Mechanisms of distributed working memory in a large-scale model of macaque neocortex

Jorge F. Mejias[1], Xiao-Jing Wang[2].

[1]Swammerdam Institute for Life Sciences, University of Amsterdam, 1098XH Amsterdam, the Netherlands.

[2]Center for Neural Science, New York University, 4 Washington Place, New York, NY 10003, USA.

Contact: j.f.mejias@uva.nl, xjwang@nyu.edu.

Recent evidence suggests that persistent neural activity underlying working memory is not a local phenomenon but distributed across multiple brain regions. To elucidate the circuit mechanism of such distributed activity, we developed an anatomically constrained computational model of large-scale macaque cortex. We found that inter-areal reverberation can support the emergence of persistent activity, even when the model operates in a regime where none of the isolated areas is capable of generating persistent activity. Persistent activity exhibits a gap of firing rate at particular cortical areas, indicating a robust bifurcation in space. Model analysis uncovered a host of spatially distinct attractor states, and yielded novel experimentally testable predictions. In the model, distributed activity patterns are resilient against simultaneous inactivation of multiple cortical areas, while dependent on a structural core. This work provides a theoretical framework for identifying large-scale brain mechanisms and computational principles of distributed cognitive processes.

Introduction

With the advances of brain connectomics and physiological recording technologies like Neuropixels[1,2], a major new challenge in Neuroscience is to investigate biological mechanisms and computational principles of cognitive functions that engage many interacting brain regions. Here, the goal is no longer to identify one local parcellated brain region that contributes to or is crucial for a particular function, but how a large-scale brain system with many interacting parts underlie behavior. Currently, there is a dearth of theoretical ideas and established models for understanding distributed brain dynamics and function.

A basic cognitive function which has recently been shown to involve multiple brain areas is working memory, the brain's ability to retain and manipulate information in the absence of external inputs. Working memory has been traditionally associated with persistent neural firing in localized brain areas, such as those in the frontal cortex[3–8] and computational models uncovered the involvement of local recurrent connections and NMDA receptors in the encoding of memory items via selective persistent activity[9–12]. However, self-sustained neural persistent activity during working memory has been found in multiple brain regions; and often such highly engaged areas appear in coactivation[13–16]. These modeling efforts have been limited to exploring

either the emergence of persistent activity in local circuits, or in two interacting areas at most[17,18]. It is presently not known what biophysical mechanisms could underlie a distributed form of memory-related persistent activity in a large-scale cortex.

In this study, we tackle this challenge by building and analyzing an anatomically-constrained computational model of the cortical network of macaque monkey, and investigate a novel scenario in which long-range cortical interactions support distributed activity patterns during working memory. The anatomical data is used to constrain the model at the level of long-range connections but also at the level of local circuit connectivity. In particular, the model incorporated differences between individual cortical areas, by virtue of macroscopic gradients of local circuit properties in the large-scale network. The emerging distributed patterns of persistent activity involve many areas across the cortex, that engage temporal, frontal and parietal areas but not early sensory areas, in agreement with recent evidence[15]. Persistent firing rates of cortical areas across the hierarchy display a gap, indicative of the existence of a robust bifurcation in cortical space. Furthermore, the distributed patterns emerge even when individual areas are unable to maintain stable representations, and increase the robustness of the network to distractors and simulated inactivation of areas. The concept departs from the classical view of working memory based on local attractors, and sheds new light into recent evidence on distributed activity during cognitive functions.

Results

Our computational model includes 30 areas distributed across all four neocortical lobes (Fig. 1a; see Online Methods for further details). The inter-areal connectivity is based on quantitative connectomic data from tract-tracing studies of the macaque monkey[19–21] (Extended Data Fig. 1). For simplicity, each of the cortical areas is modeled as a neural circuit which contains two selective excitatory populations and one inhibitory population[9,22] (Fig. 1b). In addition, the model assumes a macroscopic gradient of synaptic excitation[23–25] (Extended Data Fig. 2). This gradient is introduced by considering that the number of apical dendritic spines, loci of excitatory synapses, per pyramidal cells increases[26] along the cortical hierarchy[20,27] (Fig. 1c). This gradient of area-specific connection strength was applied to both local recurrent and long-range excitatory connections. In particular, we denote the maximal strength of local and long-range excitation for the area at the top of the cortical hierarchy by J_{max}, which is an important parameter of the model (see below). To allow for the propagation of activity from sensory to association areas, inter-areal long-distance connections target more strongly excitatory neurons than inhibitory neurons for feedforward pathways, but biased in the opposite direction for feedback pathways, in a graded fashion[28] (Fig. 1b).

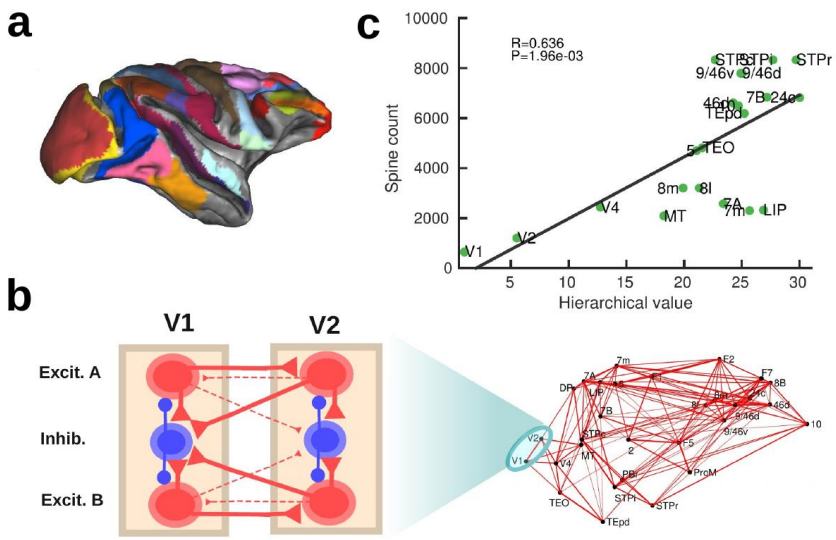


Figure 1: Scheme and anatomical basis of a multi-regional macaque neocortex model. (a) Lateral view of the macaque cortical surface with areas in color. (b) In the model, inter-areal connections are calibrated by mesoscopic connectomic data[19], each parcellated area is modeled by a population firing rate description with two selective excitatory neural pools and an inhibitory neural pool[29]. Recurrent excitation within each selective pool is not shown for the sake of clarity of the figure. (c) Correlation between spine count data[26] and anatomical hierarchy as defined by layer-dependent connections[20].

Distributed WM is sustained by long-range cortical loops

In local circuit models of working memory (WM)[10,11], areas high in the cortical hierarchy make use of sufficiently strong synaptic connections (notably involving NMDA receptors[12]) to generate self-sustained persistent activity. Specifically, the strength of local synaptic reverberation must exceed a threshold level (in our model, the local coupling parameter J_S must be larger than a critical value of 0.4655), for an isolated local area to produce stimulus-selective persistent activity states that coexist with a resting state of spontaneous activity (operating in a multistable regime rather than in a monostable regime, see Fig. 2a). However, there is presently no conclusive experimental demonstration that an isolated cortical area like dorsolateral prefrontal cortex (dlPFC) is indeed capable of generating mnemonic persistent activity. In this study, we first examined the scenario in which all areas, including dlPFC (9/46d) at the top of the hierarchy, have J_S values below the critical value for multistability (so $J_S \leq J_{max} < 0.4655$) and are connected via excitatory long-range projections of global coupling strength G (we set J_{max} =0.42 and G=0.48 unless specified otherwise)(Fig. 2a). In this case, any observed persistent activity pattern must result from inter-areal connection loops. In a model simulation of a visual delayed response task, a transient visual input excites a selective neural pool in the primary visual cortex (V1), which yielded activation of other visual areas such as MT during stimulus presentation (Fig. 2b, upper left). After stimulus withdrawal, neural activity persists in multiple areas across frontal, temporal and parietal lobes (Fig. 2b, lower right). This activation pattern was stimulus specific, so only the neural pool selective for the shown stimulus in each cortical area displayed elevated persistent activity (Fig. 2c; Extended Data Fig. 3).

