhaps further detail in a cited paper which however is not yet published or otherwise available. That being said, the result with multiple concurrent representations during the MB based blocks is somewhat compelling.

Thank you for recognizing the importance of multiple concurrent representations during MB-based blocks. We realized that our previous explanation of the modeling results lacked sufficient detail. To address this, we have carefully revised the text throughout and Figure 6 to ensure that all analyses are clearly explained. We have also reorganized Figure 6 to highlight the most relevant results and revised the text to make it self-contained, ensuring that it is independent of the accompanying theory paper on CogLink.

In particular, we have expanded our explanation of how the model results support multiple concurrent representations. Below, we provide a revised excerpt illustrating this mechanism: “Our CogLink network enabled us to investigate the mechanisms underlying MB and MF switching. Analysis of MDl 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. 6f). In the CogLink network, distinct contextual populations of MDl neurons selectively modulate disjoint populations in the orbitofrontal cortex (OFC), regulating both their activity and plasticity (see Methods; Supplementary Fig. 7a, b) 26. 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. 7a, b). As a result, switching behavior relied on overwriting the existing mapping in the same OFC population. Specifically, in Fig. 6g, we observed a “crossing" 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. "

These revisions ensure that all necessary details of the computational model are included in the manuscript, allowing the reader to fully understand the results without relying on the supplementary material or the theory paper. We hope this clarification fully addresses the reviewer's concerns.

I am also curious if the authors have any intuition for why results were somewhat assymetric – caudate for MF and dlPFC for MB – whereas their modeling assumes these both to be driven by distinct neocortical regions (OFC and dlPFC) and presumably both with their own BG connections.

Thank you for raising this question. We agree that BG and OFC-MDm circuits are all components of the MF reinforcement learning (RL) network, but they likely serve distinct and complementary roles. Specifically, we propose that the BG supports rapid value-based decision-making, whereas the OFC facilitates a more deliberate, value-based evaluation process. This distinction aligns with findings from Balewski et al. (Neuron, 2022), which demonstrated that caudate nucleus (CdN, BG) activity predicts choice direction rapidly after trial onset, independent of response times, and correlates with the first saccade direction. This suggests that BG contributes to rapid decision-making. In contrast, the OFC exhibits flip-flopping dynamics in value representations that are independent of eye movements but predictive of choice time, indicating a slower, deliberative valuation process. These findings support the idea that the BG and OFC reflect parallel but distinct valuation processes that contribute to value-based decision-making in complementary ways. We have less intuition regarding the specific contributions of dmPFC and dlPFC-MDl circuits to MB decision-making, and further work is needed to clarify their distinct roles.

References:

Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. (2011) Model-based influences on humans' choices and striatal prediction errors. Neuron. 69(6):1204-15. doi: 10.1016/j.neuron.2011.02.027.

Gläscher J, Daw N, Dayan P, O'Doherty JP. (2010) States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. Neuron. 66(4):585-95. doi: 10.1016/j.neuron.2010.04.016.

Balewski ZZ, Knudsen EB, Wallis JD. (2022) Fast and slow contributions to decision-making in corticostriatal circuits. Neuron. 110(13):2170-2182.e4. doi: 10.1016/j.neuron.2022.04.005.

Responses to reviewers' comments

We would like to thank the editor and reviewers for their thoughtful evaluation of our revised manuscript and for providing further valuable comments. We have carefully considered all suggestions and made additional revisions accordingly. We hope that these changes address the remaining concerns and meet the reviewers' expectations.

Below, we provide a point-by-point response to the reviewers' comments. Reviewer comments are presented in black, followed by our responses in blue. All changes made to the manuscript have been highlighted.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

I have reviewed the revised material, and I am pleased to confirm that they are satisfactory. I extend my gratitude to the authors for their meticulous consideration of all my concerns. The revised work has significantly strengthened from its previous version.

I have only one minor suggestion related to one of the new paragraphs in the discussion: Gueguen et al. published a study (Gueguen et al. Nat. Comms 2021) in which it was shown that the ventromedial and dorsolateral prefrontal cortex encode different types of prediction errors (reward and punishment prediction errors). Could the authors modify the sentence 'reward prediction errors are associated with the ventral striatum' (L 501-502) and incorporate the results of Gueguen et al.?

We are pleased that the reviewer found our revisions satisfactory and are grateful for the insightful feedback, which has significantly contributed to improving our manuscript.

In response to the reviewer's suggestion regarding the integration of findings from Gueguen et al. (Nat. Commun. 2021), we fully agree that their study offers valuable insights. Specifically, their work compellingly demonstrates that the sign (positive vs. negative) and domain (reward vs. punishment) of PE signals are differentially encoded in the vmPFC and dlPFC.

We have incorporated this point into our discussion as follows:

> "While RPEs are commonly associated with the ventral striatum, recent findings by Gueguen et al. suggest that the vmPFC and lOFC contribute to encoding reward PEs, whereas the insula and dlPFC are more involved in punishment PEs, indicating a distinct cortical-subcortical organization for PE processing."

