Response to Reviewers

Title: Nucleus Basalis Stimulation Enhances Working Memory by Stabilizing Attractor Networks in Prefrontal Cortex

Authors: Qi et al.

We are grateful to the reviewers for their careful reading of our manuscript, their insightful comments, and their overall positive evaluation. Some common concerns were present in both reviews. We have performed extensive additional work to address all issues raised by the reviewers (*in italics below).* This work led us to draw additional insights, and we have qualified some of our conclusions. We believe the manuscript is much stronger as a result.

Summary:   
*We all liked the strong integration of modeling and biology and think that the paper has the potential to influence our understanding of how ACh alters computation in prefrontal cortex. That said, we all agreed that a key point needs to be addressed regarding the need to quantify the quality of the population representations using decoding before/during NB stim. In particular, it is key to determine if the information content (inferred via decoding or some other approach) goes down or up with NB stimulation (and we collectively think that it could go either way depending on whether stimulation influences signal strength, noise covariance, etc.). This is particularly true because currently it is difficult to square the drop in selectivity with the improved behavior.*  
Response: The editor’s comments were well taken. We have addressed these issues and explain below.

Reviewer #1:   
*Qi and colleagues investigate the effects of NB electrical stimulation on performance and on PFC neuronal activity during a working memory task. They find that NB stimulation improves performance, but increases PFC activity non-selectively, leading to a reduced spatial selectivity and broader tuning. Using a computational model, they offer a possible explanation to reconcile these results. The question is relevant both from the Neuroscience and clinical perspectives. The observed performance improvement and increased PFC spiking are relevant results worth publishing. However, it is difficult to reconcile the decrease in neuronal selectivity with the behavioral improvement. Yet the authors seem to make an effort to justify how PFC selectivity decreases could lead to behavioral improvements with arguments that are not entirely convincing, instead of acknowledging the conflicting results and perhaps proposing alternative mechanisms involving other brain areas potentially affected by the stimulation. Besides this, the manuscript has multiple other issues that need to be solved (major and minor), which I describe in detail.   
  
MAJOR COMMENTS:   
1) In several instances, the authors claim that tuning broadening, which accompanies a decrease in selectivity, could somehow be associated with an improvement in performance (e.g., pg. 10). This claim is very controversial, since it completely ignores the basic principle that, in the presence of neuronal activity variability, lower selectivity implies LESS accurate information read-out. If the authors really believe that, somehow, the broader tuning observed in the stimulation condition leads to more accurate readout of information at the end of the delay, they should test this by, for example, measuring population decoding of the memorized location throughout the delay (or measuring this information at the individual neurons level), and comparing it between the stimulation and control conditions. My prediction is that the lower selectivity in the stimulation condition will lead to lower accuracy in decoding the memorized location compared to control, and that broader tuning will not mean better decoding.   
Given the inconsistency between the behavioral improvement in the working memory task and the decrease in neuronal selectivity caused by NB stimulation, I suggest that the authors discuss alternative explanations to the behavioral improvement involving effects of stimulation on areas OTHER than PFC. After all, working memory is a widely-distributed function that likely involves several areas other than PFC, and the authors never tested NB stimulation effects elsewhere.*

Response: In response to comments from all reviewers, we have qualified our conclusions and revised the text. We now indicate that selectivity of neuronal responses declined across conditions. Decoding of stimulus location, as evaluated by the Percentage of Explained Variance, also decreased. This result is consistent with pharmacological studies that relied with high doses of cholinergic stimulation, and explains the decrease in performance for conditions in which decoding of the precise stimulus location is essential. However, for conditions in which stimuli were highly discriminable, and performance depended on the ability to filter distractors and implement the correct rule, NB stimulation improved performance and enhanced the accuracy of behavioral responses. These results can be explained in the context of the bump-attractor model that predicts more stable stimulus representations.

*2) Most of the neuronal statistical analyses in the manuscript are performed only at the level of the population (across all neurons). Given that it is well known that lateral prefrontal neurons vary widely in their functional properties (in contrast to, say, visual areas), it seems essential to perform these analyses also at the individual neuron level using variability across trials for statistical purposes. Below are examples of some of those analyses: a) While the absence of neuronal effects in the inter-trial interval and in the phasic responses was observed in an across-neurons analysis, this trend might not apply to all neurons individually. The absence of across-neurons effects could even be due to some neurons having an enhancement and others a suppression. The authors should run statistical analyses at the individual neuron level to show the percentage of neuron showing effects, and the sign of the effects.*

Response: The reviewer’s point is well taken. We now plot results from individual neurons in Fig. 4 and Fig. S6 and report statistics in the text.