Given that selective persistent activity is also found in somatosensory WM tasks[6], we further test our model and simulate a simple somatosensory WM task by transiently and selectively stimulating a neural pool in primary somatosensory cortex. As in the case of visual stimulation, this leads to the emergence of a distributed persistent activity pattern of equal selectivity as the input (Extended Data Fig. 4). This shows that the distributed working memory mechanism is valid across different sensory modalities.

For simplicity purposes, and to explore the minimal ingredients of a distributed WM mechanisms, our model has considered NMDA receptors as the only excitatory dynamics. However, AMPA dynamics may also be of importance[30], and can be easily introduced leading to a good behavior of the model for shorter durations of the stimulus presentation (Extended Data Fig. 5). Across all these simulation results, we observed cross-area variations of neural dynamics: while areas like TEpd displayed a sharp binary jump of activity, areas like LIP exhibited a more gradual ramping activity, resembling temporal accumulation of information in decision-making[31].

Robust bifurcation in space and cortical gradients

Consistent with experimental observations[15], early sensory areas such as V1 and MT did not display persistent activity. In our model this depends on two factors. First, local synaptic coupling of early sensory areas at the bottom of the hierarchy are weak, as indicated by the fact that dendritic spine counts are one order of magnitude smaller in those areas than in prefrontal areas[26] (Extended Data Fig. 2, Extended Data Table 1). Second, a certain level of preferential targeting of inhibitory neurons by top-down projections helps to prevent indiscriminate persistent activation across all cortical areas[15]. As Fig. 2d shows, introducing a small bias of feedforward projections towards excitatory neurons and feedback projections towards inhibitory neurons reduces the overall mean firing rate of early sensory areas within a distributed attractor state, while maintaining the activity levels of areas high in the hierarchy at appropriate values. Note that, following previous work[28], the preferential feedback targeting is introduced as proportional to the fraction of laminar projections[21], and therefore it results in a very small bias for most of the projections (see Online Methods for further details). Preferential targeting of feedback projections could explain observed top-down inhibitory control effects[32] and it also allows the distributed WM patterns to exist for a wide range of global coupling strength G for long-range synaptic strengths (Fig. 2e).

When we plotted the firing rate of stimulus-selective persistent activity across 30 areas along the hierarchy, our results revealed a gap that separated the areas displaying persistent activity and those that did not (Fig. 2f). This is a novel type of bifurcation or abrupt transition of behavior that takes place in space, rather than as a function of a network parameter like in Fig. 2a. As a matter of fact, the relevant parameter here is the strength of synaptic excitation that varies across cortical space, in the form of a macroscopic gradient. The bifurcation is robust to different considerations, such as different arrangement of areas (Extended Data Fig. 6) and varying the maximum strength of inhibition-targeting projections between frontal areas (see Online Methods). For example, by imposing a maximum strength (C_{FEF}) of projections targeting FEF inhibitory populations from other frontal areas, we see in Fig. 2f that the bifurcation holds for a wide range of C_{FEF} values, and the effect is a minor alteration only for the exact areas at the

bifurcation threshold. Therefore, the bifurcation is robust and does not require precisely tuning a parameter to a threshold value.

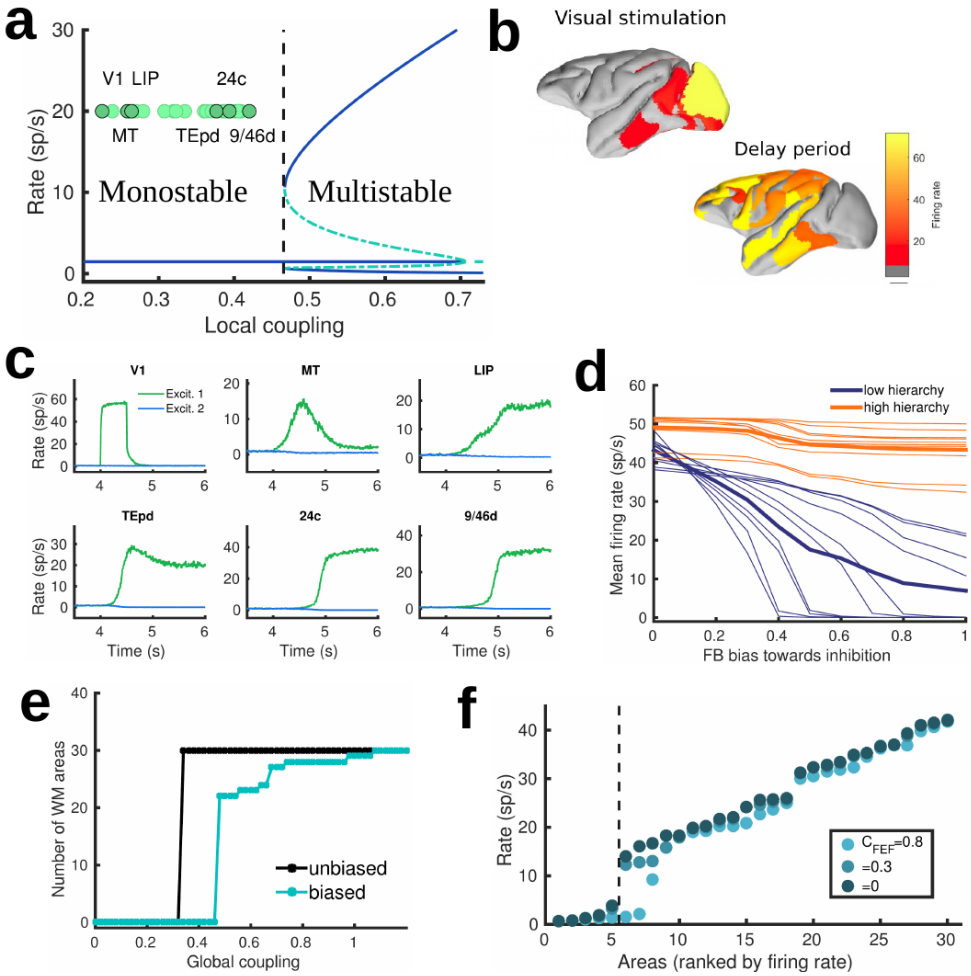


Figure 2: Distributed WM sustained via long-range loops in cortical networks. (a) Bifurcation diagram for an isolated area. Here, all areas (green circles) are in the monostable regime when isolated. (b) Spatial activity map during visual stimulation (upper left) and delay period (lower right). (c) Activity of selected cortical areas during the WM task, with a selective visual input of 500 ms duration. (d) Firing rate of areas at the bottom and top of the hierarchy (10 areas each, thick lines denote averages) as a function of the feedback preference to inhibitory neurons. (e) Number of areas showing persistent activity in the example distributed activity pattern vs global coupling strength with (turquoise) and without (black) the assumption of feedback biased towards inhibition. (f) Firing rate for all areas ranked by firing rate and G=0.52. Colors denote different maximum strengths of inhibition to frontal eye field areas 8l and 8m from all other frontal areas.

While the macroscopic gradient of excitability is an important property of the model, the particular values of excitatory strength assigned to each area are not relevant for the phenomenon

(Extended Data Fig. 7a). However, the distributed working memory patterns are markedly affected when the anatomical structure of the cortical network is changed –for example, by randomly shuffling the projection strength values (Extended Data Fig. 7b). This suggests that salient statistical features of the structure embedded in the cortical network play a role in the emergence of distributed activity patterns. The model also predicts the emergence of a hierarchy of time scales across cortical areas (Extended Data Fig. 8), in agreement with experimental findings[33] and supporting and improving previous computational descriptions[23].