Reviewer #2 (Remarks to the Author):

The authors have undertaken a substantial revision in response to my concerns and those of the other reviews. The resulting manuscript is substantially improved— the new analyses and explanations clarify the findings and impacts. I am particularly impressed with the use of the biophysical model to explain the DCM findings and to make new RSA and behavioral predictions. Overall, I am enthusiastic about the revision, and in general the synthesis of methods and approaches to interrogate the role of the MD nucleus in learning.

We sincerely appreciate the reviewer's thoughtful comments and are pleased to read that the revised manuscript is considered substantially improved.

However, I do have some additional concerns that were raised in my reading of the revision and response letter:

1) Most significantly, I am concerned about statistical circularity in performing RSA analyses on functionally-defined ROIs. The authors should ideally use anatomical or independent ROIs, or very carefully justify why the RSA analyses are statistically independent.

We thank the reviewer for raising this important point. In our study, we defined independent regions of interest (ROIs) using a univariate contrast of *Switching* > *Steady State* ($p < 0.001$, uncorrected). The subsequent RSA analyses were then confined to the Switching period. Specifically, we extracted unsmoothed response patterns from Switching trials to construct representational dissimilarity matrices (RDMs). These trials were categorized into four distinct types: cue1–'Go', cue2–'NoGo', cue1–'NoGo', and cue2–'Go'. As these trial types and contrasts were not used for ROI definition, we believe there is no issue of circularity in our RSA analyses.

This approach aligns with established practices in the field. For example, Haxby et al. (2001, *Science*) used a similarity matrix approach to examine category-specific representations (faces vs. objects) in ventral temporal cortex, with ROIs defined based on univariate contrasts of category differences. Similarly, Barron et al. (2020, *Cell*) defined hippocampal ROIs using a univariate contrast of correct > incorrect trials ($p < 0.01$, uncorrected) before applying RSA to test inferred outcome representations during the correct trials. These examples illustrate that defining ROIs from orthogonal univariate contrasts is a standard and accepted approach in RSA studies.

Nonetheless, to further address any concerns regarding statistical independence, we performed complementary analyses using anatomically-defined ROIs. Specifically, we defined the MD and caudate regions using the AAL atlas. As the dmPFC is not delineated in the AAL, we constructed this ROI using the Brainnetome Atlas (Fan et al., 2016, *Cerebral Cortex*), informed by Carlén's work (2017, *Science*). These anatomically-defined ROIs are illustrated in Fig. R1. We extracted response patterns from these regions and constructed RDMs, applying the same RSA procedures as described in the main manuscript to examine decision strategy representations during the *Switching* period.

The results from the anatomically-defined ROIs corroborated our original findings: we observed significant representations of the Correct Strategy in both the dmPFC and MD (one-sided permutation tests: dmPFC, effect size = 0.84, $p = 0.001$; MD, effect size = 0.77, $p = 0.002$; Fig. R2). No significant effects were found in the caudate after the correction of multiple comparison (effect size = 0.54, $p = 0.029$; Bonferroni-corrected threshold: $p < 0.025$). For the Wrong Strategy, none of the regions showed significant representation. These results are consistent with those obtained using functionally-defined ROIs.

We believe these additional analyses provide further validation of our RSA findings and underscore the robustness of our methodology. The results from the anatomically-defined ROI analyses have been added to the Supplementary Materials.

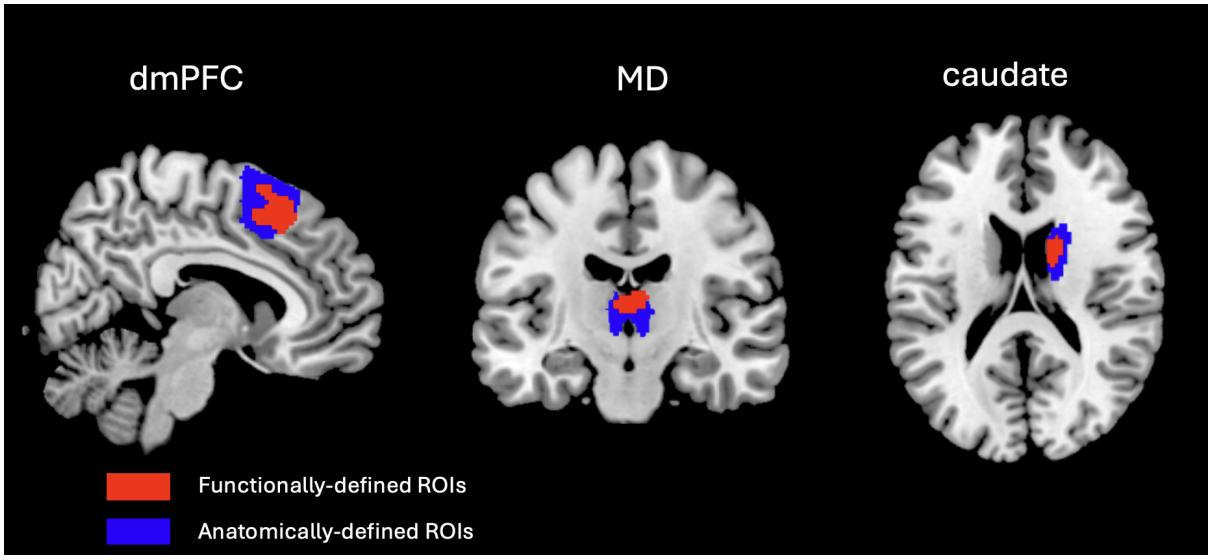


Fig. R1 Functionally defined ROIs (dmPFC, MD, and caudate based on the *Switching > Steady State* contrast) and anatomically defined ROIs derived from the AAL and Brainnetome atlases.