*b) Similarly, the results of spatial tuning are described in the main text and in Figure 4B-E,I-J only at the level of mean firing rate across neurons. For each neuron, the authors should measure the change in tuning, and test whether the change is statistically significant. They can then report the proportions of neurons with significant increases, decreases, or no changes to their tuning, and somehow display the distribution of tuning widths of all neurons for stimulation and control trials. c) Besides tuning width, the authors should quantify selectivity as the amount of information about location carried by each neuron, as many recent studies do. The selectivity index used by the authors is not informative enough, since it does not use trial-by-trial variability. To measure information, there are several methods that are easy to implement, including ANOVA percent explained variance (PEV) or ROC analysis - for individual neurons, or population-level methods such as decoding.*

Response: The reviewer is certainly right here as well, however many neurons were not well-fitted by a single Gaussian and for many of those that did we did not have sufficient power to detect a significant difference in standard deviation. Instead, we used the Selectivity Index as a proxy of the full tuning curve on a neuron-by-neuron basis. We clarify that now. We have accepted the reviewer’s suggestion on point c) and we now additionally report results of ROC analysis (Fig. S8) and PEV (Fig. S9). The results indicate lower levels of information encoded by neurons, as discussed in point 1, above.

*Minor Comments:   
1. Related to the 3-way ANOVA on pg. 5: Was each of the 3 factors within-sessions or between-sessions? This is essential information, but is not described anywhere in the manuscript. In every single ANOVA in the study, it is essential to describe whether each factor was a within- or between-sessions factor. According to the authors' description of the experimental design, the stimulation condition (off vs. on) in recording sessions is a within-sessions factor, given that each session has a performance score for both stimulation conditions - off/control and on. However, according to the Figure 2 legend, the number of sessions for control and on conditions is not the same. I assume this is because the authors included some sessions in which there was no stimulation. However, including those sessions, and not just the sessions with both on and off conditions, unnecessarily forces the stimulation condition to become a between-sessions factor rather than within-sessions. This means that the between-sessions variability can affect the overall mean performance across sessions. Given that in no-stimulation sessions the stimulating electrode was not lowered, this alone could be a confounding factor. A more appropriate approach is to only analyze sessions that included stimulation off AND on, and statistically compare performance in off and on conditions as paired within-session. This can be done with an ANOVA in which stimulation condition is a within-sessions factor. If the authors also want to include the results that use all sessions (including the ones without stimulation), they could present it as a Supplementary Figure, since it is a less powerful and less optimal analysis than the within-sessions analysis.*

Response: We now clarify that NB stimulation was a between-sessions factor of the 3-way ANOVA. Behavioral performance could be best compared in different daily sessions, alternating between days with and without stimulation. In neural recordings, stimulation always followed the control session, so performance could not be compared in the same way. The reason for this experimental design choice was that the time course of the NB stimulation effect is unknown, though pilot studies from our group suggested a long time course, that could not be effectively “washed” within a single recording session. In the original manuscript, the source of the behavioral data was explained in the paragraph preceding the text the reviewer quotes and the figure legend, exactly as the reviewer noted. We clarify further now.

*2. The current title of the manuscript is inappropriate, since it implies that the study demonstrated that the mechanism by which NB stimulation enhances working memory performance enhancement is indeed "stabilizing attractor networks in prefontal cortex". But there was no such demonstration. This mechanism is only an interpretation of the authors based on modeling work that may or may not be an accurate account of what was observed in the brain and in behavior. Please modify the title accordingly.*

Response: The reviewer’s point is well taken. We have revised the title and no longer mention attractor networks.