Long-range cortical loops support a large number of different distributed attractors

We realized that a large-scale circuit can potentially display a large number of distributed persistent activity patterns (attractors), and some of them may not be accessible by stimulation of a primary sensory area. Note that distinct attractor states are defined here in terms of their spatial patterns, which does not depend on the number of selective excitatory neural pools per area. We developed a numerical approach to identify and count distinct attractors (see Online Methods for further details). Our aim is not to exhaustively identify all possible attractors, as the activity space is too large, but to gain insight on how our estimations depend on relevant parameters such as the global coupling strength G, or the maximum area-specific synaptic strength J_{max}. Five examples of different distributed WM attractors are shown in Fig. 3a, where we can appreciate that not all distributed attractors engage cortical areas at all lobes, and that frontal areas are the ones more commonly involved.

A more detailed analysis included four cases depending on the value of the maximum area-specific synaptic strength J_{max} assumed: two of the cases had J_{max} above the bifurcation threshold for isolated areas (0.4655), and the other two had J_{max} below the bifurcation threshold. For the first two cases, having $J_{max}>0.4655$ means that at least certain areas high in the hierarchy, such as dlPFC, have strong enough local reverberation to sustain activity independently (i.e. they were 'intrinsically multistable' and able to display bistability even when isolated from the network, Fig. 3b), however, areas lower in the hierarchy like TEpd and LIP would require long-range support to participate in WM. For the last two cases, in which $J_{max}<0.4655$, none of the areas was able to display bistability when isolated, but they can contribute to stabilize distributed WM attractors as in Fig. 2b, c. In all four cases, the number of attractors turns out to be an inverted-U function of the global coupling strength G, with an optimal G value maximizing the number of attractors (Fig. 3c, curves are normalized to have a peak height of one for visualization purposes). This reflects the fact that a minimal global coupling is needed for areas to coordinate and form distributed WM attractors, but for values of G too large, all areas will follow the majority rule and the diversity and number of possible attractors will decrease. The optimal G value shifted towards lower values for increasing J_{max}, and the peak number of attractors simultaneously increasing (Fig. 3d).

Across all four cases and G values considered, we found a significant positive correlation between the number of areas involved in a given attractor and the average firing rate of these areas (Fig. 3e), which constitutes an experimentally testable prediction of the distributed model of WM. We also analyzed how distributed WM attractors were constituted for the four different cases (Fig. 3f). When the network has a high number of intrinsically multistable areas (i.e. when $J_{max}>0.4655$), attractors tend to only involve these areas and are therefore largely restricted to the areas located at the top of the hierarchy (Fig. 3f, bottom left and right panels). On the other hand,

when the network has zero or a low number of intrinsically multistable areas (i.e. J_{max}<0.4655), attractors typically involve a larger number of areas (as a larger pool of areas is needed to sustain distributed WM attractors, see top right panel in Fig. 3f) and the areas involved are more diverse in their composition (Fig. 3f, top left panel).

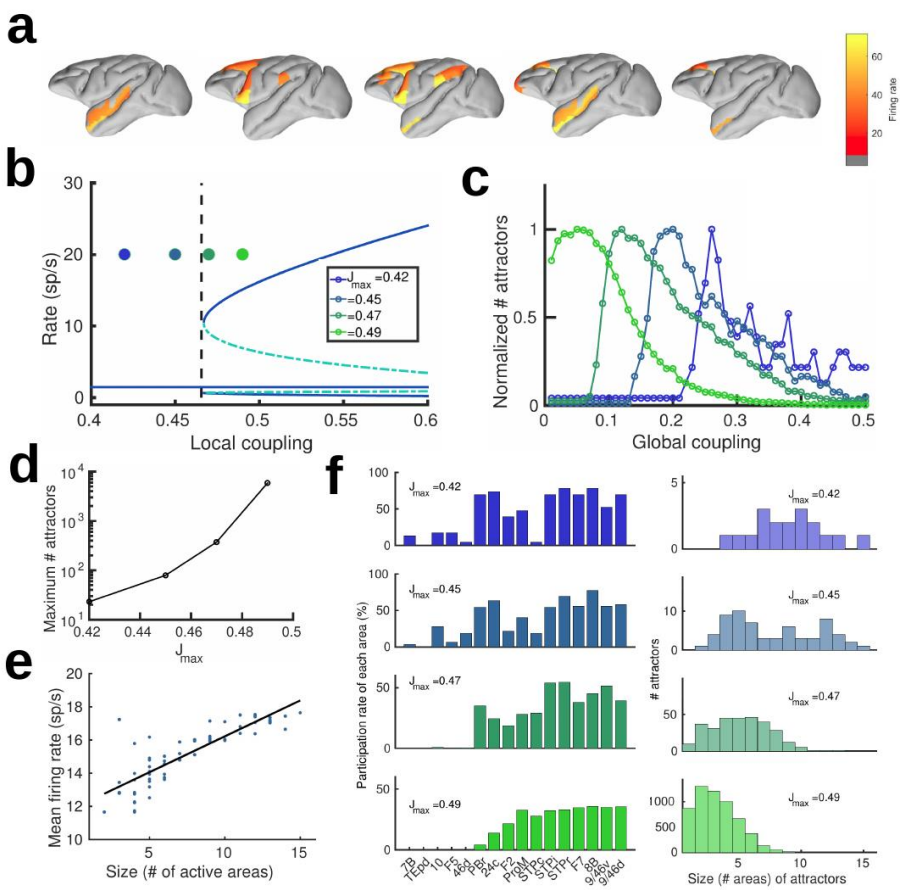


Figure 3: Distributed and local WM mechanisms can coexist in the model. (a) Five example attractors of the network (J_{max}=0.42). (b) Bifurcation diagram of an isolated local area with the four cases considered. (c) Number of attractors (normalized to one) found via numerical exploration as a function of the global coupling for all four cases. (d) Maximum (peak) number of attractors for each one of the cases. (e) Correlation between size of attractors and mean firing rate of its constituting areas for J_{max}=0.45 and G=0.2. (f) Participation index of each area (left, arranged by spine count) and distribution of attractors according to their size (right).

<u>Distributed WM is robust to distractors and deactivates with frontal inhibition</u>

The distributed nature of WM has also implications for the impact of distractors and similar sensory perturbations on maintenance of selective activity and overall performance of the network. We simulated a delayed response task with distractors (Fig. 4a), in which stimulus A is cued to be maintained in WM and stimulus B is presented as distractor during the delay period (and vice versa). We found that only strong distractors (compared to the original visual cue) are able to destabilize the memory (Fig. 4b). This is due to the robustness of a distributed attractor

compared to a local circuit mechanism, but also to the inhibitory effect of feedback projections which dampen the propagation of distractor signals (cf. responses in V1 and MT in Fig. 4b). This constitutes a fundamental difference with respect to local WM models, where distractors of the same strength as cues are able to destabilize memories[10,11].