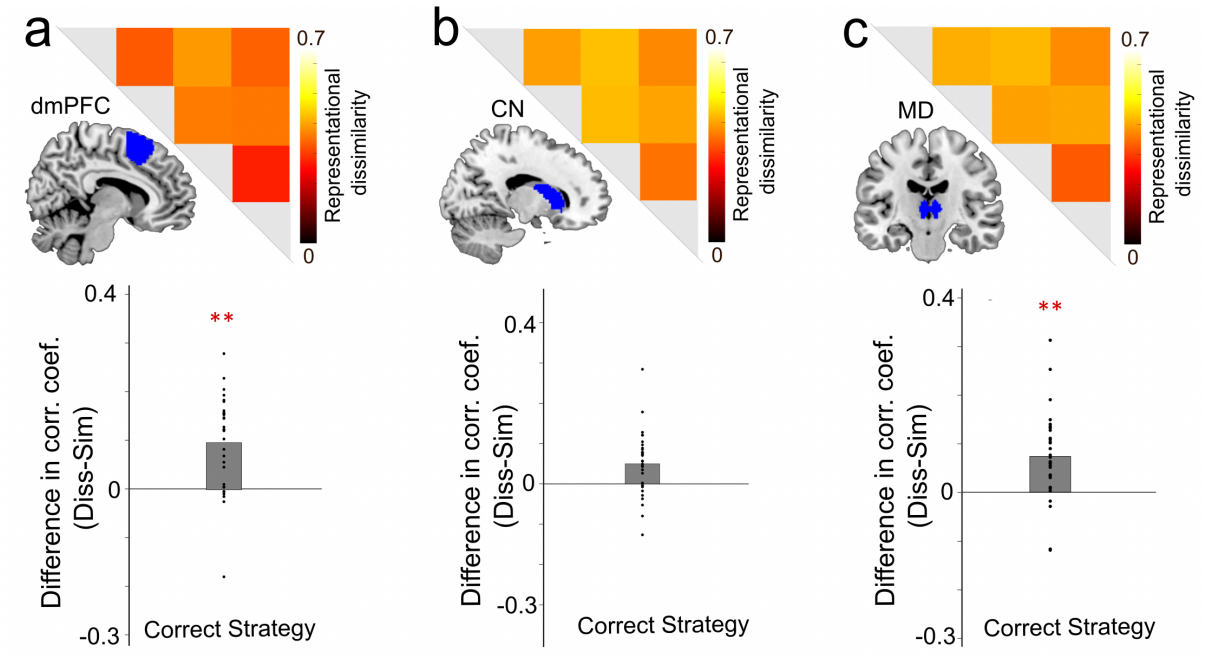


Fig. R2 The RSA results based on the anatomically defined ROIs

2) I was very confused by the description of the RSA analyses in the rebuttal letter. Figure R9 appears to be asserting a finding from a null—this should be more carefully interpreted. Figure R10 is inscrutable to me—the labels are not very informative.

We agree that Fig. R9 should be interpreted with caution, as it compares the similarity of real data to random data generated by shuffling voxel sequences. The absence of a significant difference may suggest that the response patterns in MF and MB MD thalamus are unrelated. However, we acknowledge that the conclusions from this analysis alone may be limited in strength. To address this, we conducted an additional analysis, presented in Fig. R10.

Our approach in Fig. R10 was inspired by a method used in Haxby et al. (2001, *Science*), where response patterns for each stimulus category (e.g., faces, places) were separately measured from even- and odd-numbered runs within individual subjects. By assessing the similarity between these independent response sets, the authors tested whether category information was consistently encoded.

Adopting a similar strategy, we split both the MF and MB blocks into two independent sets and extracted MD response patterns from these four conditions (MF_set1, MF_set2, MB_set1, MB_set2). We then computed the similarity between the response patterns within and across block types.

Our hypothesis was as follows: if MF-related and MB-related MD responses are truly distinct, we would expect high similarity (low dissimilarity) within the same condition (MF or MB), and low similarity (high dissimilarity) across conditions. In contrast, if the response patterns are not distinct, similarity should be high both within and across block types.

As shown in Fig. R10 (now relabeled as Fig. R3 in the response letter), the results confirmed that within-condition similarity (MF_set1 vs MF_set2; MB_set1 vs MB_set2) was significantly higher than between-condition similarity (MF_set1 vs MB_set1; MF_set2 vs MB_set2, etc.). These findings provide convincing evidence that the MD exhibits distinct and independent response patterns for MF- and MB-related processes.

To improve clarity, we have also updated the labels in the corresponding Supplementary Figure.

