nature portfolio

Peer Review File

Mediodorsal thalamus regulates task uncertainty to enable cognitive flexibility

Corresponding Author: Professor Zhe Chen

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

A version of this paper was originally rejected for publication by Nature Communications, however that decision was reconsidered after appeal by the authors.

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In this interesting paper, the authors model the mediodorsal (MD) thalamus and prefrontal cortex (PFC) using biologically grounded cell types and projections, and test the model on context-dependent decision-making tasks. The findings indicate that the MD-PFC model outperforms a PFC-only model, with potential implications for computational deficits in schizophrenia.
Overall, the paper provides meaningful insights, but the complexity of the model, with many free parameters, makes it difficult to interpret how the various components contribute to emergent dynamics and computations. Additional clarification would greatly benefit the reader.
• The manuscript does not sufficiently engage with existing literature on thalamocortical circuit models. Key concepts such as the alignment between cortical PCs and thalamus, neuron counts, and the role of thalamus in preparatory activity should be better contextualized using recent studies (e.g., Lakshminarasimhan et al. 2024, Logiaco et al. 2021, Kao et al. 2021).
• The decision to provide sensory inputs to both PFC and MD is unclear given current anatomical evidence. The authors should clarify whether MD receives direct sensory inputs or exclusively corticothalamic inputs, as this would improve the model's biological accuracy.
• Line 114: There is a lack of clarity around how the number of units in the PFC-MD and PFC-only models were matched. It is important to specify whether the performance improvement with MD is due to the network architecture or simply the higher number of units.
• The task description could be simplified to enhance readability. Consistent use of terms like "context" and "subtask" would reduce confusion and make the task setup easier to follow.
• Clarification is needed regarding whether RDM and CAC trials were interleaved or blocked, and whether the presentation of auditory and visual stimuli was mutually exclusive.
• Figures 3c-f require clearer labeling, particularly the x-axis, and the discrepancy between the text and the panels in Figure 4 should be resolved, as it may affect the interpretation of the results.
• In the schizophrenia-related simulations, it is unclear whether models with E/I imbalance were retrained before testing. This is crucial to understanding whether balance is needed for learning or simply for maintaining performance.
• Line 384: More explanation is required about the relationship between MD population size and external input dimensionality, particularly if the input size remained constant during simulations.
Methods:
• The description of the cost function, noise scale, and weight initialization is insufficient. More details in these areas would provide a clearer understanding of the modeling approach and allow the results to be more easily verified.
• It is unclear whether the input weights to PFC were fixed, which could explain the absence of context-coding in some results.
• Instead of using a single representative model for performance plots, showing the ensemble average could offer a more comprehensive picture of model behavior?
Minor Comments:
• The term "computational lesion" is ambiguous and should be clarified.
• Minor typographical errors, such as the typo in line 637 and the incorrect matrix element in line 653, should be corrected.
• Figures should be embedded within the text for easier reference, as the current layout makes the paper difficult to follow.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
This study investigates the mediodorsal (MD) thalamus's role in cognitive control, particularly in decision-making under uncertainty and cognitive flexibility. Using biologically-constrained computational models, the authors demonstrate that incorporating a feedforward MD structure into the prefrontal cortex (PFC) enhances robustness to low signal-to-noise ratios, improves working memory, and facilitates rapid context switching. The model successfully replicates key neurophysiological findings from mice, offering new insights into the computational mechanisms of the MD in regulating cue uncertainty and context switching. Additionally, the study provides experimentally testable predictions linking cognitive deficits to disrupted thalamocortical connectivity, imbalances in prefrontal excitation-inhibition, and dysfunctional inhibitory cell types.
While I found the work to be of high interest, consider this topic area to be of substantially broad impact to reach a broad audience, and laud the combination of modelling and data, I have a number of concerns with the work as presented that I hope can be remedied through a revised version.
• Model Construction: The connections of the different interneuron populations is typically distinct to different populations of pyramidal neurons, both in terms of layers and compartments. The same is true for different thalamic populations, particularly those that putatively contact distinct interneuron populations. This is far from a mere triviality, as the role of SST+ and PV+ interneurons is typically ascribed to gating dendritic vs. somatic activity, respectively, and thalamic neurons can also target distinct compartments of different classes of pyramidal neurons. For this reason, I worry that the model is highly simplistic and will likely lead to spurious and/or misleading results. This would not be an issue if the model was created at the ‘mean field’ level, however the cellular-level precision of the inhibitory neuron classes suggests that the authors intended for the model to be interpretable at the cellular level.
• Lack of Suitable Network Variants: The authors do not appear to have tested whether other, less elaborate, networks were capable of providing a similar fit to data. For instance, if RNNs without all three IN populations perform as well as the reported version, this would argue against the need for including what might be otherwise unnecessary biological detail in their model. This is (unfortunately) a common side-effect of RNNs, in that they are capable of fitting to a wide-array of functions, but at the detriment that it can be quite challenging to infer precisely how the networks were capable of said fitting.
• Model Generalizability: While the computational model is biologically constrained and successfully replicates key findings in mice, its generalizability to other species, particularly humans, is unclear. The reliance on mouse-specific thalamocortical connectivity may limit the broader applicability of the findings.
• Experimental Validation: Although the model makes experimentally testable predictions, the study does not provide direct experimental validation of these predictions. The absence of empirical data to support the model's hypotheses weakens the overall impact of the findings.
• Simplification of Neural Dynamics: The model may oversimplify the complex dynamics of the prefrontal cortex and mediodorsal thalamus interactions by focusing on specific connectivity patterns and cell types. This reductionist approach might overlook other critical factors influencing cognitive flexibility and decision-making.
• Contextual Limitations: The study primarily focuses on context-dependent decision-making and cue uncertainty. While these are important aspects of cognitive control, the work does not address how the MD thalamus might interact with the PFC in other cognitive domains, potentially limiting the scope of its conclusions.
• Lack of Behavioral Correlation: The study primarily relies on computational modeling and neurophysiological findings, with limited emphasis on how these relate to observable behavioral outcomes. A stronger connection between the model's predictions and real-world behavioral data would enhance the study's relevance and impact.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
In this manuscript, the authors generate a PFC-MD model that incorporates known macro- and micro-circuits. They found that their model recapitulated known behavioral results and neural tuning during behavior as well as generated experimentally testable predictions. Overall, the computation modeling and data are well developed and detailed, the topic of MD-PFC interactions in support of cognition is important and of wide interest, and the overall depth of analysis is tremendous. Missing from much of the first draft of the manuscript is satisfactory statistical analysis to fully justify many of the

claims that are made – having this would greatly strengthen the manuscript.

COMMENTS:
1. Figure 1: What percentage of the PFC network are the different neuron types (Exc, PV/SOM/VIP)?
2. line 128: “higher/lower density/sparsity” is confusing; recommend separating (“higher density/lower sparsity…”
3. “congruence” is not well defined
4. line 145 – there is a switch from discussing cue uncertainty to “cue conflict”, which seems like the same thing as “cue uncertaintly”, but then why the language change? Furthermore, isn’t lines 145-148 a recapitulation of the description in lines 125-128 (which is probably the more thorough of the two)?
5. Figures 1e and 1f and associated text: there is no statistical basis for the statement that rule errors increase with cue uncertainty; this is needed; simplest would be to do a two way ANOVA (factors: error type and level of uncertainty) followed by post-hoc testing
6. Same issue in Figures 1d, 1i, and 1j and associated text; need statistics here to justify the statement that the PFC-MD model outperformed the PFC-alone model
7. Figure 2d and 2e show the ratios nicely but are not really connected to percentages discussed in the corresponding text. It is only in the methods that it is mentioned that this is via rank sum test. More information would be helpful here: perhaps dark shade for significantly tuned neurons and light shade for non-tuned?
8. Figures 2f, 2g, 2h and associated text – how was it determined that 25-30% of MD1 unites modulated their firing rates with respect to cue uncertainty? Pairwise test as described in methods would not be appropriate here. In some case the differences look like they wouldn’t be significant.
-Figure 2i and associated text – it looks like almost every MD neuron codes for both rules, with the timing of the firing rate increase corresponding to different rules. Does this match real data?
9. Figure 2j – again need statistical justification for the statements about differences; 2 way ANOVA would be straightforward. The (likely) increased context coding in MD is also not mentioned in the main text.
10. Figure 2k - Not sure why this figure is in black and white. Also, the concept of decoding accuracy is difficult to wrap your head around here. Are you decoding the correct rule, or the rule they actually implement? One could speculate that under uncertainty the rule matches what is correct less often, but that the rule representation in the PFC corresponds to their actual choice (right or wrong) nonetheless.
11. Figure 3m – again need to report statistics to support statement that the principal angle shifts in relation to uncertainty
12. Figure 4 and Supplementary Figure 4 – Panels are easy to follow here but all panels need statistical analysis to justify statements about differences. 4g lists three p-values apparently using what is typically a pairwise test, so this is confusing as well. The fact that the accuracy floors, even at the highest uncertainties, are often well above 25% seems potentially problematic. Shouldn’t the floor be 25% and no lower or higher at max uncertainty (see 4b baseline, for example; even lower, 4d; down close to 20% 4e)? Furthermore, it is buried in the methods that chance accuracy is 25%, apparently because correct means correct rule AND choice. This is not obvious in the main body of the manuscript and a little bit strange because “correct” in a behavioral context (simulated or not) more typically refers to whether the choice was correct, period, irrespective of whether the organism was implementing the optimal strategy. If you’re going to stick with this convention, then it should be articulated clearly in the main body of the manuscript, and the ylabels should be renamed so the meaning is clearer.
13. The 4e and 4f insets are also difficult to comprehend and very hard to see.
14. Line 281 – what are these p-values? Post-hoc tests? A series of pairwise comparisons? Is there a multiple comparisons correction?
15. I don’t understand how 5a and 5b show that “MD inhibition” alters PFC Exc rule tuning. And again, no statistical analysis is done to support the contention that “these operations changed tunings of some PFC excitatory units during the delay period and reduced choice at the population level during the target period.”
16. line 301 – it’s a large leap from a model of PFC-MD dynamics to the neural bases of cognitive biases; probably best to remove and save for discussion.
17. line 303 “observed a more robust task performance…” I think you mean to say a more robust effect on task performance
18. in Figure 5d you now have performance that goes down to 10% (well below chance according to the methods) and only down to 50% (well above chance) depending on which simulated manipulation is being used. Both seem problematic for the reasons described above.
19. there is no legend in Figure 5 for panels 5a,b,f.
20. line 308-310 “A closer examination of single-unit and population responses revealed that bidirectional SOM-VIP amplification produced emergent rule tunings and enhanced rule discriminability (Fig 5f)”. Really not sure what the basis is for this statement.
21. line 327 “Notably, this modified corticothalamic connectivity produced similar psychometric curves (Fig 6b)…” Need to state similar to what and ideally show and statistically examine the relationship.
22. line 329 “Fig 6” which panel?
23. Figures 6e, 6f, 6g; stats marked and described as Wilcoxon in figure caption; however, there are many comparisons here; I would suggest a two-way ANOVA followed by multiple comparisons post-hoc testing. Furthermore, in the caption, there are numbers following the p-values that are parentheses; no idea what those are.
24. correct grammar in line 348
25. line 344 – I recommend briefly describing the modified version of the CAC subtask and how it introduces the new features
26. line 362 – correct grammar
27. Figure 7g – please interpret, describe, and use stats
28. Figures 8b-d and associated text – need stats
29. line 413 – recommend restating in language that more specifically describes the effects of MD inhibition and activation
30. line 510 “Our analyses of corticothalamic and thalamocortical projections suggest an asymmetric information flow and

distinct contributions in cognitive control." Please follow this by reiterating what the asymmetry and differential contributions are
31. "Our results are also consistent with the idea that the MD can facilitate GABA shift" what does this mean?
32. It would be nice if the discussion better delineated the differences in the computational basis of how the MD supports working memory vs cognitive flexibility; which subtypes, which direction in the circuit etc
33. Overall, thorough well written Methods
34. line 648: Why equal percentages of the different interneuron subtypes? This is does not match real proportions. Also, what percentage are excitatory and inhibitory neurons in the network? I ask this earlier, but this should probably be noted in both locations due to its importance.
35. line 701-702 missing word
36. Lines 721-722 – earlier you state the # of MD1 and MD2 neurons in your model is equal, but then here you use an unequal # (18 vs 12); why?
37. line 747: How are you doing three way comparisons using rank sum? Also of note, it is generally when or where in the results you used this procedure

(Remarks on code availability)
I confirmed that the code is available on github and has an accompanying readme that describes how to generate the main figures in python.

Version 3:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors did their best to address our concerns. We recommend mentioning in the text that the approach incorporating many cell types has conceptual limitations, which new supp table 1 does not fully mitigate.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have clearly undertaken an extensive set of analyses to appease the reviewers. Although they superficially answered a number of my concerns, I remain underwhelmed by the strength and specificity of the evidence provided by their approach.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)
Overall, the manuscript is greatly improved via clarifications of the model and tremendously improved statistical analysis. The authors did a thorough job with the rebuttal and associated revision. Overall, the computational modeling and data remain well developed and detailed, the topic of MD-PFC interactions in support of cognition is important and of wide interest, and the overall depth of analysis is tremendous. With the elaborations in the revision, the manuscript is much more methodologically sound, and additional detail has been provided that will enhance reproducibility. The only comment I have that does not preclude publication is below.

MINOR COMMENT:
-please re-review Figure 7g and associated text; I think it could be improved further via some light revision and/or additional stats, but it is not absolutely necessary; will leave it up to the reviewers

(Remarks on code availability)
There is a clearly written readme
Folders are well organized
There is a link in readme to trained models, but many of the folders in the google drive folders are empty

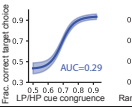
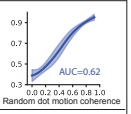
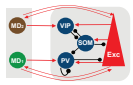
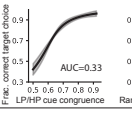
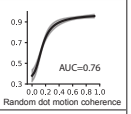
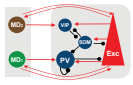
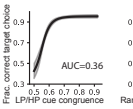
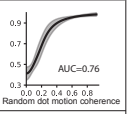
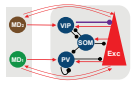
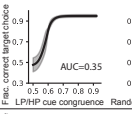
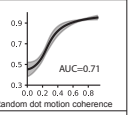
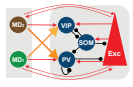
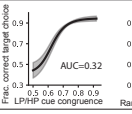
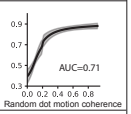
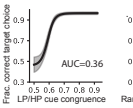
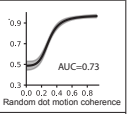
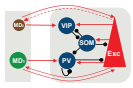
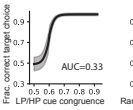
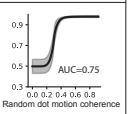
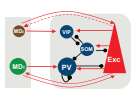
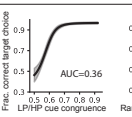
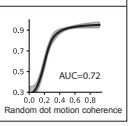
REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

In this interesting paper, the authors model the mediodorsal (MD) thalamus and prefrontal cortex (PFC) using biologically grounded cell types and projections, and test the model on context-dependent decision-making tasks. The findings indicate that the MD-PFC model outperforms a PFC-only model, with potential implications for computational deficits in schizophrenia.

Overall, the paper provides meaningful insights, but the complexity of the model, with many free parameters, makes it difficult to interpret how the various components contribute to emergent dynamics and computations. Additional clarification would greatly benefit the reader.

Thanks for your insightful comments. We agree with the reviewer that a systematic investigation of biologically-constrained models and result comparison on different model assumptions (such as the PFC-alone model, the disrupted MD model, the disrupted specific thalamocortical or corticothalamic connectivity) can lead to meaningful interpretations. Despite many free parameters and model simplifications, we believe that our modeling paradigm provides a principled framework to examine the contribution of individual components to emergent dynamics and computations. We now provide a new **Supplementary Table 1** (see below) to summarize or highlight these points.

	Network connectivity	MD unit tuning				Psychometric Curves	
		Rule	Context	Cue Uncertainty	Cue Sparsity	CAC	RDM
PFC-alone (baseline)		N/A					
Fig. 1a N_{Exc}=205 N_{PV}=17 N_{VIP}=17 N_{SOM}=17 N_{MD1}=N_{MD2}=100		✗ ✗	✓ ✓	✗ ✓	✗ ✓		
N_{Exc}=205 N_{PV}=21 N_{VIP}=15 N_{SOM}=15 N_{MD1}=N_{MD2}=100		✗ ✗	✓ ✓	✗ ✓	✗ ✓		
w/o Inh specificity		✗ ✗	✓ ✓	✓ ✓	✓ ✓		
w/o MD specificity		✗ ✗	✓ ✓	✓ ✓	✓ ✓		
no cell type N_{Exc}=205 N_{MD}=200 N_{Inh}=51		✗	✓	✓	✓		
Fig. 6a N_{Exc}=205 N_{PV}=17 N_{VIP}=17 N_{SOM}=17 N_{MD1}=120 N_{MD2}=80		✗ ✗	✓ ✓	✓ ✓	✓ ✓		
N_{Exc}=205 N_{PV}=21 N_{VIP}=15 N_{SOM}=15 N_{MD1}=120 N_{MD2}=80		✗ ✗	✓ ✓	✗ ✓	✗ ✓		

Reviewer Comment 1.1 — The manuscript does not sufficiently engage with existing literature on thalamocortical circuit models. Key concepts such as the alignment between cortical PCs and thalamus, neuron counts, and the role of thalamus in preparatory activity should be better contextualized using recent studies (e.g., Lakshminarasimhan et al. 2024, Logiaco et al. 2021, Kao et al. 2021).

Answer: Thanks for your insightful comments and pointers. We agree with the reviewer that the importance of discussing the relevant literature on thalamocortical circuit models. While several prior modeling studies mentioned by the reviewer focused on different task behaviors, the general modeling concepts using RNN are rather similar. Some analysis strategies (such as the relative cortex-thalamus size, examining the error with respect to the thalamus size, thalamo-cortical control) are similar. We now cite these important studies at several places in both Results and Methods.