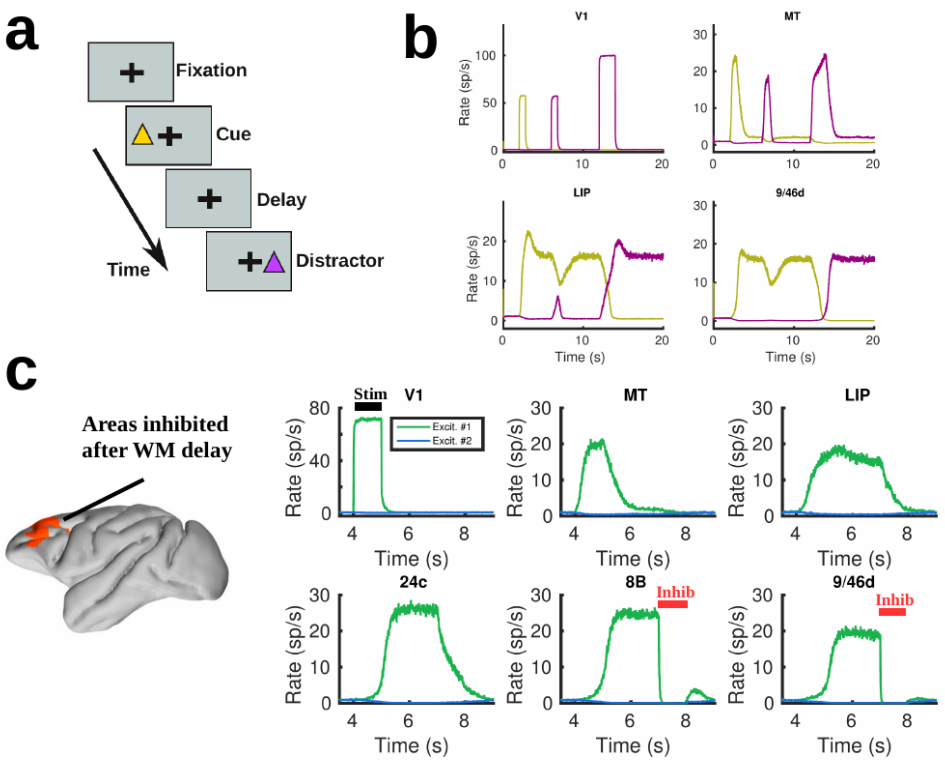


Figure 4: Resistance to distractors and shutting down persistent activity in the network. (a) Scheme of the WM task with distractors. (b) Activity traces of selected areas (color code denotes input selectivity) in response to stimuli of 500 ms duration and with G=0.26. (c) A distributed attractor is shut down by inhibiting a selective group of frontal areas (9/46v, 9/46d, F7, and 8B). The shut-down input, of strength I=0.3 and 1s duration, is provided to the nonselective inhibitory population of each of these four areas.

An important aspect of a WM model is how to shut down the persistent activity once the task has been done. In many traditional WM models, this is achieved by providing a strong inhibitory input to the whole WM circuit, which drives the activity of the selective populations back to their spontaneous state[9–11,34]. It is, however, unrealistic to expect that this approach could also be used for shutting down distributed WM patterns, as it would require a large-scale synchronous inhibitory pulse to all active areas.

We therefore explore in our model whether more spatially selective signals can shut down distributed patterns of activity. In spite of their robustness to sensory distractors as discussed above, we find that distributed WM activity patterns can be shut down with an excitatory input targeting inhibitory populations of areas high in the hierarchy. Fig. 4c shows how a visually evoked distributed WM attractor is deactivated when we deliver excitatory input to the inhibitory

populations in the top four areas of the hierarchy (9/46v, 9/46d, F7 and 8B). These prefrontal areas are spatially localized and thought to be highly engaged in WM maintenance, and therefore they are suitable candidates to control the suppression of persistent activity in other cortical areas, such as areas LIP and 24c as shown in the figure. Therefore, and in spite of engaging cortical areas across all four lobes, distributed WM attractors can be controlled and deactivated by localized inhibition to a small set of frontal areas.

Cortical inactivations have differential effects on the disruption of distributed WM

Finally, we tested the effect of inactivating cortical areas in our model during WM tasks, which can be done experimentally using optogenetic methods or otherwise by lesioning selected areas. We tested this by completely and permanently suppressing the firing rate of the inactivated area in the model, in such a way that the area becomes a sink of current and does not communicate with other areas. We found that inactivations of most individual areas have only a local effect on the overall maintenance of the distributed WM attractor, even when inactivating areas relatively high in the hierarchy such as the case of area 24c (anterior cingulate cortex). Indeed, both the spatial activity distribution (Fig. 5a, 3D brain plots on the left) and the temporal activity traces of non-silenced areas (Fig. 5a, activity traces on the right) remain largely unaffected by such inactivations. However, inactivating key prefrontal cortical areas such as 9/46d (dlPFC) can still have a strong impact on working memory maintenance for moderate values of the maximum area-specific synaptic strength J_{max} (Extended Data Fig. 9), which is in agreement with classical prefrontal lesion studies[35]. This key role of dlPFC and similar prefrontal areas also aligns well with the strong role of these areas in shutting down distributed WM patterns as explained above (Fig. 4c).

In a similar way, we studied the case of inactivating a given number of randomly selected areas in the model. We found that the number of active areas in distributed WM attractors decreases only linearly with the number of inactivated areas (Fig. 5b). When instead of selecting inactivated areas at random we silence areas in reverse hierarchical order (i.e. by inactivating top hierarchical areas first), the number of active areas decreases a bit more abruptly (Fig. 5c) and, as seen above and in Extended Data Fig. 9, can prevent the emergence of distributed WM altogether if J_{max} is sufficiently small. Inactivating a number of random areas decreases the overall firing rate of the remaining areas when participating in distributed WM attractors (Extended Data Fig. 10a).

We also carried out a more systematic evaluation of the effect of cortical inactivations, including their effect on attractors that were not accessible from sensory stimulation directly. This study revealed that inactivating most areas has only limited consequences on the total number of available distributed attractors. The exceptions were several temporal and prefrontal areas for which the overall impact is large and sometimes more than half of the initially available attractors were lost (Fig. 5d). The impact of the inactivation on the number of available attractors increases with the location of the silenced area in the hierarchy (Extended Data Fig. 10b). Interestingly, and beyond this hierarchical dependence, the temporal and prefrontal areas that had the strongest impact are part of the anatomical 'bowtie hub' of the macaque cortex identified in anatomical studies[19] (Fig. 5e). Overall, silencing areas at the center of the bowtie has a deeper impact than silencing areas on the sides of the bowtie (Extended Data Fig. 10c).

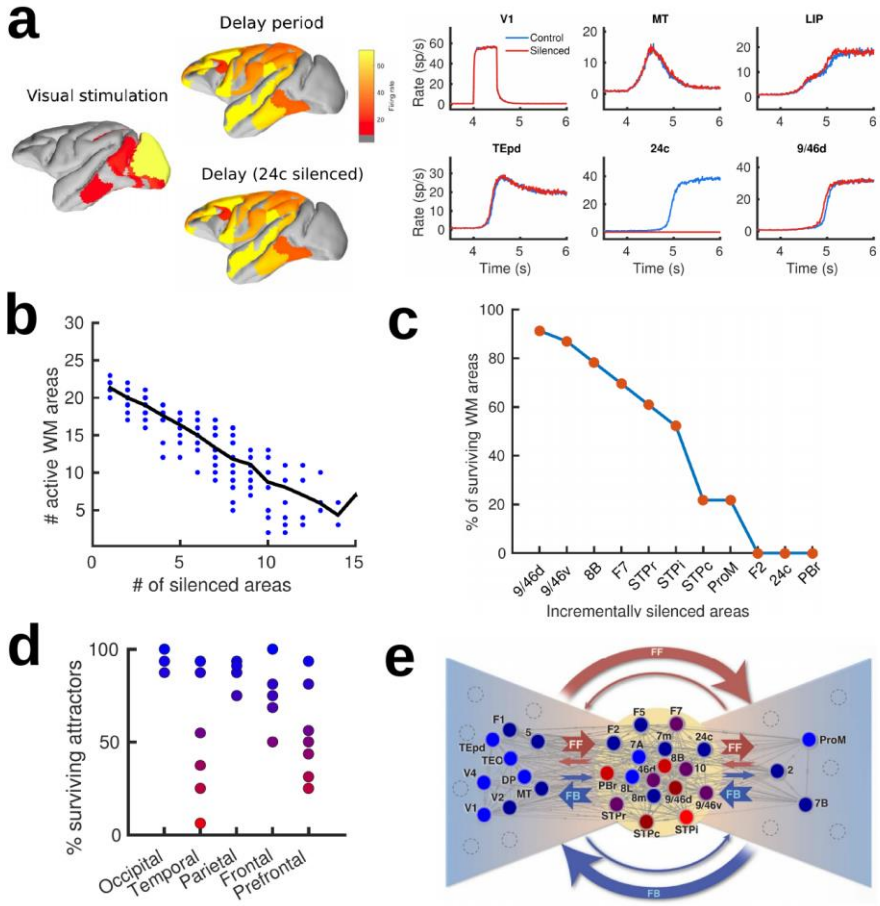


Figure 5: Effects of cortical inactivations in distributed WM patterns. (a) Spatial activity maps during stimulation and delay periods for control (top) and the case of 24c inactivation (bottom). Traces on the right show the impact for selected areas. (b) Number of active areas in the example attractor as a function of the number of (randomly selected) silenced areas. (c) The number of active WM areas decreases faster when areas are incrementally and simultaneously silenced in inverse hierarchical order. (d) Numerical exploration of the percentage of surviving attractors for area inactivations in different lobes. (e) Silencing areas at the center of the 'bowtie hub' have a stronger impact on WM (adapted from[19]). For panels (e) and (f), areas color-coded in blue (respectively, red) have the least (respectively, most) impact when silenced.