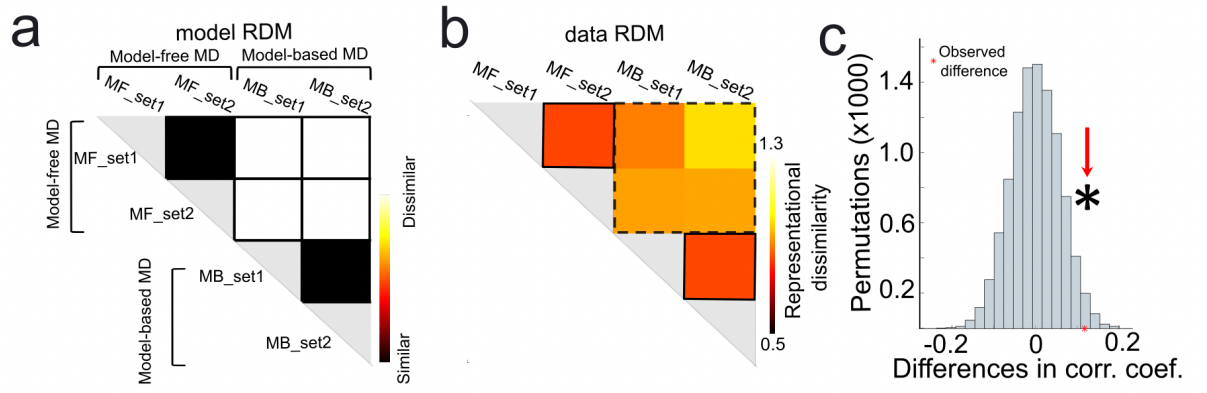


Fig. R3

3) The authors should include some caveats about interpreting correlations with their sample size (n = 32)

We thank the reviewer for the insightful comment regarding the interpretation of correlation results given our sample size.

In response, we have added the following caveat to the Results section:

“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— characterized

by increased exploratory behavior prior to reaching a stable learned state—may be associated with stronger modulation of dmPFC-to-MD connectivity."

4) Were the average modulatory effects in the DCM positive or negative?

In response to the reviewer's comment, we conducted a Bayesian parameter averaging (BPA) analysis on the winning model. This approach estimates a joint posterior distribution at the group level by combining individual participants' posterior densities. A posterior probability threshold of 90% was used to determine significant effective connectivity.

The BPA revealed significantly *positive* modulatory effects (B-matrix) for all four connections in the winning model: dmPFC → MD (0.353), striatum → MD (0.423), MD → dmPFC (0.691), and MD → striatum (0.811) (Fig. R4, left panel). These *positive* values indicate that rule reversals enhance the effective connectivity from one brain region to another. All connections exceeded the significance threshold (posterior probability > 0.9; Fig. R4, right panel).

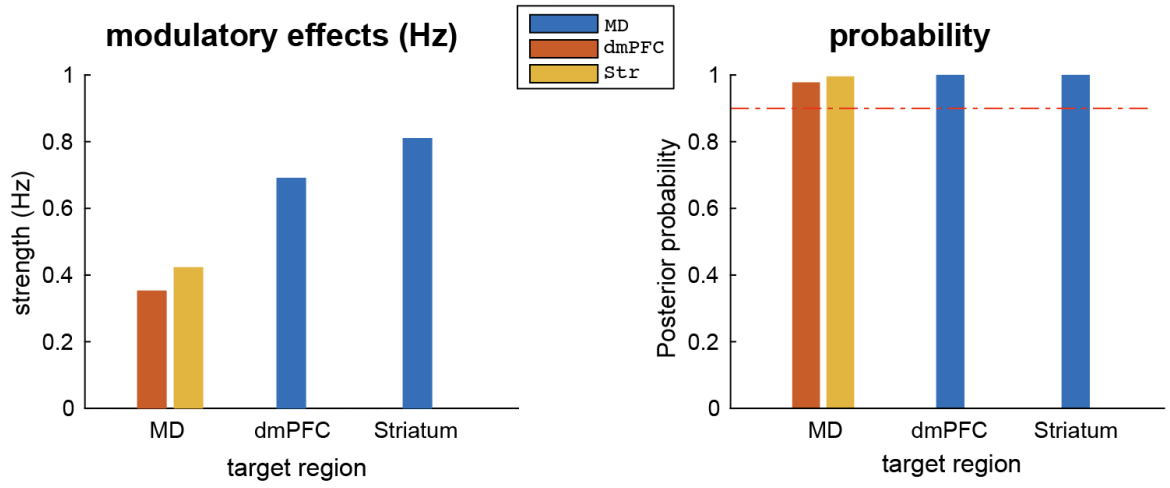


Fig. R4 Bayesian parameter averaging (BPA) for the four connections in the wining model

5) Are the SPE and RPE results collapsing across the MB and MF blocks? If so, are these result different between those blocks?

We derived reward prediction errors (RPE) from the *model-free* blocks and state prediction errors (SPE) from the *model-based* blocks. These were subsequently used as parametric regressors in GLM analyses of the fMRI data to identify associated neural signals.

In our initial response letter, we addressed a similar comment from Reviewer 1 regarding the comparison of RPE and SPE across block types. Given the overlap in concerns, we reiterate our response here. Specifically, we examined RPEs within *model-based* blocks and SPEs within *model-free* blocks to assess how these signals vary across contexts. Our analysis revealed no significant difference in RPEs, suggesting that the computation of RPE is stable across both learning strategies - whether based on direct reinforcement (model-free) or informed by task structure (model-based).