"In the literature, several thalamocortical circuit models have been proposed based on RNNs. Incorporating biological details (such as the thalamus size) and thalamocortical connectivity constraints has been proven important for learning." ***(Lines 96-98)***

"We have taken into consideration of several thalamocortical circuit models developed for different task behaviors in the literature." ***(Lines 1026-1027)***

"From a control perspective, the MD thalamus can be viewed as a controller that operates on the cortical plant, which enables the context switch. This is reminiscent of the role of thalamus in optimal anticipatory control in motor preparation preparatory" ***(Lines 1276-1278)***

Reviewer Comment 1.2 — The decision to provide sensory inputs to both PFC and MD is unclear given current anatomical evidence. The authors should clarify whether MD receives direct sensory inputs or exclusively corticothalamic inputs, as this would improve the model's biological accuracy.

Answer: Thank you for highlighting this important distinction. We agree with reviewer that the MD exhibits minimal connectivity with the sensory pathways and receives its predominantly driving input from layers V and VI of the PFC areas through the corticothalamic pathway (Mitchell and Chakraborty, *Front. Syst Neurosci.* 2013). The MD also receives some modulatory inputs from the subcortical regions such as forebrain, midbrain and brainstem. In the thalamus, only primary sensory nuclei (first order) receive direct sensory inputs.

In addition to the prefrontal regions, there is some functional connectivity between the MD subregions and different functional networks. Human diffusion MRI studies have also shown that the connectivity strength between the MD subregions and the visual network, the somatomotor network, the ventral attention network, and the dorsal attention network presents an increasing trend from MDm (medial) to MDc (central) to MDd (dorsal) to MDl (lateral) (Li et al. https://doi.org/10.1016/j.neuroimage.2022.118876). See also another independent report (Kumar et al., 2022, https://www.nature.com/articles/s42003-022-04126-w).

We also note that the PFC does NOT receive direct sensory input either. For instance, visual input usually arrives at the frontal-eye-field (FEF) in the frontal lobe through extrastriate visual areas, whereas auditory input usually arrives from the afferents of the early part of cortical auditory system (e.g., auditory belt cortex) to the PFC. However, direct sensory pathways were commonly assumed in all computational models of the PFC in the literature, without modeling the intermediate pathway. In our current PFC-MD model (Fig. 1a), we have assumed that the sensory input may arrive indirectly to the MD (through PFC), whereas the PFC also receive the sensory input indirectly.

That said, we are certainly aware of the potential impact on result interpretation induced by this assumption. To address this concern, we have further investigated the impact of different input assumptions on the results. First, in the two-context (CAC and RDM) task, we found that our results were robust by removing the direct sensory input to the MD structure.

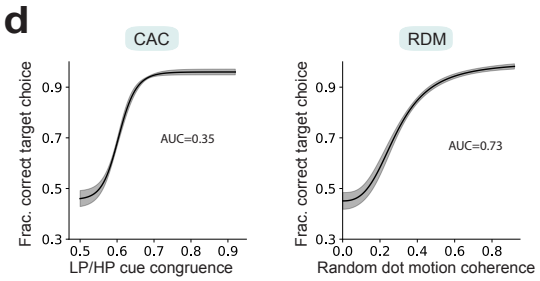
(1) The PFC-MD model performance or accuracy remained similar in terms of AUC statistics (**Response Fig. 1**, compared to Fig. 1d in the main text)

(2) PFC units remain tuned to the rule in both CAC and RDM contexts (**Response Fig. 2**, compared to Fig. 2d,2e in the main text)

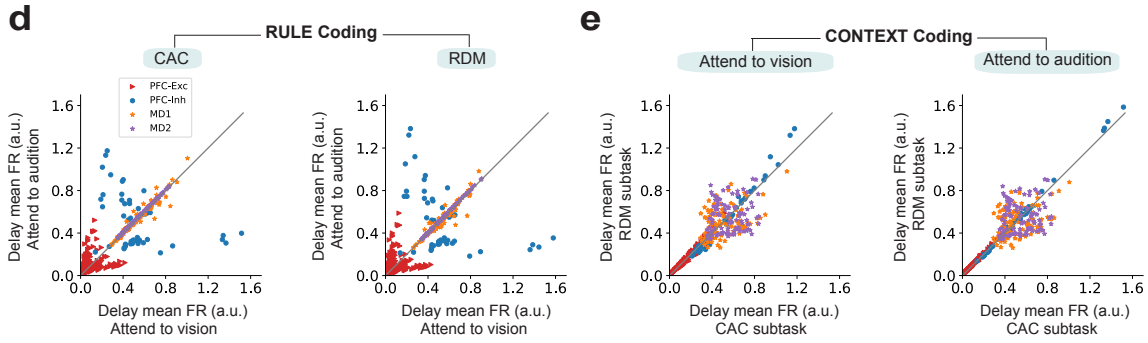
(3) MD units remain tuned to context (**Response Fig. 3**, compared to Supplementary Fig. 2d, 2e in the main text).

(4) MD units remain modulated with cue uncertainty in both CAC and RDM contexts.

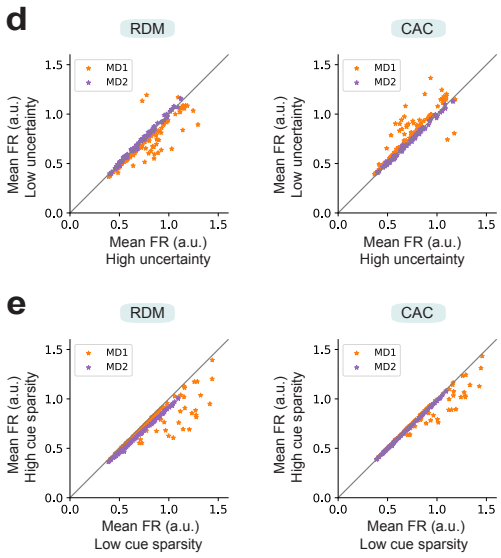
For result comparisons, we now list these three Response Figures to show that two model assumptions (i.e., MD without direct sensory input vs. MD with direct sensory input) lead to qualitatively similar results.



Response Fig. 1 (Compared to Fig. 1d in the main text (PFC-MD model, black curve)). **d** Psychometric curves of CAC and RDM contexts. The area under the psychometric curve (AUC) was reported.

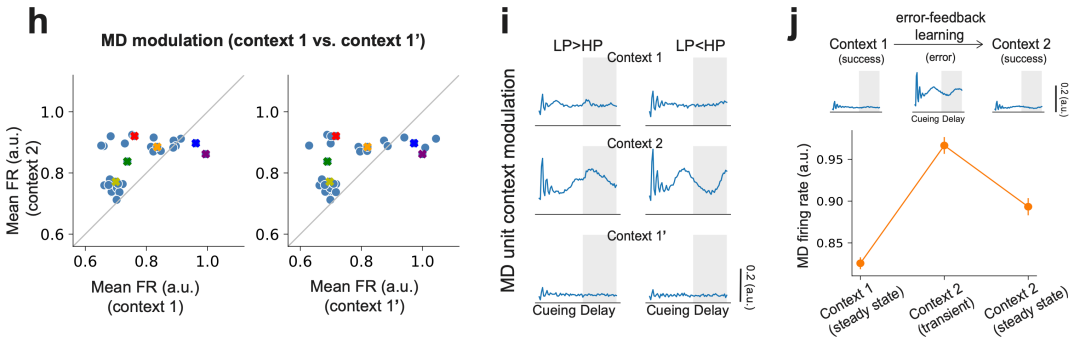


Response Fig. 2 (Compared to Fig. 2d, 2e in the main text). **d** Population statistics of mean firing rates (during the task delay period) of excitatory PFC units and MD units for encoding rule. **e** Population statistics of mean firing rate (during the task delay period) of excitatory PFC units and MD units for encoding context.



Response Fig. 3 (Compared to Supplementary Fig. 2d, 2e in the main text). **d** Population statistics of mean firing rates (during the cueing period) of MD_1 units for encoding cue uncertainty. **e** Population statistics of mean firing rates (during the cueing period) of MD_1 units for encoding cue sparsity.

Next, in the context-switching task, we also found qualitatively similar results derived from two model assumptions. See the **Response Fig. 4** (assuming that MD without direct sensory input) compared to Fig. 7h, i, j in the main text.



Response Fig. 4 (Compared to Supplementary Fig. 7 h, I, j, in the main text). **h** MD units showed context-invariant firing. **i** Turning curve illustrations of one MD unit during context switching. **j** A subset of MD units (n=8 out of 30 from one trained PFC-MD model) showed increased modulation with respect to decision error during the transient switching stage. The MD firing rate was averaged across the delay period when the change of network state became very small (i.e., steady state). Tuning curves of one MD were shown on the top.

Based on these observations, we conclude that changing the MD input assumption would NOT change our conclusion. Therefore, we have decided to keep most of result figures unchanged. Instead, we now discuss the model assumptions and possible discrepancy from biological reality in Discussion and Methods sections. For instance,

"We have also considered a scenario where there is no direct sensory input $\mathbf{u}$ in Equation 3, then Equation 4 can be further simplified. Since the results were qualitatively similar, we have only presented the results in the general case." ***(Lines 1050-1052)***

Additionally, in some mathematical equations of the Methods section (e.g., fixed-point analysis, cue-to-rule transformation), we have included discussions both conditions (with $\mathbf{V}^{in}$ and without $\mathbf{V}^{in}$**).**

Finally, we provide demo codes for PFC-MD models with two different assumptions (with vs. without direct sensory input to MD). Our shared code would allow researchers to modify the assumptions according to their need (such as exclusion of sensory input, network size, connectivity and task design) to test the variability in model predictions.

Reviewer Comment 1.3 — Line 114: There is a lack of clarity around how the number of units in the PFC-MD and PFC-only models were matched. It is important to specify whether the performance improvement with MD is due to the network architecture or simply the higher number of units.
Answer: Thanks for your comment and we apologize for the missing information. We add one sentence in the main text to clarify this point.

"The total number of units in the PFC-MD model is 256 (N_{PFC}+ N_{MD}=256), whereas the matching number of units in the PFC-alone model is 256 (N_{PFC}=256)." ***(Lines 1055-1057)***

In the single-modal context-switching task (Fig. 7a), we also systematically changed the number of MD units to compare the performance in switch time (Fig. 8b). This information has been documented in the Results and Methods sections. Finally, we have also changed various network architectures or cell-type percentages (**Supplementary Table 1**) and confirmed our results. Based on these results, we conclude that the PFC-MD performance improvement related to PFC-alone is largely due to the network architecture difference.

Reviewer Comment 1.4 — The task description could be simplified to enhance readability. Consistent use of terms like "context" and "subtask" would reduce confusion and make the task setup easier to follow.
Answer: Thanks for your insightful comments. We now use the term "context" consistently in two-modality context-dependent decision-making task (i.e., combined CAC and RDM) to avoid confusion.

We also simplify the task description to enhance readability in the Methods section (see below).

"The computational task was modified from a combination of a two-context cross-modal four alternative forced choice (4AFC) task[18] and a single-modal two alternative force choice (2AFC) task with ambiguous cues[6]. **Figure. 1b** illustrates the two cueing contexts that were mapped to the same rule: attend-to-audition or attend-to-vision. During the 800-ms cueing period, the network received cueing signal from either visual (moving dots) or auditory modality (LP/HP pulses), each of which had various degrees of cueing uncertainty. Under visual modality, the random dot motion (RDM) context used moving dots as cues, where the coherence of moving dots in time indicates the degree of cueing uncertainty: with coherence 1 meaning all dots moving with the same direction (rightward, rule 1 vs. leftward, rule 2) and coherence 0 meaning completely random direction. The dimensionality of cue input in RDM was 64, and each dimension

represents one random dot. We used the set [-1,1] to represent the direction of dot moving: -1 corresponding to leftward, -1 corresponding to rightward, and the value of other direction is selected randomly between -1 and 1. Under auditory modality, the conflicting auditory cue (CAC) context used a sequence of conflicting pulses as cues: low-pass (LP) pulse (+1) or high-pass (HP) pulse (-1). The {+1,-1} pulses were task-relevant cues, and the background was represented by 0 ('blank'). In training, the number of pulses during the 800-ms cueing period was 32, yielding a cue signal density 40 s^{-1}. The ratio between the numbers of two pulses defines the auditory cue congruence. Accumulating the cueing evidence would produce the indication of one rule (more -1's) or the other (more +1's). During the cueing period, the evidence of ambiguous cues needs to be accumulated in order to identify the rule. During 800-ms delay period of working memory, rule information needs to be preserved. During the 200-ms target period with "divided attention", the auditory and visual stimuli (independent from the cueing period) were simultaneously presented (e.g., upsweep/downsweep tone that was mapped to auditory left/right choice, respectively; yellow/blue light that was mapped to visual left/right choice, respectively), and each attended target was associated with a unique outcome of four choices. Therefore, the learning goal was to attend to the correct modality and select the one correct choice (**Fig. 1b**). When errors occurred, the task error could be ascribed to either "rule error" (or "executive error", which occurred in the cueing or delay period) or "sensory error" (which occurred in the target period), or both (**Supplementary Table 1**).

In theory, the chance-level accuracy for the 4AFC task is 25%. However, the rule error and sensory error are independent in the task design. For the rule error alone, the chance-level accuracy is 50%. To introduce a proper level of sensory error, we assumed two competing sensory inputs were stochastic and drawn from normal distributions with two different mean values but the same variance. Increasing the variance would increase the sensory error; we selected a variance value that best matched to the experimental data. In the extreme case where sensory inputs were deterministic, the model prediction would lead to only rule error." ***(Lines 937-970)***

Reviewer Comment 1.5 — Clarification is needed regarding whether RDM and CAC trials were interleaved or blocked, and whether the presentation of auditory and visual stimuli was mutually exclusive.

Answer: Thanks for your comment and apology for the missing information. The RDM and CAC trials were interleaved. Importantly, during the 800-ms cueing period, the auditory and visual cues were mutually exclusive; during the 200-ms target period ("divided attention"), auditory and visual stimuli are always presented simultaneously, which is the key component of divided attention. We now highlight these facts in the Methods section.

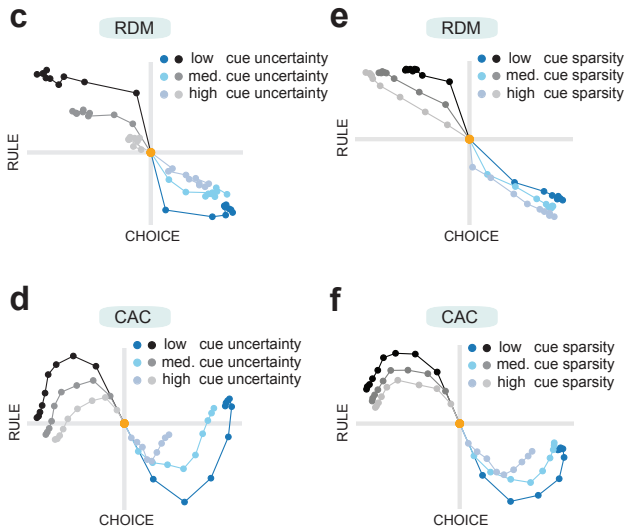
"During the 200-ms target period with "divided attention", the auditory and visual stimuli (independent from the cueing period) were simultaneously presented" ***(Lines 955-957)***

"The two contexts (CAC and RDM) were learned simultaneously and their trials were interleaved." ***(Lines 1057-1058)***

Reviewer Comment 1.6 — Figures 3c-f require clearer labeling, particularly the x-axis, and the

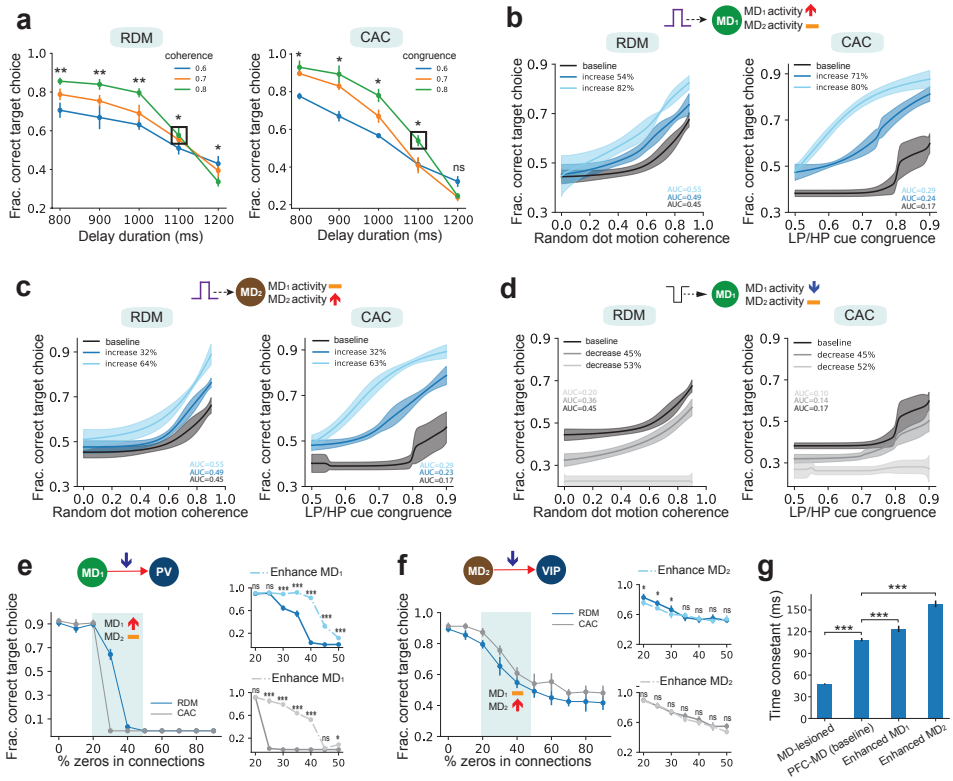
discrepancy between the text and the panels in Figure 4 should be resolved, as it may affect the interpretation of the results.

Answer: Thanks for your comment and we apologize for the missing information. We now provide labeling of x-axis in all panels and resolve all discrepancies between texts and panels in revised Fig. 3c-f below.



Response Fig. 5 (Updated Fig. 3c-f in the main text).

We have now updated all panels (Fig. 4b-g) and provided additional AUC statistics. Details of statistical tests and p-values are provided in figure legend. We have tried to resolve all discrepancies between text and figure panels. The revised figure and legend as shown below.