*3. Pg. 25: "The anatomic location of electrode penetrations was determined on the basis of MR imaging." Was this method applied to both the stimulating electrode and the recording electrodes in the prefrontal cortex? It is important to report the locations of all recorded prefrontal neurons included in the analysis, in order to know in which subregions in the lateral prefrontal cortex the recorded neurons came from. This can be done separately for each monkey by showing a figure with the 2D coordinates of the electrode penetrations within the chamber with respect to known anatomical landmarks such as the arcuate and principal sulci. Previous studies by some of the authors (Riley et al., 2018) have shown functional differences between neurons in different prefrontal subregions during working memory tasks. Therefore, knowing the precise recording locations seems important. Also, were the recorded neurons only dorsal to the principal sulcus, or also ventral. If the latter, then please modify the area name from dorsolateral prefrontal cortex (dlPFC) to lateral prefrontal cortex (LPFC).*

Response: This was a very good suggestion. We have now added a figure with anatomical penetrations. Recordings sites were localized in the posterior-dorsal and mid-dorsal subdivisions of the dlPFC, based on the Riley nomenclature, which includes both banks of the principal sulcus.

*4. To maintain consistency with the within-sessions approach suggested in Major Comment #2, I strongly recommend that the authors present Figures 2E-J not merely as an across-sessions mean for on and off conditions, but rather in a form that maintains the pairing of the performance scores within each session. For example, this can be accomplished with a scatterplot where each session is a dot and the x-y coordinates represent performance in the off and on conditions, respectively. A unity line can further help compare performance in off vs. on conditions.*

Response: The reviewer is certainly right that this would have been an ideal presentation but these results were obtained from different sessions with and without stimulation.

*5. The authors should always make a clear distinction, in the Summary and elsewhere on the manuscript, between results obtain from the real electrophysiological or behavioral data, and those obtained from their modeling approach. For example: Pg. 1, Summary: "Tuning of neuronal responses broadened, which rendered an attractor network more stable and filtered distracting visual stimuli more effectively." This statement is confusing, since it is based on the modeling section of the manuscript but is stated as if it were a result from the electrophysiological data.*

Response: The reviewer’s point is well taken. We now mention experimental and modeling results in separate sentences in the Summary, and identify simulations as such.

*6. One of the most important results in the study was that selectivity for spatial locations was decreased by NB stimulation. This was likely unexpected, given that the authors observed improvements in working memory performance. However, it is essential to include this results in the Summary and how it was at odds with performance increases. Currently, the Summary describes results in a way that ignores such decrease in selectivity and instead focuses only on the overall activity increases and on the broadening of neuronal tuning as a result of stimulation.*

Response: The reviewer’s point is well taken. We now strike a more balanced view in the Summary and mention that “, behavioral performance increased for working memory task conditions that benefited from more stable representation of stimuli in working memory and filtering of dissimilar distractors, but declined for conditions that required fine judgments of the spatial location of stimuli”.

*7. "We focused specifically on the prefrontal cortex, an area critical for working memory and cognitive plasticity (Constantinidis and Klingberg, 2016), which receives innervation from a dedicated sub-region of the Nucleus Basalis (Gielow and Zaborszky, 2017)." The latter reference, used by the authors as evidence of innervations from the Nucleus Basalis to the prefrontal cortex, is from a study in rats. There is no evidence from this study that such innervations exist in primates, specifically in macaques. Given known major differences between rodents and primates, especially with regards to the prefrontal cortex, and given the importance of this evidence to this study, it seems essential for the authors to provide any evidence of such innervations in primates from previous studies. This evidence is important to discard alternative explanations to their own results that are mainly based on indirect projections between the Nucelus Basalis and the prefrontal cortex.*

Response: The reviewer’s point is well taken. We now review in more detail the evidence that Nucleus Basalis is topographically organized across species, including in the rat, monkey, and human.

*8. Pg. 5, Results: "NB Stimulation was ineffective when the visual stimulus was ipsilateral to the stimulation (Fig. 2F,H)." Pg. 17, Discussion: "NB stimulation degraded performance when ipsilateral stimuli were held in memory and when distractors were present." The first statement - that stimulation "degraded performance" in the Discussion is inconsistent with the second statement - that stimulation "was ineffective" in the Results. This inconsistency makes the authors appear as if they either didn't quite understand their own results, or they manipulate the description of their results to favor their own interpretations in the Discussion. This could cause the readership to become skeptical of the authors' claims.*

Response: The reviewer’s point is well taken. We now make it clear in both instances that NB stimulation leads to reduced performance.