In contrast, SPEs were significantly greater during *model-based* blocks compared to *model-free* blocks (linear mixed-effects test, $t(62) = 3.10$, $p = 0.003$; see Fig. R5). This finding

indicates that discrepancies between predicted and actual environmental states are more pronounced in the context of model-based learning, particularly following rule reversals. This likely reflects the greater cognitive demands of model-based learning, which requires ongoing maintenance and updating of internal models to support flexible, goal-directed behavior.

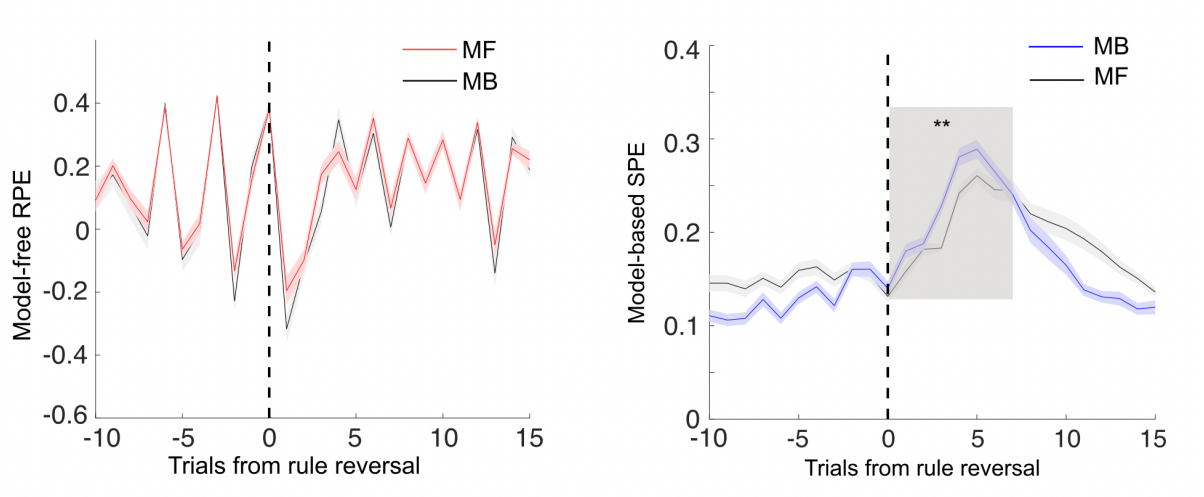


Fig. R5 The comparison of reward prediction error (RPE, left) and state prediction error (SPE, right) between the model-free and model-based blocks.

6) In the biophysical modeling, how are the MF and MB blocks defined? Is this model trained to explain the human behavior, or only to perform the task?

We thank the reviewer for this question. The definition of model-free (MF) and model-based (MB) blocks is provided in the Methods section. For clarity, we summarize it below:

> "After observing in Fig. 6c that the MDl 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 xmdl 2 were categorized as model-free (MF) blocks (n = 192)."

We chose to use MDl activity rather than the log posterior odds ratio of behavior as the classification criterion because behavioral data can exhibit high variability, even when driven by the same underlying strategy. In contrast, neural representation (as shown in Fig. 6b) can provide a more stable and accurate reflection of the underlying state.

Regarding the model training process, the CogLink model is trained solely to perform the probabilistic reversal task and is not directly fitted to human data. Despite this, the model successfully replicates both the behavioral signatures and the neural patterns observed in human subjects and produces testable behavioral and neural predictions, demonstrating its ability to capture essential aspects of the human decision-making process.

We hope this clarifies our approach and the rationale behind our classification criteria.

7) The results from the mouse data should include some more justification that mice could learn the task and that the split into switch and steady-state blocks is justified.

We have added further justification regarding mice performance in the revised manuscript. Specifically, only sessions in which animals achieved >65% correct responses during the Steady-State period were included in the analysis. We then assessed performance before and after rule reversals, finding that accuracy remained above 80% in both cases. This indicates that the mice successfully learned the task during initial acquisition and were also able to adapt following reversals.

Additionally, we defined switch onset as the first trial in which the cue-side mapping was reversed, and considered the switch complete when mice made three consecutive correct responses under the new rule.

This information has been incorporated into the manuscript as follows:

"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. 3). 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)…"

8) The statistical reporting has been much improved, but there are still cases where only p-values are reported. Please adopt a consistent style, with all information required according to the reporting summary. Additionally, the way the multiple comparisons correction is presented is somewhat unorthodox and confusing.

We have thoroughly reviewed the statistical reporting throughout our manuscript and have now ensured that all relevant values are provided in a consistent and transparent format. The newly added statistical details have been highlighted in the revised manuscript.

Additionally, we have revised the section on multiple comparisons correction to improve clarity. For permutation tests and correlation analyses conducted in the context of Representational Similarity Analysis (RSA), we applied Bonferroni corrections by dividing the significance threshold (0.05) by the number of comparisons (0.05/*n*). Specifically, we performed two comparisons (0.05/2) for the permutation tests (Correct vs. Wrong strategy models) and three comparisons (0.05/3) for the correlation analyses (switch offset, slope, lapse), separately for each brain region.