Response Fig. 6 (Updated Fig. 4 in the main text).
a Increasing the duration of task delay period reduced the task performance under cue uncertainty in both contexts. Error bar denotes SD (n=10). Statistical tests were conducted between green and blue curves (pointwise) for each delay duration condition (*p<0.05, **p<0.01, rank-sum test). The task condition marked with "□" was used in the illustrations of the three remaining panels. **b,c** Increasing the MD_1 or MD_2 population firing rate during an elongated delay period improved the working memory and the psychometric curves in two contexts. Black curve denotes the baseline, and the number of % denotes the relative increase in subpopulation firing rate. With MD_1 activation, the derived AUC increased from the baseline (RDM: dark blue vs. black, p=0.016; light blue vs. black, p=0.009; CAC: dark blue vs. black, p=0.0001; light blue vs. black, $p=0.7x10^{-4}$). With MD_2 activation, the derived AUC increased from the baseline (RDM: dark blue vs. black, p=0.028; light blue vs. black, p=0.009; CAC: dark blue vs. black, p=0.0001; light blue vs. black, p=0.0002). **d** Decreasing the MD_1 population firing rate degraded the working memory and the psychometric curves of two contexts. With MD_1 suppression, the derived AUC increased from the baseline (RDM: dark grey vs. black, p=0.014; light grey vs. black, p=0.009; CAC: dark grey vs. black, p=0.0003; light grey vs. black, $p=0.7x10^{-4}$). **e** *Left:* Weakening MD_1→PV connections (by increasing the percentage of zeros) during the cueing period quickly reduced the task performance in both RDM and CAC contexts. *Right:* increasing MD_1 activity (dotted line) could rescue each context's performance under a wide range of connectivity conditions (shaded area: 20%-50% of zeros). Error bar denotes SD (n=10). Statistical tests were conducted between dark blue and light blue curves (pointwise) for condition (* p<0.05, ***p<0.001, rank-sum test). **f** Weakening MD_2→VIP connections only degraded the task performance slowly in both RDM and CAC contexts, but increasing MD_2 activity had no or little effect on the change in task performance. Statistical tests are similar to panel (**e**). **g** Impact of network time constant under different thalamocortical connectivity manipulations. Error bar denotes SD (n=20 independently trained PFC-MD models). The MD_1 and MD_2 enhancement corresponded to the experiments where respective thalamocortical connections were strengthened during working memory. All statistical tests were rank-sum tests against the PFC-MD baseline (MD-lesioned vs. baseline: p<0.0001; enhanced MD_1 vs. baseline: p<0.0001; enhanced MD_2 vs. baseline: p<0.0001; Bonferroni-corrected rank-sum test)

We also revise the main text as below.

"To demonstrate the role of MD in working memory maintenance[18], we selected the condition of coherence/congruence of 0.8 in respective RDM/CAC contexts (see the box symbol in **Fig. 4a**) and increased the firing rate of MD_1 or MD_2 subpopulation by a multiplicative gain during an elongated 1100-ms delay period while keeping the other MD subpopulation unchanged. Such MD activations led to a slight boost in rule maintenance and subsequent decision-making accuracy under a wide range of cue uncertainty levels. Notably, it required less firing rate elevation for MD_2 than MD_1 to achieve the same level of task improvement or AUC statistic (**Fig. 4b,c**), probably because the MD_2→VIP pathway has an amplification effect on increasing the target sensitivity for working memory maintenance. In contrast, suppressing the MD_1 activity while keeping the MD_2 activity unchanged during the delay period had a negative effect on task accuracy or AUC statistic (**Fig. 4d**)." ***(Lines 359-369)***

"To distinguish the role of two specific thalamocortical projections (MD_1→PV and MD_2→VIP) in regulating prefrontal computation, we disconnected one of two thalamocortical projections or set the respective thalamocortical connections to zeros (i.e., "computational lesion") while keeping the other intact. We found that weakening MD_1→PV connectivity (by setting a small percentage of connections to zeros) reduced the performance rapidly for both RDM and CAC contexts (**Fig. 4e**). Within a narrow range (20-50% sparsity, shaded area in the left panel of **Fig. 4e**), enhancement of phasic MD_1 activity could boost task performance in a wide range of sparsity conditions (**Fig. 4e**, dashed lines of two smaller panels)." ***(Lines 407-414)***

Reviewer Comment 1.7 — In the schizophrenia-related simulations, it is unclear whether models with

E/I imbalance were retrained before testing. This is crucial to understanding whether balance is needed for learning or simply for maintaining performance.

Answer: Thanks for your insightful comments. It is important to clarify the following facts. In the schizophrenia-related simulations, we have conducted two types of experimental testing: one is related to the impact of noise sensitivity on working memory with an E/I imbalanced model, the other is related to the impact of E/I imbalance on switching deficit.

In the first experimental testing (Fig. 9a), we first trained the PFC-MD model to learn the task in context1 (control model), and then manipulate E/I imbalance on the trained model in the testing phase, resulting in the decreasing in task performance. Activating MD subpopulation will improve the task performance. Therefore, the model with E/I imbalance was NOT retrained, meaning that balance is needed for maintaining task performance.

In the second experimental testing (Fig. 9b, c), we first trained the PFC-MD control model to learn the task in context1, based on this pretrained model, we further trained it to learn the context2. This switching time is treated as the baseline. We further investigated the impact of E/I imbalance on switching time deficit based on the pretrained model. Specifically, the E/I imbalanced pretrained model was trained learning the task in context2. We found that the E/I imbalanced model caused switching deficit during the reversal context-switching task compared to the baseline. Therefore, the balance is needed for learning. We now provide more details in the Methods section (see below).

"**Computer simulations in a schizophrenia mouse model**

We conducted computer simulations to reduce prefrontal cortical PV inhibition in the PFC-MD model in an attempt to mimic one of known changes in schizophrenia. In the CAC context switching task, we first trained the PFC-MD model to learn the task in *Context 1* (control model), and then manipulated E/I imbalance on the trained model (imbalanced model). We compared their task performances in the presence of noise during working memory. We further activated MD subpopulation and reexamined the impact on task performance. Next, during context switching, we retrained both the control model and imbalanced model to learn *Context 2* and computed their switch time for comparison. We further activated MD subpopulation and reexamined the impact on the switch time. Finally, we repeated computer simulations in a cross-modal cueing context-switching task setting (e.g. CAC→RDM)." ***(Lines 1111-1120)***

Reviewer Comment 1.8 — Line 384: More explanation is required about the relationship between MD population size and external input dimensionality, particularly if the input size remained constant during simulations.

Answer: Thanks for your insightful comments. To detect the impact of MD on the learning speed of context switching, we varied the size of MD population (N_{MD}=2,8,16,30) and find that more MD units can improve the switching speed. During the process, we kept the number of PFC and input dimensionality unchanged. The external input dimensionality remained constant, which was determined by the task a priori. We now add one sentence to clarify this fact in the main text.

"We systematically varied the size of MD population (N_{MD}=2,8,16,30; N_{PFC}=256) and retrained the respective PFC-MD networks, during which the input dimensionality remained constant (N_{in}=6)." ***(Lines 674-676)***

Methods:

Reviewer Comment 1.9 — The description of the cost function, noise scale, and weight initialization is insufficient. More details in these areas would provide a clearer understanding of the modeling approach and allow the results to be more easily verified.

Answer: Thanks for your insightful comments and we agree with reviewer that this information is important for enhancing result reproducibility (in addition to the open-source software). We have now added more details in the Methods section.

"

$$\tau \dot{\mathbf{x}}^{\mathrm{PFC}} = -\mathbf{x}^{\mathrm{PFC}} + \mathbf{W}^{\mathrm{rec}}\mathbf{r}^{\mathrm{PFC}} + \mathbf{W}^{\mathrm{in}}\mathbf{u} + \sqrt{2\tau\sigma_{rec}^2}\boldsymbol{\xi}. \qquad (1)$$

where τ denotes the time constant (we used $\tau = 20$ ms), $\boldsymbol{\xi}$ denotes additive N_{PFC}-dimensional Gaussian noise, each independently drawn from a standard normal distribution, and $\sqrt{\sigma_{rec}} = 0.05$ defines the scale of the noise standard deviation; $\mathbf{W}^{\mathrm{rec}}$ is an $N_{\mathrm{PFC}} \times N_{\mathrm{PFC}}$ matrix of recurrent connection weights, and $\mathbf{W}^{\mathrm{in}}$ denotes an $N_{\mathrm{PFC}} \times N_{\mathrm{in}}$ matrix of connection weights from the input to network units (N_{in} =72), which includes four types of noisy input,

$$\mathbf{u} = \left(\boldsymbol{u}_{\mathrm{RDM}}, \boldsymbol{u}_{\mathrm{CAC}}, \boldsymbol{u}_{\mathrm{context}}, \boldsymbol{u}_{\mathrm{sensory}}\right) + \mathbf{u}_{\mathrm{noise}}$$
$$\mathbf{u}_{\mathrm{noise}} = \sqrt{2\tau\sigma_{in}^2}\boldsymbol{\xi}$$

Here the stimulus input $\boldsymbol{u}_{\mathrm{RDM}}$ denotes cueing input in the RDM context (N_{RDM}=64) and $\boldsymbol{u}_{\mathrm{CAC}}$ denotes cueing input in the CAC context (N_{CAC}=2). The context input $\boldsymbol{u}_{\mathrm{context}}$ indicates which task the network is performing in the current trial (N_{context}=2). The $\boldsymbol{u}_{\mathrm{sensory}}$ represents sensory input during the target period (N_{sensory}=4). The input noise strength is $\sqrt{\sigma_{in}} = 0.05$. " ***(Lines 977-988)***

"**Training procedure.** We trained the PFC-alone and PFC-MD models by supervised learning. The two contexts, CAC and RDM, were learned simultaneously. The CAC or RDM tasks consisted of a batch size of 64 simulated trials used for training (N_{trials}=64). The number of training trials for two rules (attend-to-audition vs. attend-to-vision) was balanced. The model was trained by minimizing a loss function using back-propagation through time (BPTT). The loss function $\mathcal{L}$ is computed by mean-squared-error (MSE) averaged over N_{trials} trials:

$$\mathcal{L} = \frac{1}{N_{trials}} \sum_{n=1}^{N_{trials}} \sum_{i}^{N_{out}} \mathbf{m}_i (\mathbf{z}_i - \hat{\mathbf{z}}_i)^2$$

where $\mathbf{z}_i$ denotes the *i*-th target output, and $\hat{\mathbf{z}}_i$ denotes the corresponding estimated network output, $\mathbf{m}_i$ denotes the error mask, which is a non-negative matrix that determines whether the squared error at time point t and output units i need to be computed. In the response epoch, $\mathbf{m}_i = 0$, in the other task epoch, $\mathbf{m}_i = 1$. We applied L_2 regularization in the loss function with a default regularization parameter 0.1. The model parameters were optimized using Adam optimizer, with default configuration of hyperparameters (learning rate parameter 0.0005; weight decay regularization parameter 0.1). The optimization continued

until the desired accuracy level of 0.98 was reached. The recurrent connection matrices {$\mathbf{W}^{\mathrm{rec}}$, $\mathbf{W}^{\mathrm{MD}\to\mathrm{PFC}}$, $\mathbf{W}^{\mathrm{PFC}\to\mathrm{MD}}$} were initialized using Kaiming initialization with the parameter $a = 6$ (He et al., 2015), which was designed to solve the vanishing and exploding gradient problems in practice. The input weights were initialized using a uniform distribution between -1 and 1, and the output weights were initialized with a Gaussian distribution of mean 0 and standard deviation $0.4/\sqrt{256}$." ***(Lines 1054-1075)***

Reviewer Comment 1.10 — It is unclear whether the input weights to PFC were fixed, which could explain the absence of context-coding in some results.

Answer: Thanks for your insightful comments. The input weights to PFC were not fixed; only the intracortical connection $\mathbf{W}^{\mathrm{rec}}$ was fixed. We now clarify this fact in the main text.

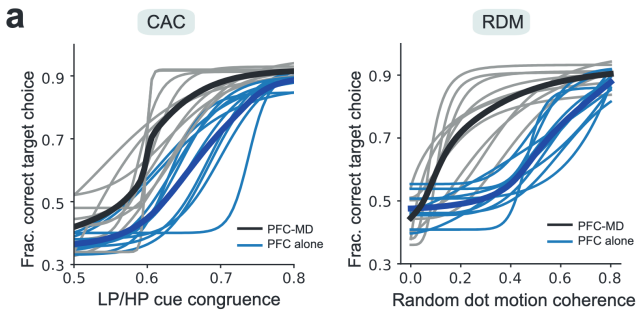
"Specifically, we adapted all other synaptic connections (except PFC-to-PFC connections) based on inter-trial error feedback to learn context switching on a trial-by-trial basis." ***(Lines 553-555)***

We have indeed conducted experiments where the input weights to PFC were fixed (Fig. 8d), the conceptual findings remained unchanged, but the modified setup could affect the relative switch time.

Reviewer Comment 1.12 — Instead of using a single representative model for performance plots, showing the ensemble average could offer a more comprehensive picture of model behavior?

Answer: Thanks for your insightful comments. We agree with the reviewer that ensemble average may provide a more comprehensive picture, despite the challenge of presenting all results by ensemble average.

In the revised Supplementary Fig. 1a, we have now provided ensemble average curves (thick solid black/blue curves) when comparing the performance of PFC-MD model (black) versus PFC-alone model (blue).



Response Fig. 7 (Updated Supplementary Fig. 1a in the main text). **a** Selected psychometric curves of PFC-MD and PFC-alone models (n=10 for each group) for both CAC and RDM contexts. The thick solid curves represent the respective ensemble averages.

Additionally, statistical characterizations of model behavior in terms of convergence speed were all based on ensemble average. These have been previously shown in Fig. 2j, 2k, Fig. 3k,3l, Fig. 4g, Fig. 7b, Fig. 8b, Fig. 9a,9b,9c,9d.

Note that we used the same PFC-MD model while comparing performances under different cue conditions (such as different cue uncertainty or sparsity levels, such as Fig. 1i) or under different MD activation/suppression conditions (such as Fig. 4b-d). In those experiments, no ensemble average was presented because the individual effects could vary among the trained models.

Minor Comments:

Reviewer Comment 1.12 — The term "computational lesion" is ambiguous and should be clarified.

Answer: Thanks for your insightful comments. Our term “computational lesion” refers to removing the connectivity between neurons and thalamocortical pathways. We simply by setting some connection weights to zeros. We now add a note in the main text.

“To distinguish the role of two specific thalamocortical projections (MD_1→PV and MD_2→VIP) in regulating prefrontal computation, we disconnected one of two thalamocortical projections or set their thalamocortical connections to zeros (i.e., “computational lesion”) while keeping the other intact.” ***(Lines 408-409)***

Reviewer Comment 1.13 — Minor typographical errors, such as the typo in line 637 and the incorrect matrix element in line 653, should be corrected.

Answer: We apologize for the oversight. They are now corrected.

$$\underbrace{\begin{pmatrix} 1 & + & + & + & - \\ + & 1 & + & + & - \\ + & + & 1 & + & - \\ + & + & + & 1 & - \\ + & + & + & + & -1 \end{pmatrix}}_{\mathbf{W}^{\mathrm{rec}}} = \underbrace{\begin{pmatrix} 1 & + & + & + & + \\ + & 1 & + & + & + \\ + & + & 1 & + & + \\ + & + & + & 1 & + \\ + & + & + & + & 1 \end{pmatrix}}_{\mathbf{W}^{\mathrm{rec},+}} \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{pmatrix}}_{D}$$

$$\mathbf{W}^{\mathrm{rec}} = \mathbf{W}^{\mathrm{rec,plastic}} \odot \mathbf{M}^{\mathrm{rec}} = \begin{pmatrix} w_{\mathrm{E}\to\mathrm{E}} & w_{\mathrm{PV}\to\mathrm{E}} & w_{\mathrm{SOM}\to\mathrm{E}} & w_{\mathrm{VIP}\to\mathrm{E}} \\ w_{\mathrm{E}\to\mathrm{PV}} & w_{\mathrm{PV}\to\mathrm{PV}} & w_{\mathrm{SOM}\to\mathrm{PV}} & w_{\mathrm{VIP}\to\mathrm{PV}} \\ w_{\mathrm{E}\to\mathrm{SOM}} & w_{\mathrm{PV}\to\mathrm{SOM}} & w_{\mathrm{SOM}\to\mathrm{SOM}} & w_{\mathrm{VIP}\to\mathrm{SOM}} \\ w_{\mathrm{E}\to\mathrm{VIP}} & w_{\mathrm{PV}\to\mathrm{VIP}} & w_{\mathrm{SOM}\to\mathrm{VIP}} & w_{\mathrm{VIP}\to\mathrm{VIP}} \end{pmatrix} \odot \begin{pmatrix} 1 & -1 & -1 & 0 \\ 1 & -1 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 1 & 0 & -1 & 0 \end{pmatrix}$$

Reviewer Comment 1.14 — Figures should be embedded within the text for easier reference, as the current layout makes the paper difficult to follow.

Answer: Thank you for your suggestion. We have rearranged the figures for better readability.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Thanks very much for your insightful co-reviewed feedback.

Reviewer #3 (Remarks to the Author):

This study investigates the mediodorsal (MD) thalamus's role in cognitive control, particularly in decision-making under uncertainty and cognitive flexibility. Using biologically-constrained computational models, the authors demonstrate that incorporating a feedforward MD structure into the prefrontal cortex (PFC) enhances robustness to low signal-to-noise ratios, improves working memory, and facilitates rapid context switching. The model successfully replicates key neurophysiological findings from mice, offering new insights into the computational mechanisms of the MD in regulating cue uncertainty and context switching. Additionally, the study provides experimentally testable predictions linking cognitive deficits to disrupted thalamocortical connectivity, imbalances in prefrontal excitation-inhibition, and dysfunctional inhibitory cell types.

While I found the work to be of high interest, consider this topic area to be of substantially broad impact to reach a broad audience, and laud the combination of modelling and data, I have a number of concerns with the work as presented that I hope can be remedied through a revised version.

Thanks very much for constructive suggestions. In response to your suggestions, we have tried to elaborate the model and provide additional results to improve the result presentation.

