nature portfolio

Peer Review File

Flexible gating between subspaces in a neural network model of internally guided task switching

Editorial Note: This manuscript has been previously reviewed at another journal that is not operating a transparent peer review scheme. This document only contains reviewer comments and rebuttal letters for versions considered at *Nature Communications*.

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The manuscript is improved by inclusion of additional analyses and information. Authors have added figure 1e from Kamigaki et al. 2011 and show it with Figure 1d to indicate that monkeys and the trained RNN could shift between rules after a single error. The authors proceed to mention that because the RNN has a supervised learning, it could achieve such a high performance following a single error. In addition, they run new analyses showing that the network behaves differently (requires more trial and errors) if a criterion (less than 100%) is set for the shifting between rules (Figure 1f). The authors also mention that in this condition there is not a clear separation between the two subpopulations of excitatory neurons in the PFC module, however do not present the related data. In fact, setting a shift criterion is normally considered in various rule shifting tasks (including the WCST). Therefore, it is unclear why the related data is not shown. It appears necessary to distinguish these two different behaviors of the network.

The authors have added informative statements in the revised manuscript to explain the discrepancy between their RNN performance and the results of multiple studies showing that the shift between rules could not be achieved after a single error: ""Our networks switch rules in just one trial (Figure 1c, d). This fast switching agrees with the monkey behavior in some studies [21, 9, 27], but other studies report that monkeys switch rules using on average tens of trials....In this case, the dissected network mechanism is compatible with what we reported here, albeit not as clear cut.". "There are other reasons that may contribute to the suboptimality of behavior during rule switching."

These statements clearly raise the problem of explaining the behavior of biological organisms facing the necessity for rule-based action selection in a rapidly changing environment with the proposed artificial RNN models. Therefore, the scope of this proposed RNN remains limited. It is helpful for readers to understand these limitations of the proposed RNN in explaining many studies in monkeys performing rule-shifting tasks.

The problem in finding the correspondence in performance between the proposed RNN and monkeys is somehow related to the way the authors have classified and interpreted previous studies in humans and monkeys.

The rule shift tasks can be generally classified as cued and uncued tasks and the proposed RNN needs to distinguish between different context (cued vs uncued):

(A) In cued rule shift tasks, the shift in the relevant rule is explicitly or implicitly informed and therefore the agent (RNN or neurocircuitries) have clear cut information to modify the network and

shift between rules. In this context, rapid shifting with a single error or even in the absence of an error (if the cue is given) might be possible.

(B) In uncued rule shift tasks, the rule shift or the relevant rule is not informed (notified) and therefore various additional mechanisms (and related neurocircuitries) need to be recruited (such as working memory, assessment of behavioural outcome) to infer the rule shift and maintain the relevant rule.

In uncued rule shift tasks (such as Buckley et al. (2009) and Mansouri et al. 2006), rapid shifting is highly demanding in terms of executive control processes and therefore is rarely seen. The proposed RNN does not clearly explain these situations/tasks (and related processes) in uncued rule shift tasks.

However, the RNN can explain the behavior of young adult humans performing WCST and monkeys performing cued rule shift task. In Kamigaki et al. study, the spatial position of color and shape information differed and therefore the monkeys could orient their attention toward the peripheral or central objects to perform the task. This acted as an implicit cue to the relevant rule. This was confirmed by the significant difference in monkeys' eye position between the two rules/blocks (Kamigaki et al. Journal of Cognitive Neuroscience 23 2011); Kamigaki et al. also raise this possibility (Page 2518): "Moreover, our analysis indicated that gaze eccentricity was significantly different between the shape and color dimensions during the sample and choice periods, therefore, we cannot rule out the alternative possibility that the monkeys might have used a strategy for solving our task to shift their gaze or spatial attention between the small color object and the large shape object.". Therefore, this study can be considered as a cued rule shift task and therefore a rapid shifting between rules (after rule change) is expected. The proposed RNN also behaves similarly.

I suggest the authors clearly distinguish between cued and uncued rule shift tasks. This will greatly improve the manuscript and make the message clear in terms of its scope and translatability to real biological conditions. Without such a distinction, the context in which the neurocircuitry (or RNN) is performing remain unclear and therefore the behavior of RNN remains unrealistic or at most only explains a part of the story without clearly explaining why biological neurocircuitries behave differently.

Minor issue:

I suggest the authors explain the study by Stanislas Dehaene, Jean-Pierre Changeux; Cerebral Cortex, Volume 1, 1991, Pages 62–79) in the Discussion section. The attractor model for maintaining rule (in the WCST and other rule-shifting tasks) is not a new concept. Authors need to acknowledge the previously proposed models in which the attractor models for shifting and maintaining rules were predicted. Then, they can clearly explain what has been added in their new proposed RNN.