*9. Figure 5 has unnecessarily too many panels, making it harder to follow. The results can be summarized with less panels. For example, Figures 5A-D are never described the Results section nor anywhere in the main text. The authors should divide the panels thematically into two figures or send some of the secondary panels to the Supplementary Figures.*

Response: The reviewer’s point is well taken. We have split the figure into two, and reduced the panels in other figures as well. The paper now contains 20 total figures. We describe in more details the panels that do remain in the figure in question (now 6).

*10. Pg. 14: "LFP recordings in the NB stimulation period were characterized primarily by an increase in power in the beta frequency range (Fig. 7C, D)." The beta band is typically around 12-25 Hz. None of the frequency bands in Fig. 7D closely corresponds to the beta band. If they intend to specifically test power in the beta band, then they should use a range corresponding to this band. Furthermore, the authors never mentioned that stimulation decreased in the alpha/beta band (8-14 Hz), as seems apparent in figure 7D. They should report this as well.*

Response: The reviewer is certainly right. We tested an a priori hypothesis about modulation of power in different frequency bands as put forward by Lundqvist et al., Neuron, 2016, and referred to the same frequency bands as they defined them (beta as 20-35 Hz and gamma as 45-100 Hz). As the reviewer points out, however, this definition poorly corresponds to established power bands. We now refer to the 20-35 Hz as high-beta/low-gamma power. We do report that stimulation caused a decrease in 8-14 Hz power.

*11. Pg. 12-13: "the neural effects we observed might have been entirely the result of changes in upstream, sensory cortical areas that were not active during the delay period of the task, but which were propagated and maintained in the prefrontal cortex." ... "attractor dynamics impose a fixed bump width in the absence of selective input during the delay period." The authors somehow assume that any inputs into PFC are not selective during the delay. However, spatially-selective signals during the delay period exist in multiple areas other than PFC, and could represent inputs into PFC. Please correct these statements to account for this.*

Response: We agree entirely with the reviewer, and this statement was meant to refer to sensory areas not generating delay period activity (as noted in the quote), which we contrasted with activity of e.g. parietal areas, that project to prefrontal cortex, in the last sentence of the paragraph. Nonetheless, the reviewer’ point is well taken and we have clarified this wording further.

*12. On page 4, the authors should describe to the potentially-broad readership their immunohistochemstry approach (ChAT) in more detail and explain why it confirms electrode placement.*

*13. Figure 2E-J: The figure legend is lacking a description of what the blue and orange color bars represent. This information is only present in one small corner of the figure, but this also needs to be described in the legend itself.*

Response: We have revised accordingly.

*14. The description of the ANOVA results on page 5 is currently quite confusing: it is not always clear which monkey each result and statical comparison refers to or whether they refer to both monkeys combined.*

Response: We have clarified that the 3-way ANOVA results presented were performed separately for each monkey.

*15. Figure 4 legend: Anytime the authors refer to mean firing rate, they should specify what the mean is taken over. If it is across neurons, this needs to be said.*

Response: It is indeed mean across neurons, and we have revised accordingly.

*16. For S6A-B, the method of "plotting the p-value for the main effect of task on firing rate (Fig. S6A-B)" is not an appropriate way to measure effects. In that case, the ANOVA PEV would be best. Also, a visual comparison between S6A and S6B is not accurate or easy to visualize. Instead, consider displaying the difference between A and B as a function of time and neuron number in one single plot.*

Response: The reviewer’s point is well taken. We have eliminated this figure.