Reviewer Comment 3.1 — Model Construction: The connections of the different interneuron populations is typically distinct to different populations of pyramidal neurons, both in terms of layers and compartments. The same is true for different thalamic populations, particularly those that putatively contact distinct interneuron populations. This is far from a mere triviality, as the role of SST+ and PV+ interneurons is typically ascribed to gating dendritic vs. somatic activity, respectively, and thalamic neurons can also target distinct compartments of different classes of pyramidal neurons. For this reason, I worry that the model is highly simplistic and will likely lead to spurious and/or misleading results. This would not be an issue if the model was created at the 'mean field' level, however the cellular-level precision of the inhibitory neuron classes suggests that the authors intended for the model to be interpretable at the cellular level.

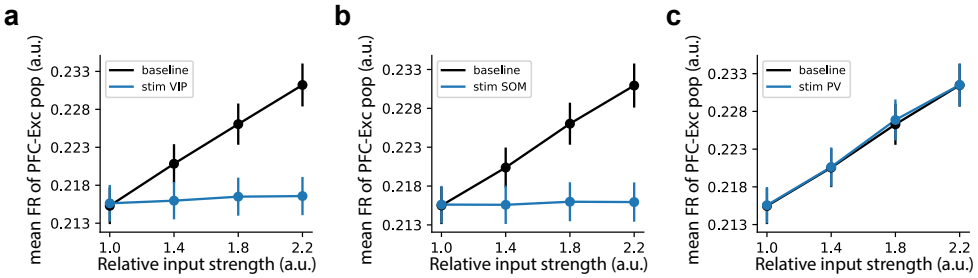
Answer: Thanks for the insightful comment. Our model construction was motivated from the mouse experiment (Mukherjee et al. 2021 Nature), in that MD1 (kainite receptor GRIK4-expressing) has a primary target on PV+ interneurons in the PFC, whereas MD2 (dopamine receptor-expressing) has a primary target on VIP+ interneurons. The difference in cortical interneuron cell-types determines their connectivity between themselves (Fig. 1a). For instance, SOM and VIP neurons mutually inhibit each other. SOM inhibits PV, and there is self-inhibition among PV neurons. We believe that our model is the first biologically-realistic cellular-level PFC-MD model developed for understanding the role of MD in regulating prefrontal computation to enable cognitive flexibility. This will have advantages over the mean-field model as the individual neuron tuning, population dynamics, subspace analysis would not be possible to characterize.

We also agree with the reviewer that our PFC-MD model was simplified related to the real biological circuit in that (i) we did not model the cortical layer structure; (ii) we did not model thalamic inhibition (from the TRN); (iii) we used a "point neuronal model" (i.e., not a compartment model), therefore we did not distinguish the somatic and dendrite targets from the interneurons; (iv) the current model is a rate model.

Nevertheless, we believe in the philosophy that "all models are wrong, some are useful". Despite the simplification, our model may serve as the first step to incorporate more biological details in the future model extension (such as spiking neural network models). We now present a detailed discussion on model limitations at the end of Discussion section.

"As a reductionist approach, our current PFC-MD model has also several methodological limitations. First, the current model is simplified related to the real biological circuit in that it does not account for cortical layer structure or thalamic inhibition, and it does not distinguish distinct somatic and dendrite targets from the cortical interneurons. Second, the current model uses a rate-based point-neuronal model instead of a spiking-based biologically realistic compartment-based model, and cannot model the local field potential (LFP) oscillations and between-region coherence. Third, the current model does not explicitly model inputs from cortico-cortical communications, top-down or modulatory input from other areas of the brain. Nevertheless, compared to the mean-field model, our cell-type specific model can validate MD and PFC tunings with respect to network architecture and thalamocortical connectivity; additionally, the reductionist approach may provide a principled paradigm to explore the key computational mechanisms of the PFC-MD circuit during cognitive flexibility and presents new paths for future investigations. For instance, we may extend the current model to a value or reward-based decision-making setting to match more broadly with animal or human experiments. Biologically inspired neural plasticity rule can be further imposed. We may also modify the decision-making task to model different dimensions of cognitive control and test the model generalizability across species and task behaviors." ***(Lines 916-932)***

Finally, to demonstrate the importance of three major interneuron cell types and their specificity of in input gain control, here we have done additional experiments to examine the role of different PFC interneuron cell types in input gain control. Specifically, we have plotted the "input-output mapping" between the input strength and the target output in terms of the mean firing rate (FR) of PFC excitatory population (see Response Fig. 8). We found that both VIP and SOM cell types serve the role of input gain control by modulating their activities to maintain a steady output, whereas the PV cell type cannot. These results suggest that distinct interneuron populations generated from this simplified cellular-based model can uncover key important mechanisms, which mean-field models are able to reveal.

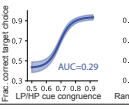
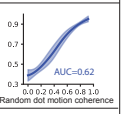
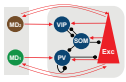
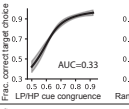
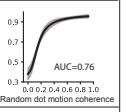
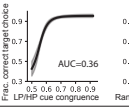
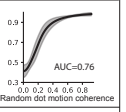
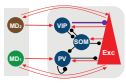
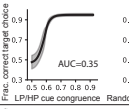
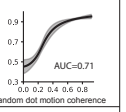
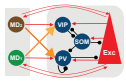
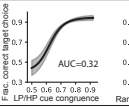
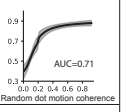
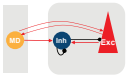
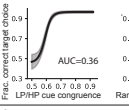
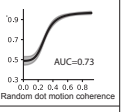
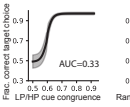
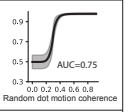
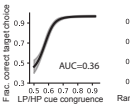
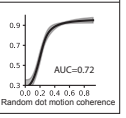


Response Fig. 8 The input strength was scaled proportionally in a task-free manner, and the output was characterized by the population response of excitatory PFC units. In baseline conditions, increasing relative input strength increased the output proportionally, whereas activating or suppressing individua interneuron cell type may potentially modulate the gain (or not). VIP and SOM cell types can modulate the gain (panels a,b), whereas PV cell type cannot (panel c). In all panels, error bar denotes s.e.m. (n=205 units).

Reviewer Comment 3.2 — Lack of Suitable Network Variants: The authors do not appear to have tested whether other, less elaborate, networks were capable of providing a similar fit to data. For instance, if RNNs without all three IN populations perform as well as the reported version, this would argue against the need for including what might be otherwise unnecessary biological detail in their model. This is

(unfortunately) a common side-effect of RNNs, in that they are capable of fitting to a wide-array of functions, but at the detriment that it can be quite challenging to infer precisely how the networks were capable of said fitting.

Answer: Thanks for the insightful comment. In our experiment, we did try to use other (less elaborate) network assumptions to learn the same task, for instance, RNN without E/I constraints or RNN without three of interneuron cell-types. Our observations are that although these models can perform the task with the same performance accuracy, the results in behavioral and neural representations are less interpretable. It is important to point out that incorporating interneuron cell-types and specific thalamocortical pathways (MD1→PV and MD2→VIP) has made the computational dissection of the PFC-MD model accessible. Additionally, we could change the assumption of the model assumption and see the impact on the MD or PFC tuning. We now include a **new Supplementary Table 1** (see below) to highlight these findings.

	Network connectivity	MD unit tuning				Psychometric Curves	
		Rule	Context	Cue Uncertainty	Cue Sparsity	CAC	RDM
PFC-alone (baseline)		N/A					
Fig. 1a N_{Exc}=205 N_{PV}=17 N_{VIP}=17 N_{SOM}=17 N_{MD1}=N_{MD2}=100		✗ ✗	✓ ✓	✗ ✓	✗ ✓		
N_{Exc}=205 N_{PV}=21 N_{VIP}=15 N_{SOM}=15 N_{MD1}=N_{MD2}=100		✗ ✗	✓ ✓	✗ ✓	✗ ✓		
w/o Inh specificity		✗ ✗	✓ ✓	✓ ✓	✓ ✓		
w/o MD specificity		✗ ✗	✓ ✓	✓ ✓	✓ ✓		
no cell type N_{Exc}=205 N_{MD}=200 N_{Inh}=51		✗	✓	✓	✓		
Fig. 6a N_{Exc}=205 N_{PV}=17 N_{VIP}=17 N_{SOM}=17 N_{MD1}=120 N_{MD2}=80		✗ ✗	✓ ✓	✓ ✓	✓ ✓		
N_{Exc}=205 N_{PV}=21 N_{VIP}=15 N_{SOM}=15 N_{MD1}=120 N_{MD2}=80		✗ ✗	✓ ✓	✗ ✓	✗ ✓		

Our observations are that the addition of the feedforward MD structure contributes to the key difference in model robustness and performance (such as AUC), whereas MD cell-type specificity and MD population size contribute to unit tuning and context-switch speed. While the list in **Supplementary Table 1** is by no means complete, we are open to further model testing or incorporation of more biological details into the model architecture.

Reviewer Comment 3.3 — Model Generalizability: While the computational model is biologically constrained and successfully replicates key findings in mice, its generalizability to other species, particularly humans, is unclear. The reliance on mouse-specific thalamocortical connectivity may limit the broader applicability of the findings.

Answer: Thanks for your insightful comments. We agree completely with the reviewer that our model result replication and prediction have focused on task-performing mice experiments. Yet its generalizability to humans or other species (such as tree shrews or monkeys) remains untested. This represents an important direction of our current research in computational modeling effort. Recent unpublished data from the Halassa lab have shown that the MD thalamus encodes the task context in a human hierarchical decision-making task (Manuscript Preprint will become online soon). It is highly likely that the findings of behavioral and neural data will differ between species and the model cannot fully translate to other species. We now have discussed this limitation in Discussion. While not being necessarily sufficient to convince the reviewer, we would still like to mention that our preliminary data in modeling human fMRI experiments on a similar context-switching task has shown qualitative similar results in behavioral data (e.g., switching latency) and MD representation (e.g., encoding "reward surprise signal"). These results have been summarized in an abstract recently submitted to COSYNE'25. However, because of the complexity of these new tasks, the new data along with new modeling effort will be reported in a follow-up investigation.

Reviewer Comment 3.4 — Experimental Validation: Although the model makes experimentally testable predictions, the study does not provide direct experimental validation of these predictions. The absence of empirical data to support the model's hypotheses weakens the overall impact of the findings.

Answer: Thanks for your insightful comments. First, our key results (Fig. 2 and Fig. 7) have replicated the key experimental neurophysiological and behavioral findings in mouse experiments. The results of context-switch experiments have been validated by empirical behavioral data from wide-type mice and tree shrews from the Halassa lab (Mukherjee et al. Nature 2021; Lam et al., Nature, in press). We have cited these studies in the discussion. Second, in Fig. 9, our model has made experimental testable predictions, and the results were conceptually similar to (or validated by) the behavioral results in 22q11 mouse experiments. To our best knowledge, this is the first effort. We believe that these findings have served as the valuable first step to validate the model. Due to space limitation (9 main figures and 10 supplementary figures, more than 160 panels in total), a systematic investigation of 22q11 mouse data compared to the model is beyond the scope of current manuscript and will be the subject of future investigation. Finally, we hope that our continuing modeling effort will further strength this line of reasoning for thalamocortical circuits across cognitive task conditions and multiple species.

Reviewer Comment 3.5 — Simplification of Neural Dynamics: The model may oversimplify the complex dynamics of the prefrontal cortex and mediodorsal thalamus interactions by focusing on specific connectivity patterns and cell types. This reductionist approach might overlook other critical factors influencing cognitive flexibility and decision-making.

Answer: Thanks for your insightful comments. We believe in the modeling philosophy "all models are wrong, some are useful". To our best knowledge, any computational models, regardless of the modeling level (e.g., cellular, neuronal, or systems level), are all subject to abstraction and simplifications; but this does not prevent us to use the model to explore the key computational principles or mechanisms of the circuit under study. The reductionist approach inevitably ignores some biological details (such as the cortical layer

structure; thalamic inhibition; distinct somatic and dendrite targets from the interneurons). Additionally, the PFC-MD model does not explicitly model other top-down or modulatory input from other areas of the brain. We now discuss these limitations at the end of Discussion.

By taking a reductionist approach in our modeling, we also agree with reviewer that there are other critical factors influencing cognitive flexibility, such as thalamic inhibition, cortico-cortical communications, thalamocortical LFP coherence, and top-down modulation (e.g., influenced by dopamine). While it is nearly impossible to explore all these factors in a stand-alone modeling paper, we have briefly discussed these issues at the end of Discussion. Finally, we hope our continuing modeling effort will further enrich the model complexity and capture more complex PFC-MD interactions in the context of cognitive flexibility and decision-making.

Reviewer Comment 3.6 — Contextual Limitations: The study primarily focuses on context-dependent decision-making and cue uncertainty. While these are important aspects of cognitive control, the work does not address how the MD thalamus might interact with the PFC in other cognitive domains, potentially limiting the scope of its conclusions.

Answer: Thanks for your insightful comments. While we agree that cognitive control covers a wide range of domains (such as cognitive flexibility, attention, memory, and emotional processing), we also respectfully disagree with the reviewer's statement. In our opinion, it is nearly impossible to cover all aspects of cognitive control in a single study. Meanwhile, we are actively working on value or reward-based decisions and test whether a modified PFC-MD model is broadly in agreement with experimental findings, but that is a study on its own and will warrant its own publication in the future.

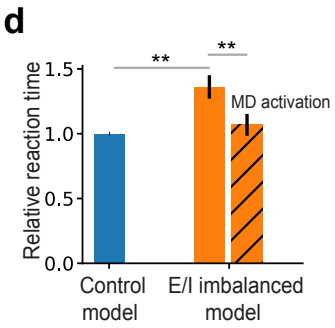
In our current experiment, we have mainly focused on the 4AFC task (or its variants) used in the mouse experiment (Schmitt et al. 2017 *Nature*), which has covered few cognitive domains (context switching, task uncertainty, and working memory). However, it does not limit us to this specific experimental setting. Recently, we have adopted a similar version of PFC-MD model to learn several versions of human cognitive tasks, such as the Wisconsin Card Sorting Test (WCST) and a human decision-making task with outcome uncertainty (task described in Wang et al. *Nat. Commun*. 14, 3552, 2023). Interestingly, we have found several model predictions conceptually like human experiments, at both behavioral and neural levels (such as MD activation in context-switching task-performing human fMRI experiments; unpublished data from the Halassa lab). Part of these new findings has been recently submitted to COSYNE'25. However, while these results are interesting to report on its own, it is beyond the scope of the current manuscript and the detailed investigation will be presented in a follow-up study.

Finally, we hope our continuing modeling effort will further support our conclusion and expand the scope of the investigation of PFC-MD interactions.

Reviewer Comment 3.7 — Lack of Behavioral Correlation: The study primarily relies on computational modeling and neurophysiological findings, with limited emphasis on how these relate to observable behavioral outcomes. A stronger connection between the model's predictions and real-world behavioral data would enhance the study's relevance and impact.

Answer: Thanks for your insightful comments. We agree with the reviewer that real-world behavioral characterizations are rich and multi-dimensional (e.g., performance accuracy, reaction time, switching time latency, etc). In our experiment, we have mainly focused on the **performance accuracy** as a form of psychometric curve as the model behavioral readout. Additionally, we presented results (Fig. 7b, Fig. 8b, 9b,c) regarding the **cue-to-rule remapping time** during context switching.

To partially address the reviewer's concern, we have done additional computer simulations to compare the reaction time between a control model and E/I imbalanced model (see a new panel in Fig. 9d shown below), this model behavioral prediction was conceptually like some findings in the literature (Ouhaz et al, 2022; Lam et al., 2022). However, to our best knowledge, the reaction time data from rodent or human experiments are still lacking to allow us to make direct comparisons. We hope our continuing modeling effort will further strengthen this weakness and provide more validation on real-world behavioral data.



Response Fig. 9 (Added a new panel as Fig. 9d in the main text).
d Reduced cortical inhibition in the trained PFC-MD model caused slower relative reaction time (**, p<0.01; signed-rank test). MD activation improved reaction time ((**, p<0.01). Error bar denotes SD (n=10).

Additionally, we also add a few sentences in the main text.
"Furthermore, in light of the observation that E/I imbalance deteriorates the reaction time in decision making[42], our computer simulations also supported that MD activation rescued the reaction-time deficit in decision-making for the E/I-imbalanced model (**Fig. 9d**)." ***(Lines 750-752)***

Ouhaz Z, Perry BAL, Nakamura K, Mitchell AS. Mediodorsal Thalamus Is Critical for Updating during Extradimensional Shifts But Not Reversals in the Attentional Set-Shifting Task. eNeuro. 2022 9(2):ENEURO.0162-21.2022.

Lam NH, Borduqui T, Hallak J, Roque A, Anticevic A, Krystal JH, Wang XJ, Murray JD. Effects of Altered Excitation-Inhibition Balance on Decision Making in a Cortical Circuit Model. J Neurosci. 2022 Feb 9;42(6):1035-1053.

Reviewer #4 (Remarks to the Author):

In this manuscript, the authors generate a PFC-MD model that incorporates known macro- and micro-circuits. They found that their model recapitulated known behavioral results and neural tuning during behavior as well as generated experimentally testable predictions. Overall, the computation modeling and data are well developed and detailed, the topic of MD-PFC interactions in support of cognition is important and of wide interest, and the overall depth of analysis is tremendous. Missing from much of the first draft of the manuscript is satisfactory statistical analysis to fully justify many of the claims that are made – having this would greatly strengthen the manuscript.

We acknowledge that all statistical analyses need to be strengthened, and we ensure that the details below are clearly explained in the revised manuscript.

COMMENTS:

Reviewer Comment 4.1 — Figure 1: What percentage of the PFC network are the different neuron types (Exc, PV/SOM/VIP)?

Answer: Thanks for your insightful comments. According to Dale's principle, the ratio of Exc and Inh (including PV, SOM, VIP) is 4:1. In our computer simulations, the total number of PFC units is 256, N_{PFC}=256, so we set N_{exc}=205, and N_{inh}=51 (N_{PV}= N_{SOM} =N_{VIP}=17). We now provide those missing details in the Methods section.

Additionally, we also varied the network size and architectures and summarize the results in new **Supplementary Table 1**.