Reviewer #2 (Remarks to the Author):

The revised manuscript has clarified most of my concerns. In particular, I appreciate that the authors have shown the self-organization of winner-take-all inhibition between PFC SST neurons. However, I still have a few minor comments.

In you response to my comment 5, I also wanted to see the role of PFC VIP that regulates the PFC SST neurons. These VIP cells do not receive any input extrinsic to the PFC module. This may be fine, as VIP cells are thought to mediate top-down input while the PFC module has no upstream regions in the present model. However, a brief comment on anatomically known long-range inputs to PFC VIP neurons will be useful for the readers.

In addition, I am curious about the role of PFC VIP cells in this model. What happens if VIP cells are impaired in the PFC module after task training? Are these cells crucial for the model's performance? These points are minor and I stated them only vaguely in my previous comment. So, I do not ask the authors to include new simulation results. The authors may consider these points in a revision if they improve the manuscript.

Reviewer #2 (Remarks on code availability):

I only know C language and could not check the code. However, all simulation results look reasonable, at least qualitatively.

Overall response

We thank the reviewers for their thoughtful feedback. Reviewers acknowledged the improvement in the revised manuscript (Reviewer 1: *"The manuscript is improved by inclusion of additional analyses and information".* Reviewer 2: *"The revised manuscript has clarified most of my concerns. In particular, I appreciate that the authors have shown the self-organization of winner-take-all inhibition between PFC SST neurons"*). According to the reviewers' feedback, we conducted additional analyses to address the concerns raised by the reviewers. These analyses led to a new Supplementary Figure 6 and a new panel in Supplementary Figure 5 (panel f), two new sections in Discussion and a rewritten paragraph in Discussion. The main revisions are the following:

1. In response to Reviewer 1's concern that the data for the connection weight matrix underlying the slow switching networks was not shown, we created a new Supplementary Figure 6 to show the relevant data. We believe this provides convincing evidence for the difference in the network structure between the slow switching models and fast switching models.

2. In response to Reviewer 1's comment that cued and un-cued rule switching should be more clearly distinguished in the manuscript, we added a section in the Discussion on the difference between cued and un-cued rule switching in terms of both behavioral performance and required neural circuitry.

3. In response to Reviewer 1's comment that the model proposed by Dehaene and Changeux should be more carefully explained and compared with our current model, we re-wrote the relevant paragraph in the Discussion to show that both models use the attractor mechanism to maintain the rule, but different mechanisms to shift between rules.

4. In response to Reviewer 2's comment on the role of VIP cells in the PFC module in this task, we conducted an additional simulation where these neurons were silenced after training and presented the result in Supplementary Figure 5f. This simulation showed no significant performance decrease when the VIP cells were silenced.

5. In response to Reviewer 2's comment on the input that VIP cells in the PFC receive in the brain, we created a section in the Discussion on the known anatomical projections to the PFC VIP cells and their functions. We acknowledge that it is an interesting future direction to include these various inputs to build a more biologically realistic model.

Below we provide a point-by-point response to the reviewers' comments.

Point-by-point response to the reviewers' comments

Reviewer #1 (Remarks to the Author):

The manuscript is improved by inclusion of additional analyses and information. Authors have added figure 1e from Kamigaki et al. 2011 and show it with Figure 1d to indicate that monkeys and the trained RNN could shift between rules after a single error. The authors proceed to mention that because the RNN has a supervised learning, it could achieve such a high performance following a single error. In addition, they run new analyses showing that the network behaves differently (requires more trial and errors) if a criterion (less than 100%) is set for the shifting between rules (Figure 1f).

Thank you for the acknowledgement of the improvement made in the revision. We would like to respectively mention that in fact, we used the lower criterion to decide when to stop training the networks: the training (i.e. adjustment of the network weights through backpropagation) was stopped when the network performed correctly for 80% of the trials. In some sense this is related to the criterion for deciding when to shift rules which is commonly used in experiments, because using a lower criterion for shifting rules would presumably lead to a lower overall performance.

The authors also mention that in this condition there is not a clear separation between the two subpopulations of excitatory neurons in the PFC module, however do not present the related data. In fact, setting a shift criterion is normally considered in various rule shifting tasks (including the WCST). Therefore, it is unclear why the related data is not shown. It appears necessary to distinguish these two different behaviors of the network.