For the other fMRI analyses, we used whole-brain family-wise error (FWE) correction in the general linear model (GLM) analyses. In cases where we had a-priori hypotheses about specific brain regions, we applied small volume correction (SVC) to restrict the search space accordingly (e.g., in prediction error–related and PPI analyses).

9) The sentence "MDl neurons exhibited strong contextual encoding (Fig. 6b), with their activity closely linked to MB behaviors, mirroring patterns observed in human subject" needs more explanation (this seems to be explained more fully in the Figure legend than in the text).

We have expanded the explanation of this sentence in the manuscript as follows:

"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 ration of model-based versus model-free strategy (r = 0.56, p = 2.00×10−5, **Fig. 6c**). This relationship aligns with human data, where the activity of MDl is significantly correlated with state prediction error in a model-based RL algorithm (**Fig. 4f**)."

10) As mentioned, I found the RSA analyses arising from the biophysical modeling valuable. However, I think the results are too strongly interpreted as written. These RSA results are not evidence of toggling, though toggling is one possible explanation for the pattern the authors observe.

We agree with the reviewer's point and have accordingly tempered our conclusions. In the revised manuscript, we now present the 'toggling' predictions of the CogLink model as one possible interpretation of the RSA results, as follows:

"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."

"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."

Reviewer #3 (Remarks to the Author):

I thank the authors for their lengthy revisions and accompanying analyses. I find that the paper is significantly better connected than before, though perhaps still slightly disjoint. My only significant issue at the moment is with a new analysis the authors have introduced to show that the behavior is indeed a mixture of model-based and model-free behavior. While I think that the idea of such an analysis would significantly strengthen the paper, I do not believe the current attempt at this is methodologically sound and should be in some form replaced.

Details below:

Behavior

The authors have significantly improved their characterization of switching behavior. The lapse rate, slope, and delay seem like a reasonable choice, and a much better parametrization to fit.

However, I am not at all sold on the new analysis that the authors have included in figure 1g and 1h. This analysis appears to have been added as a response to multiple reviewer comments that the model-free/model-based behavioral analysis was somewhat lacking. I believe that this new analysis introduces a number of new issues.

The authors claim that, as their task involves common/rare rewards, this maps onto the common/rare transition structure of a two-step task. Note that the two-step task already involves varying reward rates, but the common/rare analysis is explicitly at the level of transitions, not at the level of such rewards. In the two-step task, rare outcome for a model-based agent should lead to an explicit reversal of attribution, e.g. a rare rewarded trial should increase confidence in the unchosen option, and vice-versa. That is, knowledge of the structure of the task leads to opposite interpretations of the same outcomes.

However, that's clearly not the case here, principally because a model of the task doesn't allow me to differentiate common/rare unless I already quite confidently know which regime I'm in, which defeats the entire point. That is, assume option 1 to be the 70% rewarded option, I choose option 1, and it's unrewarded. The two-step logic is that, due to this being a rare transition, I should assign that lack of reward to option 2, and in fact increase my confidence in option 1. Quite frankly, that logic makes no sense on the current task. At best, a strong prior belief that option 1 is the better option means that I may be less sensitive to an unrewarded trial on option 1, but there's no plausible reason why I should take that lack of reward as evidence against option 2.

So while the authors certainly can perform this analysis on the current dataset, mapping onto common/rare reward outcomes, it's somewhat incomprehensible from my perspective. My own interpretation of figure 1h is that the reviewers convincingly show a very strong MF effect, but what they're interpreting as MB is not due to a model so much as recent trial history, and a tendency of rare or common unrewarded outcomes to be exactly that. (Lagged learning effects would likewise provide a problem in the two-step task, hence typical setups of reward probabilities that drift around 0.5).

More generally, I think the idea of a one-back model might be misguided here, and there may not be a great way to build intuition that doesn't involve an actual history of recent trials. That being said, I believe most of the analyses linked to behavior later rely on a comparison of something akin to an ideal observer changepoint model versus a model-free agent, so I hope that this should not cause too many problems?

We appreciate the reviewer's insightful clarification regarding the distinction between common/rare transitions in the two-step task and the common/rare outcomes in our paradigm.

As noted in our previous revision, our probabilistic reversal learning task does not incorporate transition probabilities. Therefore, dissociating model-free and model-based strategies based on behavioral patterns - such as in the two-step Markov decision task - is not appropriate in our context. We fully agree with the reviewer's concern that applying this type of behavioral analysis to our paradigm is problematic. Specifically, the logic of interpreting rare outcomes as reflecting evidence for or against unchosen options, central to the two-step task, does not translate to our design.

In light of these conceptual concerns, we have removed the analysis of recent trial history and the common/rare reward outcome analysis (previously Figures 1g and 1h). Instead, our revised manuscript now focuses on neural representational differences between blocks that align more closely with either model-free or model-based strategies, and we provide a post hoc justification for this block-wise categorization.