Reviewer Comment 4.2 — line 128: "higher/lower density/sparsity" is confusing; recommend separating ("higher density/lower sparsity…"

Answer: Thanks for your insightful comments and apologize for the confusion. We now rephrase the sentence as follows

"At the signal level, uncertainty is characterized by the cue density, or alternatively the cue sparsity (a higher density or lower sparsity level implies a greater SNR)." ***(Lines 129-130)***

Reviewer Comment 4.3 — "congruence" is not well defined

Answer: Thanks for your insightful comments. The term "congruence" refers to the "degree of agreement" among the HP and LP cues. A higher degree of congruence implies a lower level of HP/LP cue uncertainty.

In the Methods section, we defined the "congruence".

"In the CAC task, the LP/HP congruence level (range: [0.5,1.0]) was defined by a ratio: $\max\left\{\frac{\#\text{LP pulses}}{\#\text{LP pulses}+\#\text{HP pulses}}, \frac{\#\text{HP pulses}}{\#\text{LP pulses}+\#\text{HP pulses}}\right\}$, with 0.5 being maximally uncertain and 1 being maximally certain." ***(Lines 1124-1126)***

Reviewer Comment 4.4 — line 145 – there is a switch from discussing cue uncertainty to "cue conflict", which seems like the same thing as "cue uncertaintly", but then why the language change? Furthermore, isn't lines 145-148 a recapitulation of the description in lines 125-128 (which is probably the more thorough of the two)?

Answer: Thanks for your insightful comments. We now replace "cue conflict" with "cue uncertainty" to avoid the confusion. We also remove the redundant sentences.

Reviewer Comment 4.5 — Figures 1e and 1f and associated text: there is no statistical basis for the statement that rule errors increase with cue uncertainty; this is needed; simplest would be to do a two way ANOVA (factors: error type and level of uncertainty) followed by post-hoc testing

Answer: Thanks for your insightful comments. We now provide corresponding statistical tests to support the statement. Specifically, we use both nonparametric Kruskal-Wallis test to test the median is different

(for three uncertainty levels) and Bonferroni-corrected rank-sum test for different pairs. Our rationale is the following: data distribution is non-normal because of the small sample size; correct / error type and uncertainty should be independently reported as there is no interaction between them).

In Fig. 1e, all comparisons are statistically significant and shown as follows.

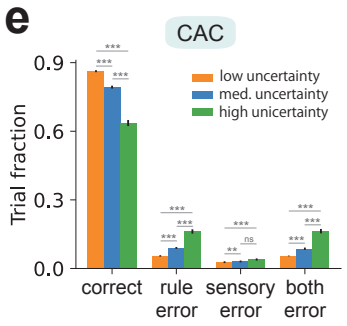
Correct group: Kruskal-Wallis test, p=0.0000; Bonferroni-corrected rank-sum test: low vs. med, p = 0.0002; low vs. high, p= 0.0002, med vs. high, p=0.0002

Rule error group: Kruskal-Wallis test, p=0.0000; Bonferroni-corrected rank-sum test: low vs. med, p = 0.0002; low vs. high, p= 0.0002, med vs. high, p=0.0002

Sensory error group: Kruskal-Wallis test, p=0.0003; Bonferroni-corrected rank-sum test: low vs. med, p = 0.0032; low vs. high, p= 0.0003, med vs. high, p=0.1736 (n.s.)

Both error group: Kruskal-Wallis test, p=0.0000; Bonferroni-corrected rank-sum test: low vs. med, p = 0.0002; low vs. high, p= 0.0002, med vs. high, p=0.0002

The revised Fig. 1e is shown below.



Response Fig. 10 (Update Fig 1e in the main text).

e Percentage of task error types in the CAC context under various cue uncertainty conditions. For each correct/error group, all pairwise p-values were computed based on Bonferroni-corrected rank-sum test. **p<0.01, ***p<0.001.

In Fig. 1f, all comparisons are statistically significant and shown as follows.

Correct group: Kruskal-Wallis test, p=0.0000; Bonferroni-corrected rank-sum test: low vs. med, p = 0.0008; low vs. high, p= 0.0008, med vs. high, p=0.0008

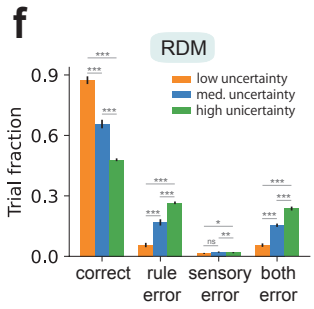
Rule error group: Kruskal-Wallis test, p=0.0000; Bonferroni-corrected rank-sum test: low vs. med, p = 0.0008; low vs. high, p= 0.0008, med vs. high, p=0.0008

Sensory error group: Kruskal-Wallis test, p=0.0003; Bonferroni-corrected rank-sum test: low vs. med, p = 0.1152 (n.s); low vs. high, p= 0.0157, med vs. high, p=0.0033

Both error group: Kruskal-Wallis test, p=0.0000; Bonferroni-corrected rank-sum test: low vs. med, p = 0.0008; low vs. high, p= 0.0008, med vs. high, p=0.0008

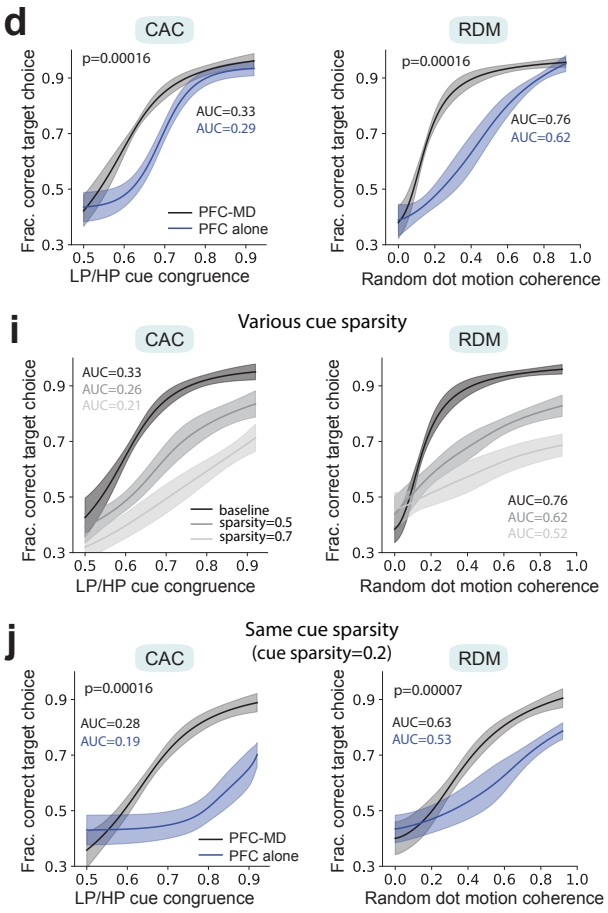
We now provide more testing details and report p-values (via *p<0.05,**p<0.01,***p<0.001) in the figure legend.

The revised Fig. 1f is shown below.



Response Fig. 11 (Update Fig 1f in the main text).
f Percentage of task error types in the RDM context under various cue uncertainty conditions. For each correct/error group, all pairwise p-values were computed based on Bonferroni-corrected rank-sum test. **p<0.01, ***p<0.001.

Reviewer Comment 4.5 — Same issue in Figures 1d, 1i, and 1j and associated text; need statistics here to justify the statement that the PFC-MD model outperformed the PFC-alone model
Answer: Thanks for your insightful comments. We now provide additional statistics to justify the statement. To compare two models, we compared the area under the psychometric curves (AUC). We report the p-value based on the non-parametric rank-sum test. The revised panels in Fig. 1d, 1i, 1j are shown below.



Response Fig. 12 (Update Fig 1d in the main text).
d Psychometric curves of RDM and CAC contexts. Shaded areas denote SD, and the lines denote the mean derived from the parameter fit of a logistic function. Each condition was repeated 10 times with different input realizations yet with identical summary statistics. The PFC-MD model (black) outperformed the PFC-alone model (blue) in the presence of intermediate-to-high cue uncertainty, achieving a greater area under the psychometric curve (AUC). p=0.00016, rank-sum test. **i** Psychometric curve of the PFC-MD model under three different cue sparsity levels in the CAC and RDM contexts. The AUC statistics decreased with an increasing sparsity level (CAC: black vs dark grey, p=0.00016; black vs light grey, p=0.00016, dark grey vs light grey, p=0.0008; RDM: black vs dark grey, p=0.00016; black vs light grey, p=0.00016, dark grey vs light grey, p=0.0008, rank-sum test). **j** Psychometric curve comparison between the PFC-MD and PFC-alone models in the CAC and RDM contexts in a low cue sparsity condition. The PFC-MD outperformed the PFC-alone model in AUC (CAC: p=0.00016; RDM: p=0.00007, rank-sum test).

Additionally, in Methods section, we have added details

"Based on the percentage of correct choices, we further fit a smooth psychometric curve function using a logistic function (**Fig. 1d**). The logistic function was characterized by four parameters $\{A, B, C, D\}$:

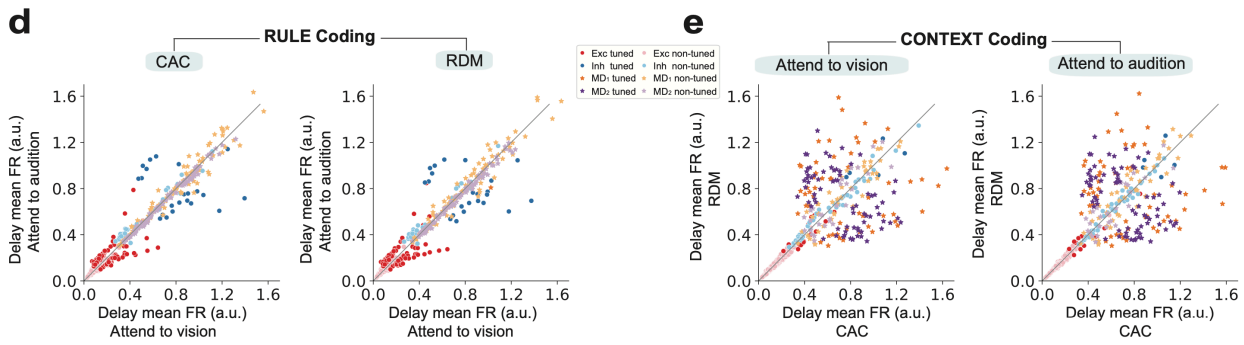
$$f(x) = \frac{A - D}{1 + \left(\frac{x}{C}\right)^{B}} + D$$

Note that in all psychometric curves, only the rightmost cue condition in the x-axis (i.e., lowest cue uncertainty) was used for training the models, and the remaining cue conditions were never used in training and only assessed in testing. To quantify the performance, we computed the area under the curve (AUC). A larger AUC value indicates a good model performance. To compare the AUC between the PFC-MD and PFC-alone models, we constructed the ensemble mean±SD of two psychometric curves and derived the corresponding AUC samples; from which we calculated the p-value based on the Mann-Whitney rank-sum test." ***(Lines 1131-1141)***

Reviewer Comment 4.7 — *Figure* 2d and 2e show the ratios nicely but are not really connected to percentages discussed in the corresponding text. It is only in the methods that it is mentioned that this is via rank sum test. More information would be helpful here: perhaps dark shade for significantly tuned neurons and light shade for non-tuned?

Answer: Thanks for your insightful comments. We now provide more criterion information regarding the percentage of rule/context-tuned neurons discussed in the text. More specifically, if a unit is said to be tuned to a task variable (rule or context), it must meet the following empirical criterion: $|\text{mean FR}_{\text{task condition1}} - \text{mean FR}_{\text{task condition2}}| > 0.25 \times \text{mean\_overall FR}$

Following the reviewer suggestion, we now revise Fig. 2d,2e to change the color of dots with dark/light shade (see below) to reflect significantly tuned/non-tuned units. We also change the text accordingly in the figure legend and Methods.



Response Fig. 13 (Update Fig 2d,2e in the main text).
d Population statistics of mean firing rates (during the task delay period) of excitatory PFC units and MD units for encoding rule. Markers in dark/light color shade represent tuned/non-tuned units, respectively. **e** Population statistics of mean firing rate (during the task delay period) of excitatory PFC units and MD units for encoding context. Markers in dark/light color shade represent tuned/non-tuned units, respectively.

Reviewer Comment 4.8 — Figures 2f, 2g, 2h and associated text – how was it determined that 25-30% of MD1 unites modulated their firing rates with respect to cue uncertainty? Pairwise test as described in methods would not be appropriate here. In some case the differences look like they wouldn't be significant.
-Figure 2i and associated text – it looks like almost every MD neuron codes for both rules, with the timing of the firing rate increase corresponding to different rules. Does this match real data?
Answer: Thanks for your insightful comments. We now change the cue uncertainty modulation criterion in the Methods and use a regression-based method to test the uncertainty tuning.

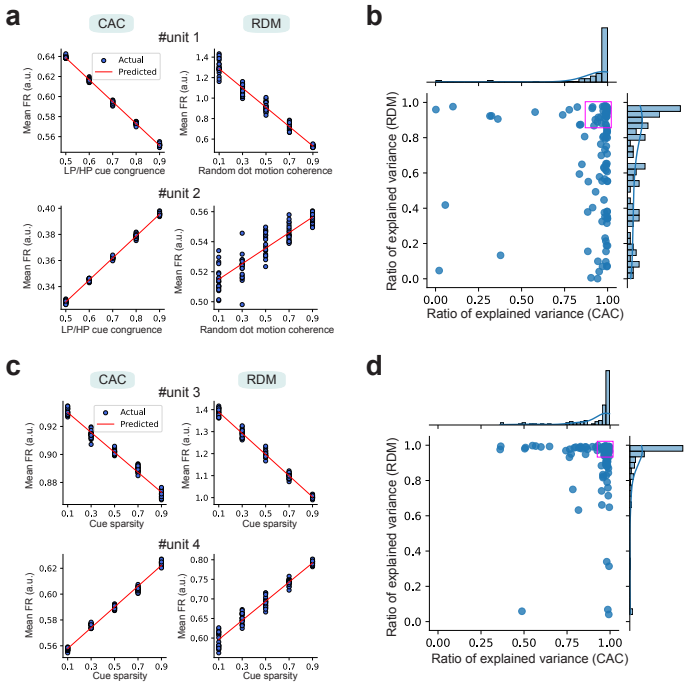
We assumed a linear model between the mean MD unit firing rate (during cueing period) and the cueing uncertainty (e.g., coherence or congruence)

$$\text{Mean FR} = \alpha + \beta \cdot \text{cue uncertainty level}$$

where α and β denote the regression coefficients. We fit the model prediction between mean FR and cue uncertainty level, and further characterize the percentage of explained variance (EV) for two different contexts (e.g., CAC and RDM contexts). The MD unit is strongly modulated with cueing uncertainty if it simultaneously meets the strong modulation criterion (e.g., %EV > 80%) for both contexts. ***(Lines 1159-1166)***

Using a criterion of 80-85% percentage threshold will yield ~30-40% of MD units that modulate their firing rates with respect to cue uncertainty.

We now update our respective Methods section and use a **new Supplementary Figure 3** (below) to illustrate our point.



Response Fig. 14 (Added a new panel as Supplementary Fig. 3 in the main text). **MD$_1$ units modulate their firing rates with respect to cue uncertainty and cue sparsity. a** The line fitting results of cue uncertainty from two MD$_1$ units. **b** The explained variance from both CAC and RDM contexts. **c** The fitting result of cue sparsity from two MD$_1$ units. **d** The explained variance from both CAC and RDM contexts.

In panels **a**,**c**, each row represents the fitting result from one MD unit. In panels **b**,**d** we summarize the % explained variance from both CAC and RDM contexts.

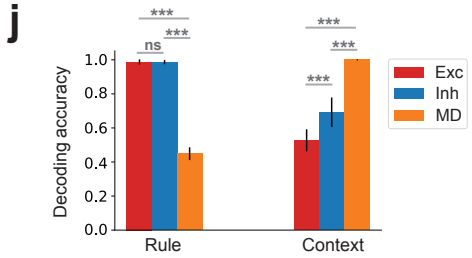
We like to clarify that in Fig. 2i, what we showed in the heatmap are the normalized peak FR of PFC excitatory units (*not* MD units). This also matched well with the experimental data of task-performing mice (Fig. 1d,f in Rikhye et al. 2018, Nat. Neurosci.).

Reviewer Comment 4.9 — Figure 2j – again need statistical justification for the statements about differences; 2 way ANOVA would be straightforward. The (likely) increased context coding in MD is also not mentioned in the main text.

Answer: Thanks for your insightful comments. We did not use 2-way ANOVA because here the rule/context decoding analyses were independently conducted. The decoding error couldn't be decomposed into two factors (rule/context and cell types). Instead, we now provide statistics derived from both nonparametric Kruskal-Wallis test and Bonferroni-corrected rank-sum test.

```
========================Kruskal-Wallis_Bonferroni_corrected_rank_sum_test========== decode: rule
Kruskal-Wallis test: Stat = 42.7297, p-value = 0.0000
Performing ranksums tests with Bonferroni correction:
Bonferroni-corrected alpha: 0.0167
Group 1 vs Group 2: Stat = 0.5410, p-value = 0.5885 --> Not significant
Group 1 vs Group 3: Stat = 5.4100, p-value = 0.0000 --> Significant
Group 2 vs Group 3: Stat = 5.4100, p-value = 0.0000 --> Significant
========================Kruskal-Wallis_Bonferroni_corrected_rank_sum_test========== decode: context
Kruskal-Wallis test: Stat = 53.2303, p-value = 0.0000
Performing ranksums tests with Bonferroni correction:
Bonferroni-corrected alpha: 0.0167
Group 1 vs Group 2: Stat = -5.1395, p-value = 0.0000 --> Significant
Group 1 vs Group 3: Stat = -5.4100, p-value = 0.0000 --> Significant
Group 2 vs Group 3: Stat = -5.4100, p-value = 0.0000 --> Significant
```

We now present the revised Fig. 2j below.



Response Fig. 15 (Update Fig 2j in the main text).
j Population decoding analysis showed that the PFC population better encoded rule, whereas the MD population better encoded context. Error bar denotes SD (n=20). ****P*<0.001, Bonferroni-corrected rank-sum test.