Thank you for the suggestion. In the revised manuscript we included a figure (Supplementary Figure 6) showing the separation of the two subpopulations of excitatory neurons in the PFC module, as well as the connectivity biases between them for the slow-switching models. As can be seen, there is a weaker separation between the two subpopulations compared to the fast-switching models. In addition, the connectivity biases between rule neurons remain greater than 0, same as the fast-switching networks. In contrast, the connectivity biases between rule neurons and conjunctive error x rule neurons in the slow-switching networks are not significantly greater than 0. This indicates that while the slow switching models has a winner-take-all subnetwork which maintains the rule, it has a more random connectivity structure

between the rule neurons and the conjunctive error x rule neurons, which causes the slowness in switching between rules.

The authors have added informative statements in the revised manuscript to explain the discrepancy between their RNN performance and the results of multiple studies showing that the shift between rules could not be achieved after a single error: ""Our networks switch rules in just one trial (Figure 1c, d). This fast switching agrees with the monkey behavior in some studies [21, 9, 27], but other studies report that monkeys switch rules using on average tens of trials….In this case, the dissected network mechanism is compatible with what we reported here, albeit not as clear cut.". "There are other reasons that may contribute to the suboptimality of behavior during rule switching."

These statements clearly raise the problem of explaining the behavior of biological organisms facing the necessity for rule-based action selection in a rapidly changing environment with the proposed artificial RNN models. Therefore, the scope of this proposed RNN remains limited. It is helpful for readers to understand these limitations of the proposed RNN in explaining many studies in monkeys performing rule-shifting tasks.

The problem in finding the correspondence in performance between the proposed RNN and monkeys is somehow related to the way the authors have classified and interpreted previous studies in humans and monkeys.

The rule shift tasks can be generally classified as cued and uncued tasks and the proposed RNN needs to distinguish between different context (cued vs uncued):

(A) In cued rule shift tasks, the shift in the relevant rule is explicitly or implicitly informed and therefore the agent (RNN or neurocircuitries) have clear cut information to modify the network and shift between rules. In this context, rapid shifting with a single error or even in the absence of an error (if the cue is given) might be possible.

(B) In uncued rule shift tasks, the rule shift or the relevant rule is not informed (notified) and therefore various additional mechanisms (and related

neurocircuitries) need to be recruited (such as working memory, assessment of behavioural outcome) to infer the rule shift and maintain the relevant rule.

In uncued rule shift tasks (such as Buckley et al. (2009) and Mansouri et al. 2006), rapid shifting is highly demanding in terms of executive control processes and therefore is rarely seen. The proposed RNN does not clearly explain these situations/tasks (and related processes) in uncued rule shift tasks.

However, the RNN can explain the behavior of young adult humans performing WCST and monkeys performing cued rule shift task. In Kamigaki et al. study, the spatial position of color and shape information differed and therefore the monkeys could orient their attention toward the peripheral or central objects to perform the task. This acted as an implicit cue to the relevant rule. This was confirmed by the significant difference in monkeys' eye position between the two rules/blocks (Kamigaki et al. Journal of Cognitive Neuroscience 23 2011); Kamigaki et al. also raise this possibility (Page 2518): "Moreover, our analysis indicated that gaze eccentricity was significantly different between the shape and color dimensions during the sample and choice periods, therefore, we cannot rule out the alternative possibility that the monkeys might have used a strategy for solving our task to shift their gaze or spatial attention between the small color object and the large shape object.". Therefore, this study can be considered as a cued rule shift task and therefore a rapid shifting between rules (after rule change) is expected. The proposed RNN also behaves similarly.

We thank the reviewer for bringing to our attention the issue about eye position of the monkeys in the Kamigaki et al. study. It is indeed true that the monkeys' eye position significantly differs across different rule blocks in that study. We also acknowledge that our RNNs do not incorporate the fine details such as gaze in the output.

However, it does not seem to us that this difference in eye position could act as an implicit rule cue, because the difference in eye position is a consequence of monkeys using different rules (color vs shape) to explore the objects. However, unlike a rule cue, it does not provide the information to the monkeys about which rule they should currently follow.

I suggest the authors clearly distinguish between cued and uncued rule shift tasks. This will greatly improve the manuscript and make the message clear in terms of its scope and translatability to real biological conditions. Without such a distinction, the context in

which the neurocircuitry (or RNN) is performing remain unclear and therefore the behavior of RNN remains unrealistic or at most only explains a part of the story without clearly explaining why biological neurocircuitries behave differently.