Discussion

The investigation of cognitive functions has been traditionally restricted to operations in local brain circuits –mostly due to the limitations on available precision recording techniques to local brain regions, a problem that recent developments are starting to overcome[1,2]. It is therefore imperative to advance in the study of distributed cognition using computational models as well, to support experimental advances. In this work, we have presented a large-scale circuit mechanism of distributed working memory, realized by virtue of a new concept of robust bifurcation in space. The distributed WM scenario is compatible with recent observations of

multiple cortical areas participating in WM tasks[13–16], even when some of these areas have not been traditionally associated with WM.

One of the main ingredients of the model is the gradient of excitation across the cortical hierarchy, implemented via an increase of excitatory recurrent connections (as hinted by the existing anatomical evidence on dendritic spines on pyramidal cells across multiple cortical areas[26]) and also by the preferential targeting of feedback projections to inhibitory cells. The importance of these macroscopic gradients and hierarchical structures has received recent attention[25,36]. It has been shown that gradients of circuit properties in line with hierarchical structures contribute to the emergence of a gradient of time scales across cortex, supporting slow dynamics in prefrontal areas[23,25,33] (see also Extended Data Fig. 8), and also that a hierarchical organization of functional states could serve as basis for WM-guided decisions and executive control[37,38]. It is possible that structural gradients would play a role not only in other cognitive functions in monkeys, but also in other animals including mice[39] and humans[40,41].

Interestingly, the model uncovered a host of distinct persistent activity attractor states, defined by their spatial distributed patterns in the large-scale cortical system, independent of the number of selective neural pools per area (Fig. 3a). Many of these persistent activity states are not produced by stimulation of primary sensory areas. It is possible that these attractors can be trigged by sensory input in computational models with more detailed descriptions of sensory features encoded in temporal or population codes. Nevertheless, these "hidden" attractor states could serve various forms of internal representations such as those that are not triggered by a particular sensory pathway –or those triggered by sensory input but are encoded differently as memories[42]. The identification of these internal representations in further detail are beyond the scope of the present paper, but uncovering their functional role should be within reach of additional experimental and computational work.

Extending the model of distributed working memory

The reported large-scale cortical network model is the first of its kind addressing a cardinal cognitive function, and it opens the door for elucidating this and similar complex brain processes in future research. Several avenues may be taken to extend the functionality of the present model. First, it is straightforward to have an arbitrary number of selective neural pools per area[11], which would increase both the selectivity to sensory inputs and the available number of distributed WM attractors. In that case, more complex connections (not necessarily A to A, B to B, etc.) can be investigated, including a distance-dependent "ring structure"[9,10] or random connections[43]. Second, the model presented here is limited to 30 cortical areas, and can be expanded to include both additional cortical areas and subcortical structures relevant for working memory (such as thalamic nuclei[44]) as their connectivity data become available. Third, the model can be improved by incorporating more biological details such as cortical layers[28], contributions of different neuromodulators, and various types of inhibitory neurons.

Relationship with other mechanisms of working memory

In the description adopted here, we have considered that working memory was maintained via selective persistent activity underlying attractor dynamics, as in most traditional WM approaches[9,34]. However, the proposed principle for distributed WM, and in particular the

hypothesis that WM arises due to cooperation between spatially distributed brain areas via long-range recurrent interactions, is also compatible with other mechanisms[45]. For example, recent work suggests that cortex can generate stable WM representations while relying on time-varying, rather than persistent, firing rates[46–48]. There is no reason to think that the encoding of memory items could not use the complex spatiotemporal interactions between brain areas instead of just local interactions. Likewise, silent WM[48–50], which relies on short-term synaptic reinforcements to sustain information in WM, could also be implemented in a large-scale, distributed scenario. A large-scale implementation of WM also pairs well with recent hypotheses in which memory selectivity is reached via dynamical flexibility instead of content-based attractors, since the wide number and heterogeneity of long-range projections would reduce connection overlap and alleviate the limit capacity of these models[43]. Frameworks of WM in which oscillations play an active role, for example regarding WM-guided executive control[38], may benefit from using distributed WM approaches, given the usefulness of previous large-scale brain models in explaining oscillatory phenomena in the macaque brain[28].

Experimental predictions provided by our model

The distributed WM model presented here yields several experimentally testable predictions in monkey (and potentially rodent) experiments, including (i) a reversed gradient of inhibition; (ii) strong large-scale interactions to sustain distributed WM patterns; (iii) a positive correlation between the number of areas involved in a WM task and their average firing rate of persistent activity (Fig. 3e); and (iv) robustness of working memory encoding against distractors via inter-areal inhibitory feedback (Fig. 4a, b). This model and its extensions can serve as a computational platform to elucidate complex experimental observations, such as optogenetic manipulation across cortical areas in behaving animals performing a working memory task. It will also be interesting to see whether the same model is able to account for decision-making processes as well as working memory[22,29].

Conceptually, this work revealed a novel form of bifurcation in space as a mechanism to generate differential functions across different cortical areas, a concept that is likely to be generalizable for understanding how distinct cortical areas endowed with a canonical circuit organization are at the same time suited for differential functions[25].

References

1. Jun, J. J. *et al.* Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551**, 232–236 (2017).
2. Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364**, (2019).
3. Fuster, J. M. Unit activity in prefrontal cortex during delayed-response performance: neuronal correlates of transient memory. *J. Neurophysiol.* **36**, 61–78 (1973).
4. Funahashi, S., Bruce, C. J. & Goldman-Rakic, P. S. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J. Neurophysiol.* **61**, 331–349 (1989).

5. Goldman-Rakic, P. S. Cellular basis of working memory. *Neuron* **14**, 477–485 (1995).

6. Romo, R., Brody, C. D., Hernández, A. & Lemus, L. Neuronal correlates of parametric working memory in the prefrontal cortex. *Nature* **399**, 470–473 (1999).

7. Kopec, C. D., Erlich, J. C., Brunton, B. W., Deisseroth, K. & Brody, C. D. Cortical and Subcortical Contributions to Short-Term Memory for Orienting Movements. *Neuron* **88**, 367–377 (2015).

8. Inagaki, H. K., Fontolan, L., Romani, S. & Svoboda, K. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature* **566**, 212–217 (2019).

9. Wang, X.-J. Synaptic Basis of Cortical Persistent Activity: the Importance of NMDA Receptors to Working Memory. *J. Neurosci.* **19**, 9587–9603 (1999).

10. Compte, A., Brunel, N., Goldman-Rakic, P. S. & Wang, X.-J. Synaptic Mechanisms and Network Dynamics Underlying Spatial Working Memory in a Cortical Network Model. *Cereb Cortex* **10**, 910–923 (2000).

11. Brunel, N. & Wang, X.-J. Effects of Neuromodulation in a Cortical Network Model of Object Working Memory Dominated by Recurrent Inhibition. *J Comput Neurosci* **11**, 63–85 (2001).