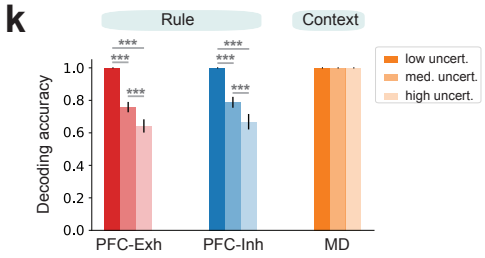
We also mention that MD population is much better in context coding than two PFC subpopulations.

"The rule information could be reliably decoded from the PFC, but not from the MD; whereas the context information could be inferred perfectly from the MD, but less reliably from the PFC" ***(Lines 274-275)***

Reviewer Comment 4.10 — Figure 2k - Not sure why this figure is in black and white. Also, the concept of decoding accuracy is difficult to wrap your head around here. Are you decoding the correct rule, or the rule they actually implement? One could speculate that under uncertainty the rule matches what is correct less often, but that the rule representation in the PFC corresponds to their actual choice (right or wrong) nonetheless.
Answer: Thanks for your insightful comments and apologize for the confusion. We now provide the revised Figure 2k in a color form (see below). To clarify, we were decoding the correct target rule instead of the rule being implemented. In the special case when accuracy is 100%, they will be the same.
We also provide statistics derived from nonparametric Kruskal-Wallis test and Bonferroni-corrected rank-sum test separately for the "PFC-Exh" and "PFC-Inh" groups (see below).

```
========================Kruskal-Wallis_Bonferroni_corrected_rank_sum_test========== cell: exc
Kruskal-Wallis test: Stat = 51.5045, p-value = 0.0000
Performing ranksums tests with Bonferroni correction:
Bonferroni-corrected alpha: 0.0167
Group 1 vs Group 2: Stat = 5.4100, p-value = 0.0000 --> Significant
Group 1 vs Group 3: Stat = 5.4100, p-value = 0.0000 --> Significant
Group 2 vs Group 3: Stat = 5.1260, p-value = 0.0000 --> Significant
========================Kruskal-Wallis_Bonferroni_corrected_rank_sum_test========== cell: inh
Kruskal-Wallis test: Stat = 53.3680, p-value = 0.0000
Performing ranksums tests with Bonferroni correction:
Bonferroni-corrected alpha: 0.0167
Group 1 vs Group 2: Stat = 5.4100, p-value = 0.0000 --> Significant
Group 1 vs Group 3: Stat = 5.4100, p-value = 0.0000 --> Significant
Group 2 vs Group 3: Stat = 5.4100, p-value = 0.0000 --> Significant
```

The revised Fig. 2k is shown below with revised figure legend.



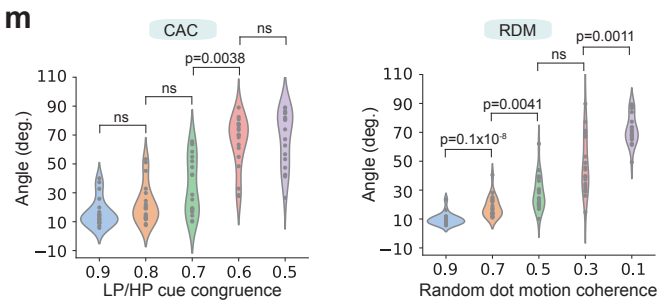
Response Fig. 16 (Update Fig 2k in the main text).
k Increased in cue uncertainty caused decreased rule decoding accuracy for the PFC, but did not affect context decoding accuracy for the MD. Error bar denotes SD (n=20). In rule decoding, statistical tests were conducted independently for each group; all paired comparisons were statistically significant. ***p<0.001, Bonferroni-corrected rank-sum test.

Reviewer Comment 4.11 — Figure 3m – again need to report statistics to support statement that the principal angle shifts in relation to uncertainty

Answer: Thanks for your insightful comments. We now provide revised Fig. 3m and corresponding statistics and p-values based on nonparametric Bonferroni-corrected rank-sum tests to support the statement (see below).

```
============================================================ CAC_task
Kruskal-Wallis test: Stat = 46.3884, p-value = 0.0000
Performing ranksums tests with Bonferroni correction:
Bonferroni-corrected alpha: 0.0050
Group 1 vs Group 2: Stat = -1.7853, p-value = 0.0742 --> Not significant
Group 1 vs Group 3: Stat = -3.4895, p-value = 0.0005 --> Significant
Group 1 vs Group 4: Stat = -4.4903, p-value = 0.0000 --> Significant
Group 1 vs Group 5: Stat = -4.7067, p-value = 0.0000 --> Significant
Group 2 vs Group 3: Stat = -2.0288, p-value = 0.0425 --> Not significant
Group 2 vs Group 4: Stat = -3.9764, p-value = 0.0001 --> Significant
Group 2 vs Group 5: Stat = -4.4092, p-value = 0.0000 --> Significant
Group 3 vs Group 4: Stat = -2.8944, p-value = 0.0038 --> Significant
Group 3 vs Group 5: Stat = -3.3001, p-value = 0.0010 --> Significant
Group 4 vs Group 5: Stat = -0.9468, p-value = 0.3438 --> Not significant
============================================================ RDM_task
Kruskal-Wallis test: Stat = 71.5779, p-value = 0.0000
Performing ranksums tests with Bonferroni correction:
Bonferroni-corrected alpha: 0.0050
Group 1 vs Group 2: Stat = -4.4903, p-value = 0.0000 --> Significant
Group 1 vs Group 3: Stat = -5.0584, p-value = 0.0000 --> Significant
Group 1 vs Group 4: Stat = -5.2207, p-value = 0.0000 --> Significant
Group 1 vs Group 5: Stat = -5.4100, p-value = 0.0000 --> Significant
Group 2 vs Group 3: Stat = -2.8673, p-value = 0.0041 --> Significant
Group 2 vs Group 4: Stat = -4.3551, p-value = 0.0000 --> Significant
Group 2 vs Group 5: Stat = -5.3018, p-value = 0.0000 --> Significant
Group 3 vs Group 4: Stat = -2.4075, p-value = 0.0161 --> Not significant
Group 3 vs Group 5: Stat = -4.2469, p-value = 0.0000 --> Significant
Group 4 vs Group 5: Stat = -3.2731, p-value = 0.0011 --> Significant
```

The revised Fig. 3m is shown below with revised figure legend



Response Fig. 17 (Update Fig 3m in the main text).
m Quantification of the principal angle between MD and PFC subspaces with respect to various levels of cue uncertainty. There was an increasing trend in principal angle with increasing cue uncertainty (n=20 for

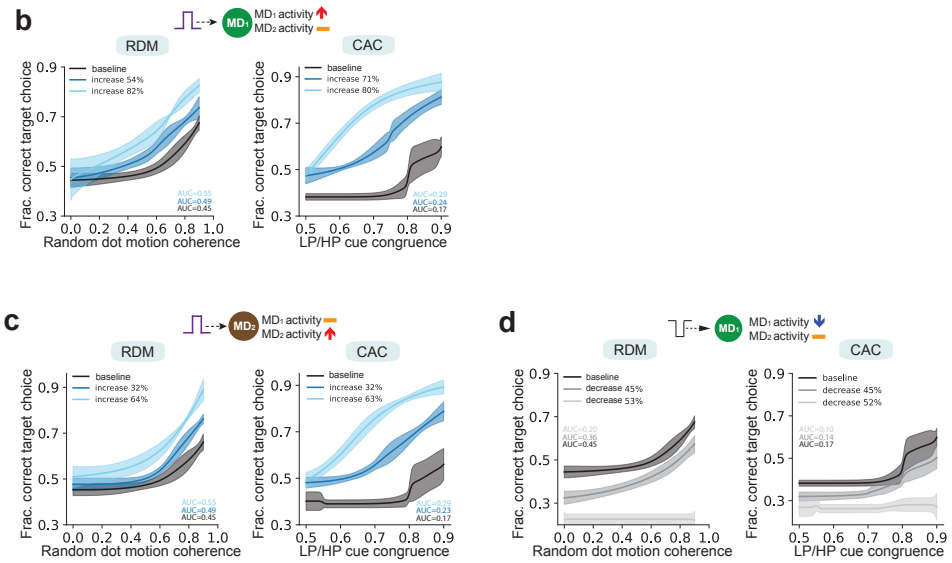
each condition). Comparisons between neighboring conditions were highlighted (Bonferroni-corrected rank-sum test).

Reviewer Comment 4.12 — Figure 4 and Supplementary Figure 4 – Panels are easy to follow here but all panels need statistical analysis to justify statements about differences. 4g lists three p-values apparently using what is typically a pairwise test, so this is confusing as well. The fact that the accuracy floors, even at the highest uncertainties, are often well above 25% seems potentially problematic. Shouldn't the floor be 25% and no lower or higher at max uncertainty (see 4b baseline, for example; even lower, 4d; down close to 20% 4e)? Furthermore, it is buried in the methods that chance accuracy is 25%, apparently because correct means correct rule AND choice. This is not obvious in the main body of the manuscript and a little bit strange because "correct" in a behavioral context (simulated or not) more typically refers to whether the choice was correct, period, irrespective of whether the organism was implementing the optimal strategy. If you're going to stick with this convention, then it should be articulated clearly in the main body of the manuscript, and the ylabels should be renamed so the meaning is clearer.

Answer: Thanks for your insightful comments. Our responses are as follows:

1) We now provide statistical tests and p-values between the area under the psychometric curves (AUC) to justify the statistical differences.

   Revised Fig. 4b, c, d are shown below

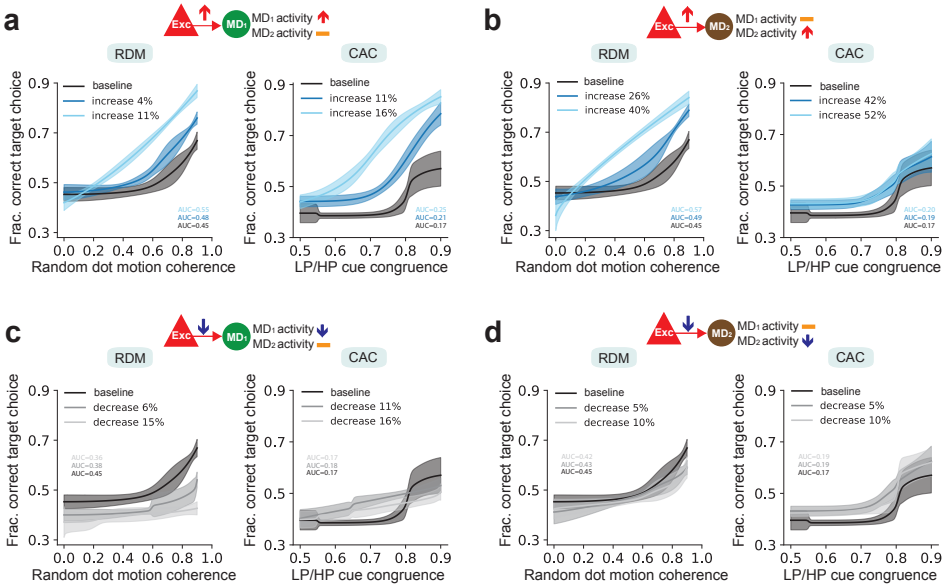


Response Fig. 18 (Update Fig 4b,c,d in the main text).

b, c Increasing the MD_1 or MD_2 population firing rate during an elongated delay period improved the working memory and the psychometric curves in two contexts. Black curve denotes the baseline, and the number of % denotes the relative increase in subpopulation firing rate. With MD_1 activation, the derived AUC increased from the baseline (RDM: dark blue vs. black, p=0.016; light blue vs. black, p=0.009; CAC: dark blue vs. black, p=0.0001; light blue vs. black, $p=0.7\times10^{-4}$). With MD_2 activation, the derived AUC increased from the baseline (RDM: dark blue vs. black, p=0.028; light blue vs. black, p=0.009; CAC: dark blue vs. black, p=0.0001; light blue vs. black, p=0.0002).

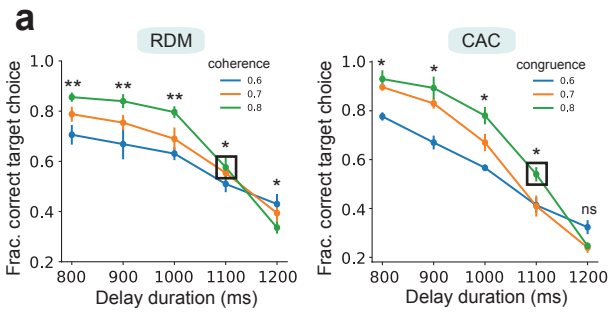
d Decreasing the MD_1 population firing rate degraded the working memory and the psychometric curves of two contexts. With MD_1 suppression, the derived AUC increased from the baseline (RDM: dark grey vs. black, p=0.014; light grey vs. black, p=0.009; CAC: dark grey vs. black, p=0.0003; light grey vs. black, $p=0.7x10^{-4}$).

Revised Supplementary Fig 4 is shown below

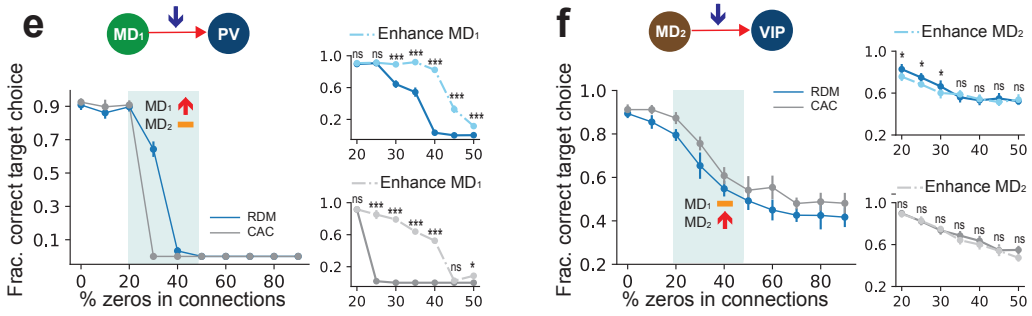


Response Fig. 19 (Update Supplementary Fig 5 in the main text). Changing directed prefrontal Exc→MD connection strengths of the PFC-MD model affect MD activity and task performance. **a** Strengthening PFC→MD_1 connection strengths (shown by a relative increase with respect to the baseline) increased the MD_1 firing rate while keeping the MD_2 firing rate unchanged, which further boosted working memory maintenance or improved task performance with elongated delay duration (RDM: dark blue vs. black, p=0.028; light blue vs. black, p=0.009; CAC: dark blue vs. black, p=0.0001; light blue vs. black, p=0.0001).
b Strengthening PFC→MD_2 connection strengths increased the MD_2 activity while keeping the MD_1 activity unchanged, which further boosted working memory maintenance during elongated delay (RDM: dark blue vs. black, p=0.016; light blue vs. black, p=0.009; CAC: dark blue vs. black, $p=0.7x10^{-4}$; light blue vs. black, $p=0.7x10^{-4}$).
c In contrast, weakening PFC→MD_1 connection strengths (shown by a relative decrease with respect to the baseline) decreased the respective MD activity, causing a degraded or inconsistent task performance (RDM: dark grey vs. black, p=0.009; light grey vs. black, p=0.009; CAC: dark grey vs. black, p=0.003; light grey vs. black, p=0.16).
d Weakening PFC→MD_2 connection strengths decreased the respective MD activity, causing a degraded or inconsistent task performance (RDM: dark grey vs. black, p=0.25; light grey vs. black, p=0.047; CAC: dark grey vs. black, $p=0.7x10^{-4}$; light grey vs. black, p=0.0002).

2) We also revise Fig 4a, 4e and 4f (see below) accordingly with more statistics.



Response Fig. 20 (Update Fig 4a in the main text).
a Increasing the duration of task delay period reduced the task performance under cue uncertainty in both contexts. Error bar denotes SD (n=10). Statistical tests were conducted between green and blue curves (pointwise) for each delay duration condition (*p<0.05, **p<0.01, rank-sum test). The task condition marked with "□" was used in the illustrations of the three remaining panels.



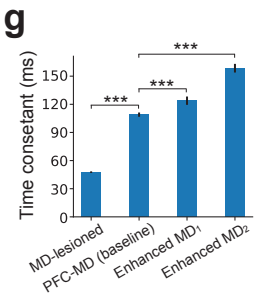
Response Fig. 21 (Update Fig 4e,f in the main text).
e *Left:* Weakening MD_1→PV connections (by increasing the percentage of zeros) during the cueing period quickly reduced the task performance in both RDM and CAC contexts. *Right:* increasing MD_1 activity (dotted line) could rescue each context's performance under a wide range of connectivity conditions (shaded area: 20%-50% of zeros). Error bar denotes SD (n=10). Statistical tests were conducted between dark blue and light blue curves (pointwise) for condition (*p<0.05, ***p<0.001, rank-sum test).
f Weakening MD_2→VIP connections only degraded the task performance slowly in both RDM and CAC contexts, but increasing MD_2 activity had no or little effect on the change in task performance. Statistical tests are similar to panel **e**.

3) In revised Fig. 4g, we only aimed to compare the pairwise comparison with the baseline model (i.e., PFC-MD). After Bonferroni-corrected rank-sum test, we can show the statistical significance (see screenshot below).

```
Bonferroni-corrected rank-sum test results:
pfc_md vs pfc: U-statistic = 400.00000, corrected p-value = 0.00000
  --> Significant difference after Bonferroni correction
pfc_md vs enhance md1: U-statistic = 0.00000, corrected p-value = 0.00000
  --> Significant difference after Bonferroni correction
pfc_md vs enhance md2: U-statistic = 0.00000, corrected p-value = 0.00000
  --> Significant difference after Bonferroni correction
pfc vs enhance md1: U-statistic = 0.00000, corrected p-value = 0.00000
  --> Significant difference after Bonferroni correction
pfc vs enhance md2: U-statistic = 0.00000, corrected p-value = 0.00000
  --> Significant difference after Bonferroni correction
enhance md1 vs enhance md2: U-statistic = 400.00000, corrected p-value = 0.00000
  --> Significant difference after Bonferroni correction

Process finished with exit code 0
```

We also revise Fig. 4g (see below)



Response Fig. 22 (Update Supplementary Fig 4g in the main text).
g Impact of network time constant under different thalamocortical connectivity manipulations. Error bar denotes SD (n=20 independently trained PFC-MD models). The MD_1 and MD_2 enhancement corresponded to the experiments where respective thalamocortical connections were strengthened during working memory. All statistical tests were rank-sum tests against the PFC-MD baseline (MD-lesioned vs. baseline: $p<0.0001$; enhanced MD_1 vs. baseline: $p<0.0001$; enhanced MD_2 vs. baseline: $p<0.0001$; Bonferroni-corrected rank-sum test).