We thank the reviewer for this suggestion. In the revised manuscript we added the following text in the Introduction about the difference between cued and un-cued rule switching tasks:

"… This is different from cued rule switching paradigms where an external cue provides information to the subjects about the relevant rule for each trial (e.g. [2, 3])."

"In contrast, in cued rule switching tasks, the first two computations (rule maintenance and rule updating) are not required, since an external cue about the rule is provided during every trial."

In the Discussion section we added a section "Cued versus un-cued rule switching" to further elaborate on the different computational requirements for un-cued rule switching tasks compared to cued rule switching tasks, and the difference in animal performance during these two types of tasks.

"The task our networks were trained on is an example of un-cued rule switching task, where the rule is not explicitly cued on each trial but needs to be inferred from the feedback of previous trials. This differs from cued rule switching, another commonly studied rule-based task paradigm where the task rule is explicitly cued during each trial (e.g. [3, 2]). Compared to cued rule switching, additional mechanisms and neural circuitries are needed to perform un-cued rule switching. These include the mechanisms and neural circuitries for inferring the rule shift, and maintaining the relevant rule across multiple trials.

This difference in the required neural mechanisms leads to a difference of task performance between cued and un-cued rule switching: while it is possible to shift rules with few errors in cued rule switching tasks, such behavior is highly demanding in un-cued rule switching tasks [10, 8]. Below we will further discuss the limitation of our work in terms of the dynamics of the behavior during rule switching."

Minor issue:

I suggest the authors explain the study by Stanislas Dehaene, Jean-Pierre Changeux; Cerebral Cortex, Volume 1, 1991, Pages 62–79) in the Discussion section. The attractor model for maintaining rule (in the WCST and other rule-shifting tasks) is not a new concept. Authors need to acknowledge the previously proposed models in which the attractor models for shifting and maintaining rules were predicted. Then, they can clearly explain what has been added in their new proposed RNN.

We thank the reviewer for this suggestion. In the revised manuscript we rewrote the part of the Discussion section to elaborate on the similarity and differences between the Dehaene and Changeux model and our model.

"This circuit mechanism bears resemblance to a previous circuit model of WCST [69]. In that model, different rules are maintained by clusters of rule-coding neurons that have self-excitatory connections and lateral inhibitory connections, similar to the winner-take-all attractor network between the rule neurons in our trained networks. On the other hand, these two models use different mechanisms for rule shifting: while in the Dehaene and Changeux model it is achieved via synaptic desensitization caused by the convergence of two signals - one that signals the recent activation of the synapse, and another that signals negative feedback, our trained networks use structured connections between rule neurons and a separate population of conjunctive error x rule to shift between rules neurons (c.f. Figure 3c). Notably, neurons that represent the conjunction of negative feedback and rule have been reported in the frontoparietal cortices of monkeys performing WCST [10, 11]."

We also added the following paragraph to compare our model to another model introduced by Rigotti et al. to perform un-cued context switching tasks.

"Rigotti et al. proposed a network model that can perform un-cued switching between attractor states (e.g. rules in WCST) using nonlinear mixed-selective neurons. In that model, a randomly connected network of neurons is connected to other populations of neurons that represent different attractors. It was shown that neurons in the random network naturally exhibit mixed selectivity of negative feedback and the current attractor state, and their interactions with the attractor neurons enable transitioning between attractor states. In our trained networks, the mixed selectivity arises from structured patterns of connections with the rule neurons (c.f. Figure 3c). It remains to be seen whether the pattern of connection strengths between the random network and the rule

neurons in the model by Rigotti et al. bear resemblance to that between the conjunctive neurons and the rule neurons in our networks."

Reviewer #2 (Remarks to the Author):

The revised manuscript has clarified most of my concerns. In particular, I appreciate that the authors have shown the self-organization of winner-take-all inhibition between PFC SST neurons. However, I still have a few minor comments.

In your response to my comment 5, I also wanted to see the role of PFC VIP that regulates the PFC SST neurons. These VIP cells do not receive any input extrinsic to the PFC module. This may be fine, as VIP cells are thought to mediate top-down input while the PFC module has no upstream regions in the present model. However, a brief comment on anatomically known long-range inputs to PFC VIP neurons will be useful for the readers.

In addition, I am curious about the role of PFC VIP cells in this model. What happens if VIP cells are impaired in the PFC module after task training? Are these cells crucial for the model's performance? These points are minor and I stated them only vaguely in my previous comment. So, I do not ask the authors to include new simulation results. The authors may consider these points in a revision if they improve the manuscript.