12. Wang, M. *et al.* NMDA receptors subserve persistent neuronal firing during working memory in dorsolateral prefrontal cortex. *Neuron* **77**, 736–749 (2013).

13. Romo, R., Hernández, A. & Zainos, A. Neuronal correlates of a perceptual decision in ventral premotor cortex. *Neuron* **41**, 165–173 (2004).

14. Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R. & Haynes, J.-D. The Distributed Nature of Working Memory. *Trends in Cognitive Sciences* **21**, 111–124 (2017).

15. Leavitt, M. L., Mendoza-Halliday, D. & Martinez-Trujillo, J. C. Sustained Activity Encoding Working Memories: Not Fully Distributed. *Trends in Neurosciences* **40**, 328–346 (2017).

16. Sreenivasan, K. K. & D'Esposito, M. The what, where and how of delay activity. *Nat. Rev. Neurosci.* **20**, 466–481 (2019).

17. Edin, F. *et al.* Mechanism for top-down control of working memory capacity. *PNAS* **106**, 6802–6807 (2009).

18. Murray, J. D., Jaramillo, J. & Wang, X.-J. Working Memory and Decision-Making in a Frontoparietal Circuit Model. *J. Neurosci.* **37**, 12167–12186 (2017).

19. Markov, N. T. *et al.* Cortical high-density counterstream architectures. *Science* **342**, 1238406 (2013).

20. Markov, N. T. *et al.* A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cereb. Cortex* **24**, 17–36 (2014).

21. Markov, N. T. *et al.* Anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. *J. Comp. Neurol.* **522**, 225–259 (2014).

22. Wang, X.-J. Probabilistic Decision Making by Slow Reverberation in Cortical Circuits. *Neuron* **36**, 955–968 (2002).

23. Chaudhuri, R., Knoblauch, K., Gariel, M.-A., Kennedy, H. & Wang, X.-J. A Large-Scale Circuit Mechanism for Hierarchical Dynamical Processing in the Primate Cortex. *Neuron* **88**, 419–431 (2015).

24. Joglekar, M. R., Mejias, J. F., Yang, G. R. & Wang, X.-J. Inter-areal Balanced Amplification Enhances Signal Propagation in a Large-Scale Circuit Model of the Primate Cortex. *Neuron* **98**, 222-234.e8 (2018).

25. Wang, X.-J. Macroscopic gradients of synaptic excitation and inhibition in the neocortex. *Nat Rev Neurosci* **21**, 169–178 (2020).

26. Elston, G. N. Specialization of the neocortical pyramidal cell during primate evolution. in *Evolution of Nervous Systems: a Comprehensive Reference* vol. 4 191–242 (Elsevier, 2007).

27. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).

28. Mejias, J. F., Murray, J. D., Kennedy, H. & Wang, X.-J. Feedforward and feedback frequency-dependent interactions in a large-scale laminar network of the primate cortex. *Sci Adv* **2**, e1601335 (2016).

29. Wong, K.-F. & Wang, X.-J. A recurrent network mechanism of time integration in perceptual decisions. *J. Neurosci.* **26**, 1314–1328 (2006).

30. Vugt, B. van, Kerkoerle, T. van, Vartak, D. & Roelfsema, P. R. The contribution of AMPA and NMDA receptors to persistent firing in the dorsolateral prefrontal cortex in working memory. *J. Neurosci.* (2020) doi:10.1523/JNEUROSCI.2121-19.2020.

31. Shadlen, M. N. & Newsome, W. T. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J. Neurophysiol.* **86**, 1916–1936 (2001).

32. Tsushima, Y., Sasaki, Y. & Watanabe, T. Greater disruption due to failure of inhibitory control on an ambiguous distractor. *Science* **314**, 1786–1788 (2006).

33. Murray, J. D. *et al.* A hierarchy of intrinsic timescales across primate cortex. *Nat Neurosci* **17**, 1661–1663 (2014).

34. Compte, A. Computational and in vitro studies of persistent activity: Edging towards cellular and synaptic mechanisms of working memory. *Neuroscience* **139**, 135–151 (2006).

35. Curtis, C. E. & D'esposito, M. The effects of prefrontal lesions on working memory performance and theory. *Cognitive, Affective, & Behavioral Neuroscience* **4**, 528–539 (2004).

36. Hilgetag, C. C. & Goulas, A. 'Hierarchy' in the organization of brain networks. *Philosophical Transactions of the Royal Society B: Biological Sciences* **375**, 20190319 (2020).

37. Muhle-Karbe, P. S., Myers, N. E. & Stokes, M. G. A hierarchy of functional states in working memory. *bioRxiv* 2020.04.16.044511 (2020) doi:10.1101/2020.04.16.044511.

38. Miller, E. K., Lundqvist, M. & Bastos, A. M. Working Memory 2.0. *Neuron* **100**, 463–475 (2018).

39. Fulcher, B. D., Murray, J. D., Zerbi, V. & Wang, X.-J. Multimodal gradients across mouse cortex. *PNAS* **116**, 4689–4695 (2019).

40. Burt, J. B. *et al.* Hierarchy of transcriptomic specialization across human cortex captured by structural neuroimaging topography. *Nat Neurosci* **21**, 1251–1259 (2018).

41. Demirtaş, M. *et al.* Hierarchical Heterogeneity across Human Cortex Shapes Large-Scale Neural Dynamics. *Neuron* **101**, 1181-1194.e13 (2019).

42. Mendoza-Halliday, D. & Martinez-Trujillo, J. C. Neuronal population coding of perceived and memorized visual features in the lateral prefrontal cortex. *Nat Commun* **8**, 1–13 (2017).

43. Bouchacourt, F. & Buschman, T. J. A Flexible Model of Working Memory. *Neuron* **103**, 147-160.e8 (2019).

44. Jaramillo, J., Mejias, J. F. & Wang, X.-J. Engagement of Pulvino-cortical Feedforward and Feedback Pathways in Cognitive Computations. *Neuron* **101**, 321-336.e9 (2019).

45. Kamiński, J. & Rutishauser, U. Between persistently active and activity-silent frameworks: novel vistas on the cellular basis of working memory. *Annals of the New York Academy of Sciences* **n/a**,.

46. Murray, J. D. *et al.* Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 394–399 (2017).

47. Lim, S. & Goldman, M. S. Balanced cortical microcircuitry for maintaining information in working memory. *Nat. Neurosci.* **16**, 1306–1314 (2013).

48. Stokes, M. G. ‘Activity-silent’ working memory in prefrontal cortex: a dynamic coding framework. *Trends in Cognitive Sciences* **19**, 394–405 (2015).

49. Mongillo, G., Barak, O. & Tsodyks, M. Synaptic theory of working memory. *Science* **319**, 1543–1546 (2008).

50. Masse, N. Y., Yang, G. R., Song, H. F., Wang, X.-J. & Freedman, D. J. Circuit mechanisms for the maintenance and manipulation of information in working memory. *Nature Neuroscience* **22**, 1159–1167 (2019).

Online Methods

Anatomical data

The anatomical connectivity data used has been gathered in an ongoing track tracing study in macaque and has been described in detail elsewhere[19–21,28]. Briefly, retrograde tracer injected into a given target area labels neurons in a number of source areas projecting to the target area. By counting the number of labeled neurons on a given source area, Markov et al. defined the fraction of labeled neurons (FLN) from that source to the target area. FLN can serve as a proxy for the "connection strength" between two cortical areas, which yields the connectivity pattern of the cortical network (Extended Data Fig. 1a, b). In addition, Markov et al. also measured the number of labeled neurons located on the supragranular layer of a given source area. Dividing this number over the total number of labeled neurons on that source area, we can define the supragranular layered neurons (SLN) from that source area to the target area (Extended Data Fig. 1c, d).

SLN values may be used to build a well-defined anatomical hierarchy[21,27]. Source areas located lower (higher) than the target area in the anatomical hierarchy, as defined in[27], display a progressively higher (lower) proportion of labeled neurons in the supragranular layer. As a consequence, the lower (higher) the source area relative to the target area, the higher (lower) the SLN values of the source-to-target projection.