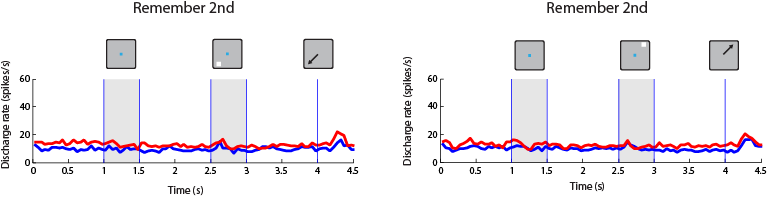
*Reviewer #2:   
This important and novel study electrically stimulates the nucleus basalis (NB) in monkeys to determine how activation of ascending cholinergic pathways changes the properties of prefrontal neurons and behavior. The authors show that NB stimulation improves performance on an oculomotor working memory task. The authors then elegantly combine neurophysiology and artificial neural network modeling to understand how changes in the coding properties of prefrontal neurons may lead to this behavioral improvement. They show a remarkable correspondence between the dynamics of a ring attractor network and neural signals in primate PFC. The authors show that stimulating NB broadens the spatial tuning curves of PFC neurons in the neural data. Simulating this change in the spatial tuning of neurons in the ring attractor network replicates an impressive array of behavioral findings observed following NB stimulation in monkeys. This includes an increase in errors when distractor stimuli are close to but not far from the remembered target, differences between remember 1st and remember 2nd tasks, as well as more fine-grained observations regarding a decrease in the variability of saccadic endpoints during NB stimulation. Integration of modeling and neural recording data is a very strong aspect of the paper, providing unique insight into the neurophysiology. The close match in dynamics between model and experiment suggest the connectional assumptions of the model reflect connectional motifs in PFC. The data establish a solid foundation for future studies investigating the neural and computational effects of NB stimulation as a potential therapeutic to improve WM performance and PFC function in human disease. The data also provide new information about how ACh acts as a neuromodulator to tune PFC circuits. Finally, in contrast to pharmacological studies that have administered cholinergic agents iontophoretically or systemically, this study characterizes the impact of modulating acetylcholine by activating ascending cholinergic pathways. Consequently, it is well grounded in the anatomy and provides information about the behavioral, neural, and computational functions of this ascending cholinergic circuit.   
  
Major Comments   
  
1. Most of the variability in performance across conditions appears to be present in the baseline (stim off) condition. Performance with stimulation on is relatively consistent across animals and conditions, about 85-95% correct in most cases (Fig. 2E, F; G, H: orange bars). There is more variability in the level of behavioral performance in the stim off control condition (blue bars). For example, in the Remember 1st task, Monkey GR is worse for ipsi than contra targets at baseline with stim off (Fig. 2E, F). In the Remember 2nd task, both animals seem better for ipsi than contra targets at baseline (Fig. 2G, H). Variation in levels of baseline performance across conditions could reflect spatial biases of the animals (which is common), and it is true that the important functional point is the relative improvement in performance with stim on across conditions, which seems consistent. But if the relative improvement of stim is driven primarily by a shift in the baseline (control condition), with stim on performance reaching a ceiling or plateau from these variable levels of baseline performance, this bears on functional interpretation of the stim effects. Stim could, for example, improve attention to eliminate spatial biases present at baseline. To address the point it could be useful to divide the data into stim off and stim on subsets, and test whether baseline performance differed significantly across conditions (ipsi/contra, remember 1st/2nd) within each subset. If performance differs across conditions in the baseline but not the stim on conditions, it would suggest that stim counteracted behavioral biases and drove performance to ceiling.*

*2. The main effect of NB stimulation was to broaden the spatial tuning curves of neurons with cue or delay period activity. This was associated with an improvement in performance both in behavior and the model. The model attributes this to the dynamics of the ring attractor network, and by extension to prefrontal local circuits that may have similar connectivity. However, a more direct question is whether NB stimulation improves or degrades the population representation of direction at the neural level. The flattening of the tuning curves suggests that NB stim reduces the strength of the direction signal in the population, but reduction in trial-to-trial variability in neuronal responses suggests that NB stim lowers neural noise, and it is not clear which of the two changes would dominate. One way to quantify this would be to decode direction (one out of eight) from population activity in PFC, to quantify changes in either percent correct decoding or the mean posterior probability associated with the correct direction with and without NB stim.*

Response: We now present results that indicate decreased information decoding from the population of prefrontal neurons (see Reviewer 1 – comment 1).

*3. The model replicates 'phantom' responses to expected but nonexistent visual stimuli using a nonselective timing signal (increasing input around the time of the stimuli but is not tuned). Did such signals exist in the neural data, e.g. were there neurons that increased firing rate at the anticipated times of the stim 1 and 2 but that were not spatially selective?*

Response: This was a very interesting question. A non-selective rise in activity starts to become evident in the population of neurons that would eventually exhibit selectivity after the fixation point turned on but before the stimulus appeared (e.g. see Fig. 5A, second panel) and then accelerates at the time of the anticipated stimulus appearance. Such a non-selective signal was largely absent from the neurons with no stimulus selectivity (see Figure X1). It appears that this signal serves a purpose for those neurons that have a role in processing visual stimuli, but do we do not have a large enough sample of non-selective neurons to investigate further.