Specifically, rather than classifying individual trials, we assigned blocks as model-free or model-based by comparing each block's empirical choice patterns with those generated by simulated model-free and model-based agents. This allowed us to categorize blocks based on which model provided a better fit for each participant's behavior. To validate this categorization, we analyzed parameters from a logistic regression model. Notably, we found that switch offset was significantly longer for model-free blocks than for model-based blocks (Fig. R5), with model-free blocks requiring more than five trials (mean ± SD = 5.3 ± 2.2) to return to chance level after rule reversal, compared to fewer than four trials (mean ± SD = 3.7 ± 1.5) for model-based blocks. This suggests that model-free strategies involve a longer exploratory phase following environmental changes. However, no significant differences were observed in slope or lapse parameters, indicating that both strategies adapt comparably once the new rule is learned.

These behavioral distinctions are consistent with theoretical differences between model-free and model-based strategies and are unlikely to reflect sampling bias, behavioral noise, or individual differences in learning performance. Thus, we believe the subsequent analyses comparing model-free and model-based behavior remain valid.

That said, as we noted in our previous revision, we recognize that the categorized blocks do not represent *pure* model-free or model-based strategies. Rather, they reflect behavior that is more consistent with the respective computational models. To ensure transparency, we have explicitly included this clarification in the Discussion section:

> "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 the model-free or model-based models."

We hope that this clarification and the revised focus of our manuscript address the reviewer's concerns while providing a more accurate representation of the underlying learning strategies in our experimental paradigm.

RSA Analysis:
Figure 3 is much clearer, and the authors have removed the somewhat confusing test of overall strategy.

Thank you.

I find fig R13 quite interesting, though somewhat confusing. Increased RDM-related activity patterns are associated with a larger offset, e.g. slower switching behavior? Isn't this somewhat unintuitive? Also unclear on the multiple comparisons number. What does 0.05/3 refer to? However, I do appreciate that the authors have carried out these tests.

Please note that in this analysis, we examined correlations between the representational **dissimilarity** strength of the *correct strategy model* (x-axis in Fig. R6) and three behavioral parameters: switch offset (s), slope (α), and lapse (ϵ). Greater dissimilarity - reflecting weaker similarity between the neural response patterns of different trial types - was associated with a larger switch offset, indicating slower behavioral adaptation following rule reversals. Since increased dissimilarity corresponds to a weaker neural representation of the correct strategy, these findings suggest that participants with a less robust representation in the dmPFC exhibited slower switching behavior. This result aligns well with our theoretical expectations.

We also apologize for not clearly stating our approach to correcting for multiple comparisons. We used Bonferroni correction by adjusting the significance threshold based on the number of comparisons (0.05 / n). As we conducted three correlation analyses—linking dissimilarity strength with switch offset (s), slope (α), and lapse (ϵ)—we applied a corrected significance threshold of $p < 0.05 / 3$.

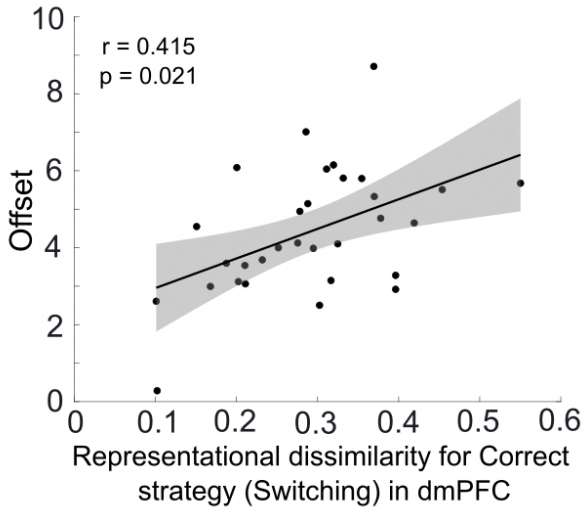


Fig. R6

Mouse Results

The authors have elaborated on the role of the MD here, linking the mouse and human studies, quite a bit better than before.

Thank you.

MF vs. MB Behavior

I'm still not sure that I follow the oscillatory behavior prior to switches. The oscillation due to over/under prediction makes sense, but I find it surprising that with a pseudo-randomized sequence there would still be so much structure across blocks. I'm assuming that the pseudo-randomized sequence was relatively random between blocks – or did it simply end up with highly similar reward sequences prior to the switch even across different blocks?

We employed a pseudo-randomized sequence in our experimental design to ensure that trial order was relatively randomized across blocks. However, we acknowledge that, due to chance, some blocks included highly similar reward sequences prior to the switch. This was not by design but an unintended consequence of the pseudo-randomization process. To illustrate this point, we plotted the proportion of trials with a 70% reward outcome across blocks (Fig. R7). As shown in the figure, the 4th and 1st trials before the rule reversal (denoted as -3 and 0 on the x-axis) were of the same type across all 12 blocks. This unintended repetition in trial structure contributed to the observed oscillatory behavior and prominent peaks preceding the switches.

Importantly, the RPE and SPE used to identify neural correlates were derived exclusively from the switching period. These analyses did not include data from trials occurring prior to the rule reversals.