Iterating these measurements across other anatomical areas yields and anatomical connectivity matrix with weighted directed connections and an embedded structural hierarchy. The 30 cortical used to build our data-constrained large-scale brain network are, in hierarchical order: V1, V2, V4, DP, MT, 8m, 5, 8l, 2, TEO, F1, STPc, 7A, 46d, 10, 9/46v, 9/46d, F5, TEpd, PBr, 7m, LIP, F2, 7B, ProM, STPi, F7, 8B, STPr and 24c. Finally, data on wiring connectivity distances between cortical areas is available for this dataset as well, allowing to consider communication time lags when necessary (we found however that introducing time lags this way does not have a noticeable impact on the dynamics of our model). The connectivity data used here is available to other researchers from core-nets.org.

The corresponding 30x30 matrices of FLN and SLN are shown in Extended Data Fig. 1b, d. Areas in these matrices are arranged following the anatomical hierarchy, which is computed using the SLN values and a generalized linear model[23,28]. Surgical and histology procedures were in accordance with European requirements 86/609/EEC and approved by the ethics committee of the region Rhone-Alpes.

In addition to the data on FLN and SLN across 30 cortical areas, we used additional data to constrain the area-to-area differences in the large-scale brain network. In particular, we have collected data on the total spine count of layer 2/3 pyramidal neuron basal dendrites across different cortical areas, as the spine count constitutes a proxy for the density of synaptic connections within a given cortical area[26]. A full list of all area-specific values of spine densities considered and their sources is given in Extended Data Table 1.We use an age correction factor meant to correct for the decrease of spine counts with age for data obtained from old monkeys. A plausible estimate would be a ~30% decrease for a 10y difference[51,52]. See Extended Data Fig. 2

for the effect of this correction on the overall gradient established by the spine count data, and the correlation of such gradient with the SLN hierarchy.

Computational model: Local neural circuit

We describe the neural dynamics of the local microcircuit representing a cortical area with the Wong-Wang model[29]. In its three-variable version, this model describes the temporal evolution of the firing rate of two input-selective excitatory populations as well as the evolution of the firing rate of an inhibitory population. All populations are connected to each other (see Fig. 1a). The model is described by the following equations:

$$\frac{dS_A}{dt} = -\frac{S_A}{\tau_N} + \gamma\,(1 - S_A)\,r_A \qquad \text{(Eq. 1)}$$

$$\frac{dS_B}{dt} = -\frac{S_B}{\tau_N} + \gamma\,(1 - S_B)\,r_B \qquad \text{(Eq. 2)}$$

$$\frac{dS_C}{dt} = -\frac{S_C}{\tau_G} + \gamma_I\,r_C \qquad \text{(Eq. 3)}$$

Here, S_A and S_B are the NMDA conductances of selective excitatory populations A and B respectively, and S_C is the GABAergic conductance of the inhibitory population. Values for the constants are τ_N=60 ms, τ_G=5 ms, γ=1.282 and γ_I=2. The variables r_A, r_B and r_C are the mean firing rates of the two excitatory and one inhibitory populations, respectively. They are obtained by solving the transcendental equation $r_i = \phi_i(I_i)$ at each time step, with Ii being the input to population I, given by

$$I_A = J_s S_A + J_c S_B + J_{EI} S_C + I_{0A} + I_{net}^A + x_A(t) \qquad \text{(Eq. 4)}$$

$$I_B = J_c S_A + J_s S_B + J_{EI} S_C + I_{0B} + I_{net}^B + x_B(t) \qquad \text{(Eq. 5)}$$

$$I_C = J_{IE} S_A + J_{IE} S_B + J_{II} S_C + I_{0C} + I_{net}^C + x_C(t) \qquad \text{(Eq. 6)}$$

In these expressions, J_s, J_c are the self- and cross-coupling between excitatory populations, respectively, J_{EI} is the coupling from the inhibitory populations to any of the excitatory ones, J_{IE} is the coupling from any of the excitatory populations to the inhibitory one, and J_{II} is the self-coupling strength of the inhibitory population. The parameters I_{0i} with i=A, B, C are background inputs to each population. Parameters are J_s=0.3213 nA, J_c=0.0107 nA, J_{IE}=0.15 nA, J_{EI}=-0.31 nA, J_{II}=-0.12 nA, I_{0A}=I_{0B}=0.3294 nA and I_{0C}=0.26 nA. Later we will modify some of these parameters in an area-specific manner (in particular J_s and J_{IE}) to introduce a gradient of properties across the cortical hierarchy. The term I^i_{net} denotes the long-range input coming from other areas in the network, which we will keep as zero for now but will be detailed later. Sensory stimulation can be introduced here as extra pulse currents of strength I_{pulse} =0.3 and duration T_{pulse} =0.5 sec (unless specified otherwise).

The last term $x_i(t)$ with i=A, B, C is an Ornstein-Uhlenbeck process, which introduces some level of stochasticity in the system. It is given by

$$\tau_{noise}\ \frac{dx_i}{dt} = -x_i + \sqrt{\tau_{noise}}\ \sigma_i\ \xi_i(t) \qquad \text{(Eq. 7)}$$

Here, $\xi_i(t)$ is a Gaussian white noise, the time constant is τ_{noise}=2 ms and the noise strength is $\sigma_{A,B}$=0.005 nA for excitatory populations and σ_C=0 for the inhibitory one.

The transfer function $\phi_i(t)$ which transform the input into firing rates takes the following form for the excitatory populations[53]:

$$\phi_{A,B}(I) = \frac{1}{2}\ \frac{aI-b}{1-\exp[-d\ (aI-b)]} \qquad \text{(Eq. 8)}$$

The values for the parameters are *a*=135 Hz/nA, *b*=54 Hz and *d*=0.308 s. For the inhibitory population a similar function can be used, but for convenience we choose a threshold-linear function:

$$\phi_C(I) = \frac{1}{g_I}\ (c_1 I - c_0) + r_0 \qquad \text{(Eq. 9)}$$

The values for the parameters are g_I=4, c_1=615 Hz/nA, c_0=177 Hz and r_0=5.5 Hz. Finally, it is sometimes useful for simulations (although not a requirement) to replace the transcendental equation $r_i = \phi_i(I_i)$ by its analogous differential equation, of the form

$$\tau_r \frac{dr_i}{dt} = -r_i + \phi_i(I_i) \qquad \text{(Eq. 10)}$$

The time constant can take a typical value of τ_r=2 ms.

<u>Computational model: Gradient of synaptic strengths</u>

Before considering the large-scale network and the inter-areal connections, we look into the area-to-area heterogeneity to be included in the model.

Our large-scale cortical system consists of N=30 local cortical areas, for which inter-areal connectivity data is available. Each cortical area is described as a Wong-Wang model of three populations like the ones described in the previous section. Instead of assuming areas to be identical to each other, here we will consider some of the natural area-to-area heterogeneity that has been found in anatomical studies. For example, work from Elston[26] has identified a gradient of dendritic spine density, from low spine numbers found in early sensory areas to large spine counts found in higher cognitive areas. This may reflect an increase of local recurrent strength as we move from sensory to association areas. In addition, cortical areas are distributed along an anatomical hierarchy[21,27]. The position of a given area within this hierarchy can be computed via a logistic regression of the SLN (fraction of supragranular layer neurons) projecting to and from that area[23].

In the following, we will assign the incoming synaptic strength (both local and long-range) of a given area as a linear function of the dendritic spine count values observed in anatomical studies,

with age-related corrections when necessary. Alternatively, when spine count data is not available for a given area, we will use its position in the anatomical hierarchy, which displays a high correlation with the spine count data, as a proxy for the latter. After this process, the large-scale network will display a gradient of local and long-range recurrent strength, with sensory/association areas showing weak/strong local connectivity, respectively. We denote the local strength value of a given area *i* in this gradient as h_i, and this value normalized between zero (bottom of the gradient, area V1) and one.