*4. The mechanism by which NB stim is proposed to give rise to phantom bumps is that stimulation makes the baseline more unstable. If this is correct, the bump should emerge at random locations in the ring encoding random directions from trial to trial (since noise in the ring drives the emergence of the bump). Was this the case in the neural data and also the activity of the artificial network?*

Response: This was a very astute observation. The emergence of the bump in the network was random, and any neuron was activated at only a small percentage of trials (this is why the absolute value of the y axis in Supplementary Figure S10C-D is quite low). We now explain that. It is more difficult to answer this question for the neural data because we did not have a large number of simultaneous recordings that would allow us to localize the position of the bump in each trial. Furthermore, the location of the first stimulus in the behavioral task was not entirely random; the first stimulus could only appear at one of two diametric locations. We saw similar activity of neurons representing both possible locations (Supplementary Fig. S3E and J), so in that sense at least, the emergence of the bump appeared to be random in neural activity as well.

*5. The artificial neural network and the PFC appeared to implement task-selectivity by different means. PFC neurons exhibited task-selective activity primarily during the 1st stimulus period (Fig. S6A and B). Task selectivity in the model was achieved by modifying connection strengths between neural populations and varying the level of excitatory input across the entire trial between tasks. It would be useful to consider these differences and how they might constrain interpretation of the current results.*

*6. Prior studies (Lundqvist et al) have reported that gamma oscillations related to working memory take the form of transient bouts ('gamma bubbles') that reflect stochastic internal network dynamics and are not aligned to task events and therefore disappear in trial-averaged data. The present LFP power analysis was done on trial averaged data. The authors may wish to note this difference in analytical approach when interpreting the influence of DB stim on gamma synchrony.*

Response: The reviewer’s point is well taken. We absolutely intend to replicate the Lundqvist et al. analysis in our dataset, but it is beyond the scope of this paper. We have rephrased the text to more accurately reflect the difference in our approach and theirs.

*Minor Comments:   
Page 5: "For monkey HE, stimulation improved performance specifically when the visual stimulus to be remembered was at a location contralateral to the site of the stimulation electrode (Fig. 2E-H)." Should this refer to (Fig. 2E, G)?*

Response: We have corrected as suggested.

*Page 8: "These results of NB stimulation moved in the same direction for both remember-first and remember-second tasks (Fig. S4A), and for both monkeys (Fig. S4B-C)." These figure citations seem incorrect as well.*

Response: The figure panel citations were as intended, but we have expanded to explain better what we referred to in each case: “These results of NB stimulation moved in the same direction for both remember-first and remember-second tasks (Fig. S5A and Fig. 4A, respectively), and for both monkeys (Fig. S6).”

*Page 10: "A broader bump leads to a more stable bump attractor, less sensitive to noise fluctuations in the network (Fig. 5E,F,G,H), thus leading to more accurate read-outs at the end of the delay (Fig. 5J,K,M,N)." Similarly seemed these figure panel citations should be checked.*

Response: We agree that the panels in question do not obviously make the point raised in the sentence. We have removed the panel citations.

*Page 15 (Discussion): It was not clear to me whether or how the following statement was supported by the data: "... and amplifying anticipatory responses, prior to the appearance of the visual stimulus."*Response: This was a reference to the “phantom bump” finding. We have rephrased for clarity.

*Methods: Neural waveforms were classified as fast spiking (FS) or regular spiking (RS) based on waveform, but this distinction does not seem to be used in the analysis or results. Was there a relation between neuronal type and NB stimulation effects? That is, were fast-spiking neurons preferentially suppressed, and regular spiking enhanced?*

Response: Yes, we analyzed FS and RS neurons separately, but found no systematic difference. We now reference this finding in the text and show the results in Supplementary Figure 12.  *Additional data files and statistical comments: Statistical analyses are rigorous and well described. Sufficient information is provided to evaluate the analyses.*

Response: Thank you.