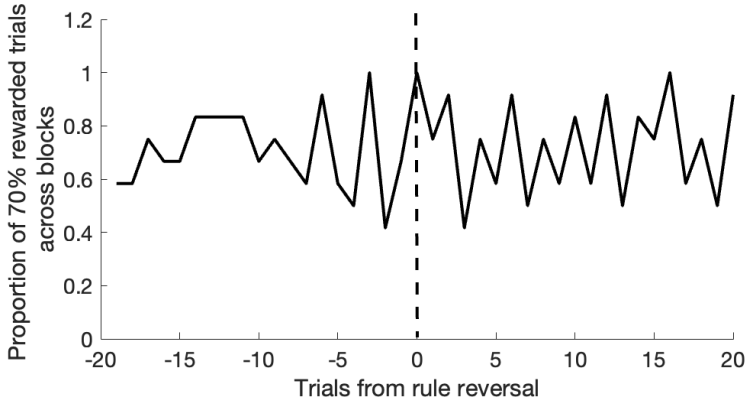


Fig. R7

I thank the authors for carrying out the metalearning analysis across blocks. It is interesting and surprising to me that the actual detection of reversal didn't improve with practice, though I suppose quite plausible.

Thank you. We agree with the reviewer's perspective. We believe the observed outcome is primarily attributable to our experimental design, which incorporated probabilistic contingencies, a novel pair of tactile stimuli for each block, and an unpredictable rule reversal point. These features were intentionally implemented to minimize carryover effects, ensuring that experience gained in one block would not influence reversal behavior in the next.

MF vs. MB Behavior

The presentation here is much improved.

Thank you.

References:

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science. 2001 Sep 28;293(5539):2425-30. doi: 10.1126/science.1063736.

Barron HC, Reeve HM, Koolschijn RS, Perestenko PV, Shpektor A, Nili H, Rothaermel R, Campo-Urriza N, O'Reilly JX, Bannerman DM, Behrens TEJ, Dupret D. Neuronal Computation Underlying Inferential Reasoning in Humans and Mice. Cell. 2020 Oct 1;183(1):228-243.e21. doi: 10.1016/j.cell.2020.08.035.

Fan L, Li H, Zhuo J, Zhang Y, Wang J, Chen L, Yang Z, Chu C, Xie S, Laird AR, Fox PT, Eickhoff SB, Yu C, Jiang T. The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. Cereb Cortex. 2016 Aug;26(8):3508-26. doi: 10.1093/cercor/bhw157.

Carlén M. What constitutes the prefrontal cortex? Science. 2017 Oct 27;358(6362):478-482. doi: 10.1126/science.aan8868.

REVIEWERS' COMMENTS

Reviewer #3 (Remarks to the Author):

I have reviewed the updated manuscript, and I am generally satisfied with the changes the authors have made.

However, I am not sure if the authors have fully revised the methods section in-line with their response? I still see discussion of the common/rare transition logic, which probably should be removed.

If I had to summarize, it seems that the authors have essentially switched to fitting whether individual blocks are best represented by a changepoint detection algorithm versus a more naive model-free approach. It certainly tests whether learners are picking up on what would be considered a model of task structure, but I'm not sure this falls within the classic reinforcement learning sense of model-based.

It may be worth clarifying how exactly this kind of model fits into the reinforcement learning literature somewhat in the text.

We appreciate the reviewer's positive assessment of our revisions. We apologize for the incomplete removal of the analysis of the interaction between reward and transition probability in the Methods section. This material has now been removed in full, and we have carefully reviewed the entire manuscript to ensure that no similar inconsistencies remain.

We agree with the reviewer that we should clarify how our "model-based" model fits within classical model-based reinforcement learning (MBRL). Classical MBRL is defined by the use of a learned internal model of the task, an agent's internal representation of environmental dynamics, to simulate experience and perform planning. In contrast, our "model-based" model extends the "model-free" model by incorporating the knowledge that the task alternates between two (loosely defined) sets of stimulus–response associations (rule reversals): one in which cue 1 favors Go (and cue 2 favors NoGo) and one in which cue 1 favors NoGo (and cue 2 favors Go). Accordingly, our model includes a confidence module that estimates the probability of a rule reversal from recent errors. This definition is similar to that detailed in a previous study (Sarafyazd et al., *Science*, 2019). We agree that our approach to defining the "model-based" model shares conceptual similarities with change-point detection, but our categorization specifically operationalizes sensitivity to task structure - that is, estimating the probability of rule reversals based on errors and adjusting choices accordingly. While this may not fully correspond to canonical model-based planning over an internal transition–reward model (Daw et al., Neuron, 2011), it captures the core distinction between structure-sensitive and reinforcement-driven strategies in this task. Thus, our framework still aligns with reinforcement learning in a broader sense: the "model-free" blocks capture habitual, reinforcement-driven choice patterns that rely solely on experienced outcomes, whereas the "model-based" blocks capture behavior that is sensitive to the task's latent structure and rule transitions. The key functional difference we emphasize is the speed of adaptation after a reversal: model-free strategies exhibit a longer exploratory phase, whereas model-based strategies adjust rapidly once the new rule is detected.

We clarified this in the Method section of our manuscript as follows:

> "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 which estimates the probability of rule reversal based on errors in recent trials. The method is similar to that detailed in the previous study58. 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). 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.”

References:

Sarafyazd, M. & Jazayeri, M. Hierarchical reasoning by neural circuits in the frontal cortex. Science 364, 6441 (2019).

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, 1204–15 (2011).