4) In theory, the chance level for the 4AFC task was 25%. However, in practice, the trained model could perform better than chance or had a choice bias. For instance, if the model learned one rule correctly, and randomly guess the choice for the other rule, the performance would be greater than 50%. Even for the randomly initialized model, the accuracy level was sometimes greater than 25%.
5) Your point regarding the "correct" interpretation is well taken. We now relabel the ylabel from "**Fraction of correct**" to "**Fraction of correct target choice**". We also articulate it clearly in the main text to make it clear.

Reviewer Comment 4.13 — The 4e and 4f insets are also difficult to comprehend and very hard to see.
Answer: Thanks for your insightful comments. We now replot these insets as separate panels to enhance the readability. See **Response Fig. 21** above.

Additionally, we provide details in the main text and figure legend for explanation.

"Within a narrow range (20-50% sparsity, shaded area in the left panel of **Fig. 4e**), enhancement of phasic MD_1 activity could boost task performance (**Fig. 4e**, dashed lines of two smaller panels). In contrast, weakening MD_2→VIP connectivity degraded the task accuracy slowly (**Fig. 4f**), and enhancement of MD_2 activity had little effect on task accuracy. This was possibly because there was no direct VIP→Exc pathway to prefrontal excitatory neurons in our model assumption." ***(Lines 412-417)***

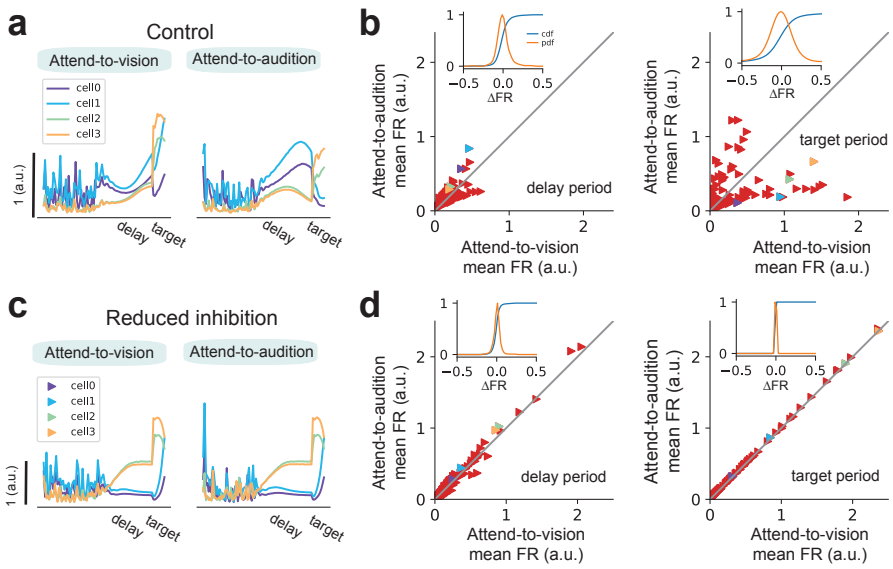
Reviewer Comment 4.14 — Line 281 – what are these p-values? Post-hoc tests? A series of pairwise comparisons? Is there a multiple comparisons correction?

Answer: Thanks for your insightful comments. We now provide p-values based on Bonferroni-corrected rank-sum test. We focus on the pairwise comparisons (compared to the PFC-MD baseline model). See **Response Fig. 22** above.

Reviewer Comment 4.15 — I don't understand how 5a and 5b show that "MD inhibition" alters PFC Exc rule tuning. And again, no statistical analysis is done to support the contention that "these operations changed tunings of some PFC excitatory units during the delay period and reduced choice at the population level during the target period."

Answer: Thanks for your insightful comments and apology for the confusion. To help clarify the differences in PFC excitatory unit rule tuning, we now add figure legend to mark the cell ID and add additional figure legend / labeling to highlight these changes (see figure below).

To support the statement, we compare the discrepancy of mean FR rate of PFC units in two rules ("attend-to-vision" - "attend-to-audition") between control and reduced inhibition conditions for both delay and target periods. We compared two cumulative distribution functions (cdf's) of $\Delta FR_{v\text{-}a}$=FR (attend-to-vision) – FR (attend-to-audition) and reported the p-values based on the two-sample Kolmogorov-Smirnov test.



Response Fig. 23 (Update Fig 5a-d in the main text).

a Tuning curves of four selective PFC excitatory units (indicated by different colors) during the control condition. These four units showed rule tunings ("attend-to-vision" vs "attend-to-audition"). **b** At the population level, mean firing rate (FR) comparison of PFC excitatory units between two rules during delay (left panel) and target (right panel) periods. Four units in the left panels are labeled in the same color in (**a**). Inset: Curves of probability density function (pdf, orange) and cumulative distribution function (cdf, blue). **c** In the presence of prefrontal E/I imbalance (e.g., reduced inhibition), PFC excitatory units shown in (**a**) decreased their discriminability in rule tuning during both delay and target periods. **d** Similar to panel (**b**), except for the reduced inhibition condition. Comparing the cdfs of $\Delta FR_{v\text{-}a}$=FR (attend-to-vision) – FR

(attend-to-audition) between panels (**b**) and (**d**) showed statistically significant differences (delay period: $p=4x10^{-7}$; target period: $p=4.8x10^{-15}$, Kolmogorov-Smirnov test)

Additionally, we also add a few sentences in the main text.

"Specifically, in the reduced inhibition condition, the mean firing rates of rule-tuned PFC excitatory units became nearly identical (**Fig. 5d**), so that the distributions of ΔFR_{v-a}=FR (attend-to-vision) – FR (attend-to-audition) dramatically differed from the control condition (delay period: $p=4x10^{-7}$; target period: $p=4.8x10^{-15}$, Kolmogorov-Smirnov test)." ***(Lines 439-443)***

Reviewer Comment 4.16 — line 301 – it's a large leap from a model of PFC-MD dynamics to the neural bases of cognitive biases; probably best to remove and save for discussion.
Answer: Thanks for your suggestion. We now remove the sentence.

Reviewer Comment 4.17 — line 303 "observed a more robust task performance…" I think you mean to say a more robust effect on task performance
Answer: Thanks for your comment. We agree with the reviewer and now rephrase the wording as "a more robust effect on task performance".

Reviewer Comment 4.18 — in Figure 5d you now have performance that goes down to 10% (well below chance according to the methods) and only down to 50% (well above chance) depending on which simulated manipulation is being used. Both seem problematic for the reasons described above.
Answer: Thanks for your insightful comments. In theory, the chance level for the 4AFC task was 25%. However, in practice, the trained model could have a choice bias to obtain better-than-chance performance. In other words, not all components are equally "wrong". For instance, if the model learned one rule correctly, and randomly guess the choice for the other rule, the performance would be greater than 50%. Even for the randomly initialized model, the accuracy level was sometimes greater than 25%.

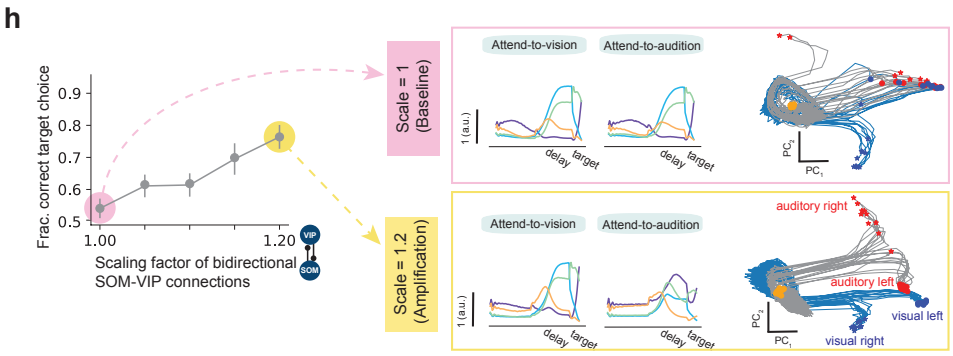
In Figure 5d, all reported results were based on random manipulations in the trained PFC-MD model. Therefore, the trained model could get worse-than-chance performance. For instance, if the model was manipulated such that both rules were mapped incorrectly, then the performance will be lower than 25%.

Reviewer Comment 4.19 — there is no legend in Figure 5 for panels 5a,b,f.
Answer: Thanks for your insightful comments. We now provide detailed legend of Fig. 5 as follows.
Fig. 5 | Parsing cognitive deficits and probing mechanistic causes in PFC-MD networks.
a Tuning curves of four selective PFC excitatory units (indicated by different colors) during the control condition. These four units showed rule tunings ("attend-to-vision" vs "attend-to-audition").
b At the population level, mean firing rate (FR) comparison of PFC excitatory units between two rules during delay (left panel) and target (right panel) periods. Four units in the left panels are labeled in the same color in (**a**). Inset: Curves of probability density function (pdf, orange) and cumulative distribution function (cdf, blue).
c In the presence of prefrontal E/I imbalance (e.g., reduced inhibition), PFC excitatory units shown in (**a**) decreased their discriminability in rule tuning during both delay and target periods.

d Similar to panel (**b**), except for the reduced inhibition condition. Comparing the cdfs of $\Delta FR_{v\text{-}a}$=FR (attend-to-vision) – FR (attend-to-audition) between panels (**b**) and (**d**) showed statistically significant differences (delay period: $p=4x10^{-7}$; target period: $p=4.8x10^{-15}$, Kolmogorov-Smirnov test)
e Two-dimensional neural trajectory representations in control and reduced inhibition conditions. Trajectories were generated from all correct and error trials. Orange dots: cue off and start of the delay period. Blue and red end points represent two rule representations, whereas dot and star symbols represent left and right choices, respectively.
f The task performance decreased with reduced prefrontal SOM→VIP and SOM→PV connectivity.
g Increasing mutual inhibition strengths between SOM and VIP neurons amplified the gain under a high cue sparsity. Activating the MD_2→VIP pathway further facilitated amplification and improved task performance. Two shaded circles indicate the two conditions illustrated in (**h**).
h In the case of RDM task of (**g**) (with cue sparsity 0.8), comparison of four selective PFC excitatory unit tunings and two-dimensional neural trajectories between scale=1 (baseline) and scale = 1.2 for bidirectional SOM-VIP connection strengths. In the latter case, single-unit rule tunings emerged, and population responses improved rule discriminability as shown in four clustered neural trajectories.

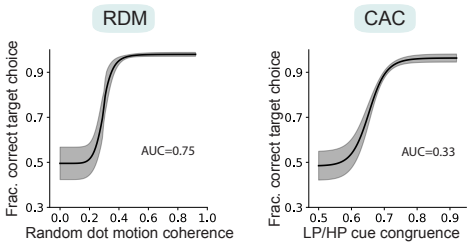
Reviewer Comment 4.20 — line 308-310 "A closer examination of single-unit and population responses revealed that bidirectional SOM-VIP amplification produced emergent rule tunings and enhanced rule discriminability (Fig 5f)". Really not sure what the basis is for this statement.
Answer: Thanks for your insightful comments and apology for the confusion. To help clarify this point, we now rearrange figure panels (see the revised Fig. 5f or the new Fig. 5h panel below). We also add a few sentences in the main text to explain our point.

"Notably, the amplification factor of 1.2 in bidirectional SOM-VIP connections produced visible changes in unit tuning and population representation compared to the no-amplification baseline (i.e., scaling factor of 1)." ***(Lines 456-458)***



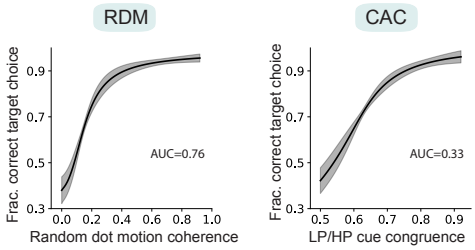
Reviewer Comment 4.21 — line 327 "Notably, this modified corticothalamic connectivity produced similar psychometric curves (Fig 6b)…" Need to state similar to what and ideally show and statistically examine the relationship.
Answer: Thanks for your insightful comments and we apologize for the confusion. To statistically examine the relationship, we report similar area under the psychometric curve (AUC) value between two conditions and report the rank-sum test for both RDM and CAC contexts.

Fig. 6b



versus

Fig. 1d



We now compare the two psychometric curve (AUC) value in the RDM context and CAC context respectively based on rank-sum tests (p=0.496 for RDM context; p=0.821 for CAC context). We now report the p-value in the main text and revise the text accordingly.

"Notably, this modified corticothalamic connectivity produced quantitatively similar AUC statistics in psychometric curve (**Fig. 6b** vs. **Fig. 1d**; p=0.496 for the RDM context; p=0.821 for the CAC context, rank-sum test)…" ***(Lines 500-502)***

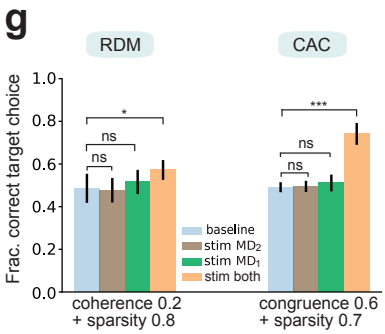
Reviewer Comment 4.22 — line 329 "Fig 6" which panel?
Answer: We apologize for the oversight. It is referred to **Fig 6d**.

Reviewer Comment 4.23 — Figures 6e, 6f, 6g; stats marked and described as Wilcoxon in figure caption; however, there are many comparisons here; I would suggest a two-way ANOVA followed by multiple comparisons post-hoc testing. Furthermore, in the caption, there are numbers following the p-values that are parentheses; no idea what those are.
Answer: We apologize for the oversight. These details are now added. The numbers following the p-values in parentheses are referred to RDM (CAC) contexts, respectively. We now rewrite the figure legend for Fig. 6e and Fig. 6f to clarify the meaning.

In Figures 6e,6f,6g, we conducted pairwise Wilcoxon rank-sum tests between the baseline model and MD manipulations. The reason is that cue uncertainty and cue sparsity were tested independently, and there are no two factors in each panel (as coherence and sparsity are tested independently; they were grouped together for visualization purpose only); the data distributions are non-normal because of small sample size; there is no multiple comparison either. We now rephrase the caption of Fig.6e and 6f as follows.

"e In the RDM context, MD_1 activation, but not MD_2 activation, improved task accuracy under cue uncertainty. MD_1 or MD_2 activation improved task performance under cue sparsity. Error bar denotes SD (n=10). Cue uncertainty comparison: Baseline vs. MD_2 activation (p=0.82); baseline vs. MD_1 activation (p=0.00016); MD_2 vs. MD_1 activation (p=0.00016). Cue sparsity comparison: Baseline vs. MD_2 activation (p=0.00016); baseline vs. MD_1 activation (p=0.00016); MD_2 vs. MD_1 activation (p=0.045), two-tailed Wilcoxon rank-sum tests.
f Similar to panel **e**, except in the CAC context. Baseline vs. MD_2 activation (p=0.16); baseline vs. MD_1 activation (p=0.00016); MD_2 vs. MD_1 activation (p=0.00043). Cue sparsity comparison: baseline vs. MD_2 activation (p=0.00016); baseline vs. MD_1 activation (p=0.00016); MD_2 vs. MD_1 activation (p=0.017), two-tailed Wilcoxon rank-sum tests." ***(Lines 525-534)***

We revise Fig. 6g and change its caption as follows



Response Fig. 24 (Update Fig 6g in the main text).
g When both cue uncertainty and cue sparsity were present, MD_1 or MD_2 activation alone couldn't improve task performance, but activation of both MD subpopulations could. Baseline vs. co-activation, *, p=0.011; ***, p= 0.00015, two-tailed Wilcoxon rank-sum tests. Other paired comparisons were n.s.

Reviewer Comment 4.24 — correct grammar in line 348
Answer: It is now corrected as follows
"a further examination of all connections between any PFC or MD units to postsynaptic PFC excitatory units revealed distinct four clusters according to their rule-tuning patterns (**Supplementary Fig. 7b**)." ***(Lines 548-550)***

Reviewer Comment 4.25 — line 344 – I recommend briefly describing the modified version of the CAC subtask and how it introduces the new features
Answer: Thanks for your insightful comments. The CAC subtask is the same as the description in Fig. 1c. What we meant is that the PFC-MD model was slightly modified because we did not need to model CAC and RDM contexts simultaneously. For this reason, we only changed the PFC-MD network input dimensionality accordingly. We now rephrase the sentence to clarify this fact.

"Specifically, we introduced cue-to-rule mapping uncertainty and sequential context switching (Context 1→Context 2→ Context 1') using the same CAC task and a slightly modified PFC-MD model with different network input dimensionality (N_{in}=6)." ***(Lines 545-546)***

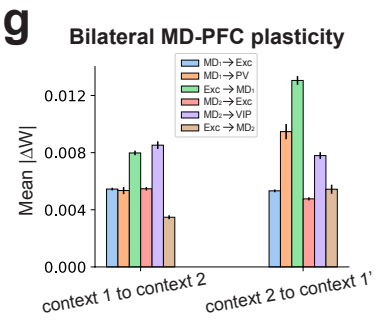
Reviewer Comment 4.26 — line 362 – correct grammar

Answer: Corrected

"Furthermore, the rule-tuning of PFC excitatory units tended to group with context-dependent E/I ratio and modified the E/I input accordingly while performing context switch successfully" ***(Lines 563-565)***

Reviewer Comment 4.27 — Figure 7g – please interpret, describe, and use stats

Answer: Thanks for your insightful comments. We now rearrange the panel (see revised Fig. 7g below) and provide error bar.



Response Fig. 25 (Update Fig. 7g in the main text). Quantification of mean synaptic plasticity (error bar: s.e.m.) in bilateral MD-PFC connections during two consecutive context-switching conditions. In general, MD_1→Exc, MD_2→Exc, and Exc→MD_2 connections showed the overall smallest changes, whereas Exc→MD_1 and MD_2→VIP connections showed the overall dominant changes.