Thank you for your suggestion. In the revision we included a Supplementary Figure 5f showing the performance of the network when the VIP cells in the PFC module were silenced. We found that the performance does not change significantly after the PFC VIP cells were inhibited. This is consistent with the fact that these cells do not receive excitatory input in our model. We agree that it would be an interesting future direction to include long-range projections from other cortical and subcortical areas to the current model.

In the revised manuscript we included the following text on the anatomically known long-range inputs to PFC VIP neurons as well as the role of PFC VIP neurons in our model:

"The SST neurons in the PFC module of our networks are targeted by the local VIP neurons. However, since the VIP neurons in our modeling setup do not receive any excitatory input, their inhibition onto SST neurons is always released. In other words, the excitatory neurons in the PFC module are in the dis-inhibited regime. Indeed, inhibiting the PFC VIP neurons in our networks did not significantly affect task performance (Supplementary Figure 5f). In reality, VIP neurons in the PFC module receive long-range connections from the mediodorsal thalamus [84] which carries information about task context [85]. In addition, their activity is also strongly modulated by negative reinforcement [15] as well as arousal level [86], which suggest that they receive long-range projections from subcortical regions such as Ventral Tegmental Area, Locus Coeruleus [87] and the Raphe Nucleus [88]. Including additional modules of these subcortical regions would be an interesting direction for future work."

Reviewer #2 (Remarks on code availability):

I only know C language and could not check the code. However, all simulation results look reasonable, at least qualitatively.

REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

The manuscript is improved by adding the additional explanations and a new supplementary figure. Authors also discuss and distinguish the cued and uncued rule-shifting tasks, which has helped to further clarify how the proposed RNN describes the behavior of neurobiological systems.

However, I am afraid the issue with interpretation of Kamigaki et al. study still remains. As mentioned previously, In Kamigaki et al. study, the spatial position of color and shape information differed and therefore the monkeys could orient their attention toward the peripheral or central objects to perform the task. The significant difference in monkeys' eye position, when they were performing the shape or color dimension (rule) tasks, was lingering from the previous trials (appeared during the sample period). This indicates that monkeys were considering the differences in the spatial position of shape and color information as a cue to the currently relevant rule. In other words, the difference in the 'spatial position of shape or color information acted as a cue' to the currently relevant rule and the significant difference in monkeys' eye position is clear evidence that monkeys noticed and used this cue. Kamigaki et al. also raise this possibility (Page 2518): "Moreover, our analysis indicated that gaze eccentricity was significantly different between the shape and color dimensions during the sample and choice periods, therefore, we cannot rule out the alternative possibility that the monkeys might have used a strategy for solving our task to shift their gaze or spatial attention between the small color object and the large shape object".

Point-by-point response to reviewer's comments

Reviewer #1 (Remarks to the Author):

The manuscript is improved by adding the additional explanations and a new supplementary figure. Authors also discuss and distinguish the cued and uncued rule-shifting tasks, which has helped to further clarify how the proposed RNN describes the behavior of neurobiological systems.

- Thank you for your acknowledgement of our revision.

However, I am afraid the issue with interpretation of Kamigaki et al. study still remains. As mentioned previously, In Kamigaki et al. study, the spatial position of color and shape information differed and therefore the monkeys could orient their attention toward the peripheral or central objects to perform the task. The significant difference in monkeys' eye position, when they were performing the shape or color dimension (rule) tasks, was lingering from the previous trials (appeared during the sample period). This indicates that monkeys were considering the differences in the spatial position of shape and color information as a cue to the currently relevant rule. In other words, the difference in the 'spatial position of shape or color information acted as a cue' to the currently relevant rule and the significant difference in monkeys' eye position is clear evidence that monkeys noticed and used this cue. Kamigaki et al. also raise this possibility (Page 2518): "Moreover, our analysis indicated that gaze eccentricity was significantly different between the shape and color dimensions during the sample and choice periods, therefore, we cannot rule out the alternative possibility that the monkeys might have used a strategy for solving our task to shift their gaze or spatial attention between the small color object and the large shape object".

- Thank you for pointing this out and sorry for not properly understanding this point earlier. It is indeed the case that monkeys can use their eye position to remind themselves about the rule from the last trial, thereby potentially decreasing the working memory requirement to maintain the rule across trials.

  In the revised manuscript we mentioned this point in the following text in the discussion section:

  "Compared to cued rule switching, additional mechanisms and neural circuitries responsible for inferring the rule shift and maintaining the relevant rule are needed to perform un-cued rule switching (although monkeys could employ alternative strategies that require less cognitive load for rule maintenance, as mentioned in [28])"