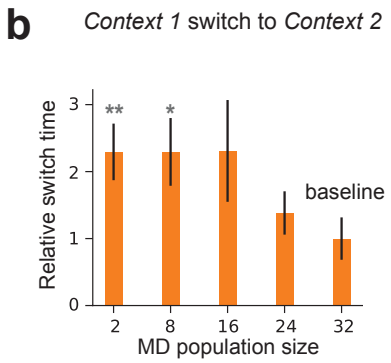
Additionally, we provide some explanation in the main text.

"In two context-switching conditions, we found that two dominant changes in MD-PFC plasticity were the Exc→MD_1 and MD_2→VIP pathways (Fig. 7g), suggesting distinct roles of corticothalamic and thalamocortical connectivity in cue-to-rule remapping, partly because of sparser Exc→MD_2 connectivity and unequal size between MD_1 and MD_2 subpopulations." ***(Lines 568-572)***

Reviewer Comment 4.28 — Figures 8b-d and associated text – need stats

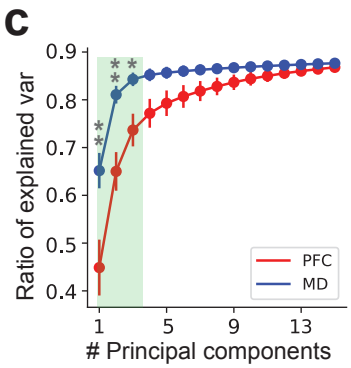
Answer: Thanks for your insightful comments. We now provide new stats for Figure 8b,8c,8d.

In Fig. 8b, we computed statistical difference between each MD population size with the baseline condition (#MD=32) based on non-parametric rank-sum tests.



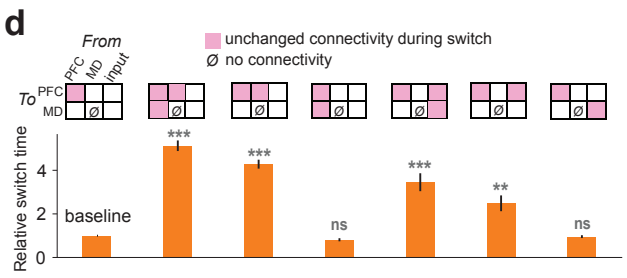
Response Fig. 26 (Update Fig. 8b in the main text). Comparison of relative context switch time with different sizes of MD population. All time was normalized with respect to the baseline (MD size of 32). Error bar denotes s.e.m. (n=10). All statistical tests were rank-sum tests against the baseline (N_{MD}=2 vs. baseline: p=0.0036; N_{MD}=8 vs. baseline: p<0.0052; Bonferroni-corrected rank-sum test).

In Fig. 8c, we computed the (point wise) statistical difference between PFC and MD regarding the ratio of explained variance based on #Principal components. We used the first three points to illustrate the differences based on non-parametric rank-sum tests.



Response Fig. 27 (Update Fig. 8c in the main text). Ratio of explained variance of PFC and MD populations derived from task-optimized PFC-MD models. Our computer simulations and principal component analysis (PCA) showed that the recurrent PFC structure had a larger dimensionality than the feedforward MD structure. Error bar denotes s.e.m. (n=10). Statistical tests were conducted between PFC and MD regarding the ratio of explained variance based on #Principal components. The first three points were used to illustrate the differences based on non-parametric rank-sum tests (p=0.003 for the first principal component; p=0.0015 for the second principal component; p=0.012 for the third principal component).

In Fig. 8d, we computed statistical differences between individual conditions with the baseline (first column) based on non-parametric rank-sum tests.



Response Fig. 28 (Update Fig. 8d in the main text). Comparison of relative context switch time with different assumptions of modifiable connectivity during switch. In the top, unilateral or bilateral PFC, MD and input connectivity patterns are shown (∅ denotes no connectivity). In total, there are five unidirectional connections that can be adapted. In the bottom, all time in the y-axis was normalized with respect to the baseline (unchanged intracortical connectivity). Error bar denotes s.e.m. (n=10). All statistical tests were rank-sum tests against baseline (**p<0.01, ***p<0.001).

Reviewer Comment 4.29 — line 413 – recommend restating in language that more specifically describes the effects of MD inhibition and activation

Answer: Thanks for your insightful comments. We now rephrase the sentence as follows.

"Furthermore, when the MD was lesioned, the neural trajectory of PFC population (Fig. 8h, middle panel, green dash curve) was markedly different from that observed during successful context switching (Fig. 8h, middle panel, grey solid curve). Specifically, in the MD lesion condition, the endpoints of trajectories that represent two rules were not separable at the first principal component (PC1) axis. However, with additional MD activation in the other pathway (Fig. 8h, right panel), the neural trajectory of PFC activity recovered to the normal successful condition, suggesting that MD activation may rescue the cognitive deficit whereas MD lesion or inhibition has an opposite effect" ***(Lines 707-714)***

Reviewer Comment 4.30 — line 510 "Our analyses of corticothalamic and thalamocortical projections suggest an asymmetric information flow and distinct contributions in cognitive control." Please follow this by reiterating what the asymmetry and differential contributions are

Answer: Thanks for your insightful comments. We now change the original sentence as follows.

"Our analyses of corticothalamic and thalamocortical projections suggest a continuous information flow and yet distinct contributions in cognitive control. Specifically, the MD→PFC projection can suppress competing cues or amplify sparse cueing signals to improve the SNR in a pathway-dependent manner (**Fig. 4**), whereas activating the PFC Exc→MD projection can achieve a similar effect (**Supplementary Fig. 4**). Together, enhancing the thalamocortical and corticothalamic loops can resolve the ambiguity in sensory cues and improve rule maintenance in working memory. During context switching, neural plasticity of thalamocortical connectivity seems to play a more crucial role in facilitating the switch than that of corticothalamic connectivity (**Fig. 8d**)." ***(Lines 847-855)***

Reviewer Comment 4.31 — "Our results are also consistent with the idea that the MD can facilitate GABA shift" à what does this mean?

Answer: Thanks for your insightful comments. We now rephrase the sentence as follows.

"Our results are also consistent with the idea that the MD projections target both prefrontal glutamatergic pyramidal neurons and GABAergic interneurons can regulate the changes in GABA activity and further change prefrontal E/I balance and functional connectivity." ***(Lines 880-882)***

Reviewer Comment 4.32 — It would be nice if the discussion better delineated the differences in the computational basis of how the MD supports working memory vs cognitive flexibility; which subtypes, which direction in the circuit etc

Answer: Thanks for your insightful comments. This has been partly explained in Fig. 8a regarding the MD's role in supporting working memory (during the delay period) and regulating cue signal-to-noise (SNR) ratio (during the cueing period). The MD's role in enabling cognitive flexibility such as in context-switching was partly explained in Fig. 8e. Following the reviewer's suggestion, we now add some discussions to further elaborate the computational roles of MD and interneuron subtypes in cognitive functions at various stages of flexible decision-making (cue integration, working memory, context switching).

"In modulating cue uncertainty during the cueing period, the MD_1 subpopulation that targets on PV interneurons seems to play a more important role in cue evidence integration (**Supplementary Fig. 2d** and **Supplementary Table 1**). In working memory during the stimulus-free delay period, activating MD_1 or MD_2 subpopulation can enhance thalamocortical communications and boost working memory (**Fig. 4b**, **c**, **g** and **Fig. 8a**). In cue-to-rule remapping, from a geometric viewpoint, the MD orients its neural subspace to accommodate context switching while maintaining context-invariant prefrontal rule representations (**Fig. 8i**)." ***(Lines 860-867)***

Reviewer Comment 4.33 — Overall, thorough well written Methods

Answer: Thanks very much for your comments.

Reviewer Comment 4.34 — line 648: Why equal percentages of the different interneuron subtypes? This is does not match real proportions. Also, what percentage are excitatory and inhibitory neurons in the network? I ask this earlier, but this should probably be noted in both locations due to its importance.

Answer: According to Dale's principle, the ratio of Exc and Inh (including PV, SOM, VIP) is 4:1. For simplicity, we have assumed the equal percentage of three different inhibitory neurons. In our computer simulations, the total number of PFC units is 256, N_{PFC}=256, so we set N_{exc}=205, and N_{inh}=51 (N_{PV}= N_{SOM} =N_{VIP}=17). We now provide those missing details in the Methods section.

According to the literature, among all cortical inhibitory neurons, the percentage of PV neurons in the cortex is 30-50% of all cortical GABAergic neurons, the percentage of SOM neurons is 30% of all cortical GABAergic neurons, and the percentage of VIP neurons is 10-15% of all cortical inhibitory neurons. Following the review on cortical interneuron (Rudy et al., 2011 *Dev. Neurobiol),* we also explore assumption of nonequal percentages of three different interneuron cell types (PV:SOM:VIP = 40%:30%:30%;); namely, N_{PV}= 21, N_{SOM} =N_{VIP}=15, and the results remain qualitatively similar (see a summary in the new **Supplementary Table 1**).

Rudy B, Fishell G, Lee S, Hjerling-Leffler J. Three groups of interneurons account for nearly 100% of neocortical GABAergic neurons. Dev Neurobiol. 2011 Jan 1;71(1):45-61. doi: 10.1002/dneu.20853.

Reviewer Comment 4.35 — line 701-702 missing word
Answer: We now change the original sentence as

"….outlier models did not show good generalization across all cue conditions (e.g., making one type of rule-specific error more frequently or showing specific choice biases); these outlier models have been excluded in our analyses." ***(Lines 1084-1086)***

Reviewer Comment 4.36 — Lines 721-722 – earlier you state the # of MD1 and MD2 neurons in your model is equal, but then here you use an unequal # (18 vs 12); why?
Answer: Thanks for your comment and we apologize for the confusion. In the first version of the model, we did assume that the size of MD_1 and size of MD_2 are equal (Fig. 1a). In the subsequent model, we tested whether the results would change if we further incorporated two biological details (from unpublished mouse data): (i) the size of MD_1 is greater than the size of MD_2; (ii) the projection of PFC Exc→MD_2 was sparser and weaker than that of Exc→MD_1. This has been mentioned in Fig. 6a figure legend.

"The modified PFC-MD model with assumed unbalanced MD subpopulations (as shown by two different circle sizes, where the size of MD_1 is greater than the size of MD_2) and sparser Exc→MD_2 connectivity than Exc→MD_1 connectivity (as shown by a weaker connection in dashed line)." ***(Lines 515-517)***

Reviewer Comment 4.37 — line 747: How are you doing three way comparisons using rank sum? Also of note, it is generally when or where in the results you used this procedure
Answer: Thanks for your valuable suggestions. To avoid multiple-way comparisons, we now change the criterion of cue uncertainty tuning criterion in the METHODS. Specifically, we assume a linear model between the mean MD unit firing rate (during cueing period) and the cueing uncertainty (e.g., coherence or congruence)

$$\text{Mean FR} = \alpha + \beta \cdot \text{cue uncertainty level}$$

where α and β denote the regression coefficients. We fit the model prediction between mean FR and cue uncertainty level, and further characterize the percentage of explained variance (EV) for two different contexts (e.g., CAC and RDM contexts). We call the MD units are strongly modulated with cueing uncertainty if it simultaneously meets the strong modulation criterion (e.g., %EV > 80%) for both contexts. See our response to #8 above.

Reviewer #4 (Remarks on code availability):

I confirmed that the code is available on github and has an accompanying readme that describes how to generate the main figures in python.

Thanks for your comments. We have provided additional code in Github to reflect additional experiments in this Revision.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The authors did their best to address our concerns. We recommend mentioning in the text that the approach incorporating many cell types has conceptual limitations, which new supp table 1 does not fully mitigate.

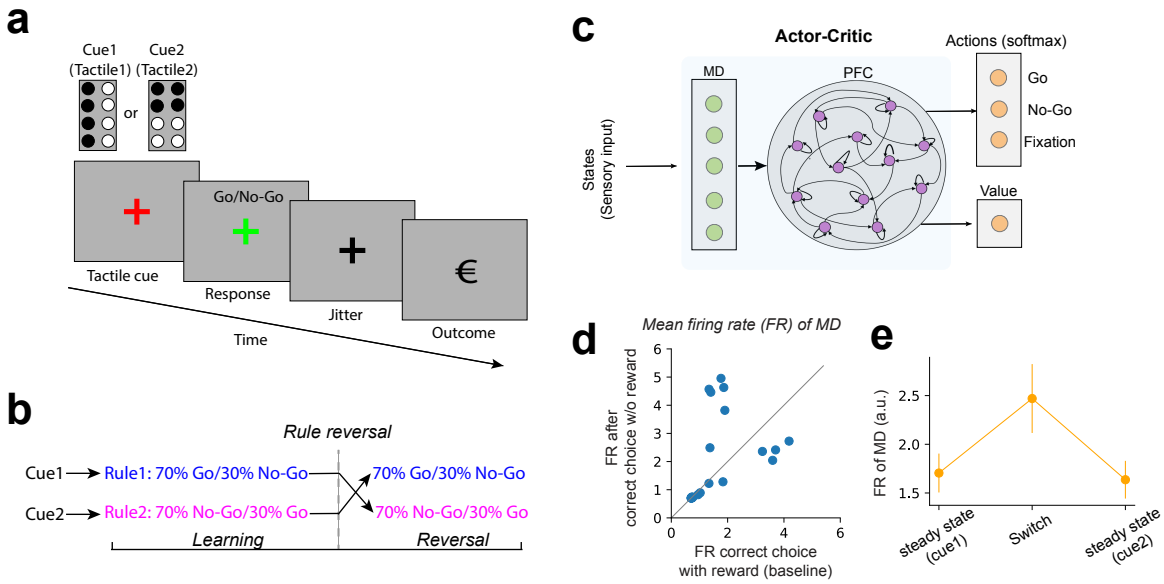
Answer: Thanks for your valuable comments. Following your recommendation, we now mention in the text (last paragraph of the paper) that our approach still has conceptual limitations, which will be addressed in the future investigation.

"Therefore, our model with only cell type specificity still has conceptual limitations." (line 919)

Reviewer #3 (Remarks to the Author):

The authors have clearly undertaken an extensive set of analyses to appease the reviewers. Although they superficially answered a number of my concerns, I remain underwhelmed by the strength and specificity of the evidence provided by their approach.

Answer: Thanks for your resourceful critiques. One the one hand, it is a bit disappointing that we were not yet able to convince the reviewer the strength of our approach. Not sure if this additional experiment would further help convincing the reviewer: we have succeeded in using reinforcement learning to train the proposed PFC-MD model in a human decision-making task with outcome uncertainty (task described in Wang et al. *Nat. Commun*. 14, 3552, 2023). To date, we have found some MD units increase firing rate to encode reward prediction error, which matched the unpublished human fMRI MD activation results during context switching.



Response Fig. 1 Probabilistic Go/NoGo reversal learning task in humans [1]. **a** Timeline illustrating a single trial. **b** In this block, Cue1 (Tactile1 stimuli) was associated with 'Go' in 70% of the trials, while Cue2 (Tactile2 stimuli) was associated with 'No-Go' in 70% of the trials. **c** Actor-Critic reinforcement learning based on the PFC-MD model. **d** Scatter plots of selected MD unit firing rates in the presence of unexpected RPE (or representing the "surprise signal"); note that some units increased or decreased firing rates compared to the baseline ("no surprise"). **e** MD population (n=32) showed increased modulation with respect to the predict error during the transient switching stage. The MD population firing rate was averaged overall all units during the delay period when the change of network state became very small (i.e., steady state). Error bar denotes s.e.m.

On the other hand, we are motivated to further substantially expand the model by considering several extensions: (i) incorporating cortico-cortical communications in goal-driven reward-based decision-making, for instance, the PFC-ACC or PFC-OFC communications in the context of modeling the transthalamic pathway; (ii) incorporating new subcortical inputs from basal ganglia (BG) structures that differentially project to different MD subpopulations in the context of reinforcement learning for reward prediction error feedback; (iii) investigating the impact of impaired cortico-thalamic-BG circuit rewarding mechanism on behaviors in the context of schizophrenia. Taken together, we sincerely hope the continuing effort of modeling these connectivity and pathways based on the foundation work proposed in our current paper can further improve both strength and specificity and share the modeling tool with a large neuroscience community.

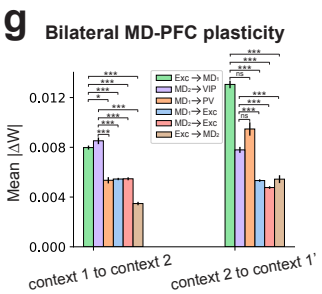
Reviewer #4 (Remarks to the Author):

Overall, the manuscript is greatly improved via clarifications of the model and tremendously improved statistical analysis. The authors did a thorough job with the rebuttal and associated revision. Overall, the computational modeling and data remain well developed and detailed, the topic of MD-PFC interactions in support of cognition is important and of wide interest, and the overall depth of analysis is tremendous. With the elaborations in the revision, the manuscript is much more methodologically sound, and additional detail has been provided that will enhance reproducibility. The only comment I have that does not preclude publication is below.

Answer: Thanks for your valuable comments.

MINOR COMMENT:
-please re-review Figure 7g and associated text; I think it could be improved further via some light revision and/or additional stats, but it is not absolutely necessary; will leave it up to the reviewers.

Answer: Thanks for your valuable comments. Figure 7g result was generated from one trained PFC-MD model, and the error bar was computed based on the number of specific connectivity. We now rearrange the bar plot and provide additional stats in the figure legend.



Response Fig. 2g Quantification of mean synaptic plasticity (error bar: s.e.m.) in bilateral MD-PFC connections during two consecutive context-switching conditions. In general, MD1→Exc, MD2→Exc, and Exc→MD2 connections showed the overall smallest changes, whereas Exc→MD1 and MD2→VIP connections showed the overall dominant changes. All pairwise p-values were computed based on rank-sum test (*p<0.05, ***p<0.001).

Reviewer #4 (Remarks on code availability):

There is a clearly written readme
Folders are well organized
There is a link in readme to trained models, but many of the folders in the google drive folders are empty

Answer: Thanks for your comments and we apologize for the incorrect uploads. We have provided additional files to the Google drive.