We assume therefore a gradient of values of J_s, with its value going from J_{min} to J_{max}. Having large values of J_s for association areas affects the spontaneous activity of these areas, even without considering inter-areal coupling. A good way to keep the spontaneous firing rate of these areas within physiologically realistic limits is to impose that the spontaneous activity fixed point is the same for all areas[18]. To introduce this into the model, we take into account that the solutions in the spontaneous state are symmetrical: S_A=S_B=S (we assume zero noise for simplicity). The current entering any of the excitatory populations is then (assuming I_{0A}=I_{0B}=I_0):

$$I = (J_s + J_c)S + J_{EI}S_C + I_0 \qquad \text{(Eq. 11)}$$

Assuming a fast dynamics for r_C and S_C (mediated by GABA) as compared to S_A and S_B (mediated by NMDA) we can obtain the approximate expression for S_C:

$$S_C \simeq \tau_G\, \gamma_I\, r_C = 2SJ_{IE}\zeta + \beta \qquad \text{(Eq. 12)}$$

with

$$\zeta = \frac{\tau_G\, \gamma_I\, c_1}{g_I - J_{II}\, \tau_G\, \gamma_I\, c_1} \qquad \text{(Eq. 13)}$$

$$\beta = \tau_G\, \gamma_I\, \frac{c_1\, I_{0C} + g_I\, r_0 - c_0}{g_I - J_{II}\, \tau_G\, \gamma_I\, c_1} \qquad \text{(Eq. 14)}$$

The equation for the excitatory current has then the form

$$I = (J_s + J_c)\, S + 2\, J_{EI}\, J_{IE}\zeta\, S + J_{EI}\, \beta + I_0 \qquad \text{(Eq. 15)}$$

To maintain the excitatory input (and therefore the spontaneous activity level S) constant while varying J_s across areas, we just have to keep the quantity $J_s + J_c + 2\, J_{EI}\, J_{IE}\, \zeta \equiv J_0$ constant (for the original parameters of the isolated area described above, we obtain J_0=0.2112 nA). A good choice, but not the only one, is to assume that the excitatory synapses to inhibitory neurons, J_{IE}, also scales with the ranks and with J_s accordingly:

$$J_{IE} = \frac{1}{2\, J_{EI}\, \zeta}\, (J_0 - J_s - J_c) \qquad \text{(Eq. 16)}$$

This linear relationship ensures that the spontaneous solution is the same for all areas in the network. Note that deviations from this linear relationship would simply lead to different areas having slightly different spontaneous activity levels, but it does not substantially affect our main results.

Since J_{IE} needs to be non-negative, the linear relationship above imposes a minimum value of J_{min}=0.205 nA for J_s. The particular maximum value of J_s, namely J_{max}, will determine the type of WM model we assume. Since the bifurcation point of an isolated area is at J_s=0.4655 nA for this set of parameter values, setting J_{max} below that value implies that all areas in the network are monostable in isolation. In this situation, any persistent activity displayed by the model will be a consequence of a global, cooperative effect due to inter-areal interactions. On the other hand, having J_{max} above the bifurcation point means that some areas will be multistable when isolated, e.g. they will be intrinsically multistable and compatible with classical WM theories.

Unless specified otherwise, we assume a range of J_{min}=0.21 nA and J_{max}=0.42 nA (i.e. below the critical value), so that the model displays distributed WM.

Computational model: Inter-areal projections

We now consider the inter-areal projections connecting isolated areas to form the large-scale cortical network. Assuming that inter-areal projections stem only from excitatory neurons (as inhibitory projections tend to be local in real circuits) and that such projections are selective for excitatory neurons, the network or long-range input term arriving at each of the populations of a given area *x* from all other cortical areas is given by

$$I_{A,net}^{x} = G \sum_{y} W^{xy} SLN^{xy} S_{A}^{y} \qquad \text{(Eq. 17)}$$

$$I_{B,net}^{x} = G \sum_{y} W^{xy} SLN^{xy} S_{B}^{y} \qquad \text{(Eq. 18)}$$

$$I_{C,net}^{x} = \frac{G}{Z} \sum_{y} W^{xy} (1 - SLN^{xy})(S_{A}^{y} + S_{B}^{y}) \qquad \text{(Eq. 19)}$$

Here, a superindex denotes the cortical area and a subindex the particular population within each area. The sum in all equations runs over all cortical areas of the network (N=30). Excitatory populations A and B receive long-range inputs from equally selective units from other areas, while inhibitory populations receive inputs from both excitatory populations. Therefore, neurons in population A of a given area may be influenced by A-selective neurons of other areas directly, and by B-selective neurons of other areas indirectly, via local interneurons.

G is the global coupling strength, which controls the overall long-range projection strength in the network (G=0.48 unless specified otherwise). Z is a factor that takes into account the relative balance between long-range excitatory and inhibitory projections. Setting Z=1 means that both excitatory and inhibitory long-range projections are equally strong, but this does not guarantee that their effect is balanced in the target area, due to the effect of local connections. Following previous work[18], we choose to impose a balance condition that guarantees that, if populations A and B have the same activity level, their net effect on other areas will be zero –therefore highlighting the selectivity aspect of the circuits. Again, deviations from this balance condition do not strongly affect our results besides the appearance of small differences between populations A and B. Considering that the transfer function of inhibitory populations is linear and their approximately linear rate-conductance relationship, it can be shown that

$$Z = \frac{2c_1\tau_G\gamma_I J_{EI}}{c_1\tau_G\gamma_I J_{II} - g_I} \qquad \text{(Eq. 20)}$$

Aside from global scaling factors, the effect of long-range projections from population *y* to population *x* is influenced by two factors. The first one, W^{xy}, is the anatomical projection strength as revealed by tract-tracing data[19]. We use the fraction of labelled neurons (FLN) from population *y* to *x* to constrain our projections values to anatomical data. We rescale these strengths to translate the broad range of FLN values (over five orders of magnitude) to a range more suitable for our firing rate models. We use a rescaling that maintains the proportions between projection strengths, and therefore the anatomical information, that reads

$$W^{xy} = k_1\,(FLN^{xy})^{k_2} \qquad \text{(Eq. 21)}$$

Here, the values of the rescaling are k_1 =1.2 and k_2 =0.3. The same qualitative behavior can be obtained from the model if other parameter values, or other rescaling functions, are used as long as the network is set into a standard working regime (i.e. signals propagate across areas, global synchronization is avoided, etc.) FLN values are also normalized so that $\sum_y FLN^{xy} = 1$. In addition, and as done for the local connections, we introduce a gradient of long-range projection strengths using the spine count data: $W^{xy} \rightarrow \lambda^x\, W^{xy}$, where λ_x=1 for the target area *x* with the maximal spine count, and decreases for other areas with the same slope as the gradient of the local connectivity presented above.

The second factor that needs to be taken into account is the directionality of signal propagation across the hierarchy. Feedforward (FF) projections that are preferentially excitatory constitute a reasonable assumption which facilitate signal transmission from sensory to higher areas. On the other hand, having feedback (FB) projections with a preferential inhibitory nature contributes to the emergence of realistic distributed WM patterns (Fig. 2d, e) (see also previous work[21,32]). This feature can be introduced, in a gradual manner, by linking the different inter-areal projections with the SLN data, which provides a proxy for the FF/FB nature of a projection (SLN=1 means purely FF, and SLN=0 means purely FB). In the model, we assume a linear dependence with SNL for projections to excitatory populations and with (1-SLN) for projections to inhibitory populations, as shown above.

Following recent evidence of frontal networks having primarily strong excitatory loops[54], it is convenient to ensure that the SLN-driven modulation of FB projections between frontal areas is not too large, so that interactions between these areas are never strongly inhibitory. In practice, such constraint is only necessary for projections from frontal areas to 8l and 8m (which are part of the frontal eye fields) and has little effect on the behavior of our model otherwise. The introduction of this cap, C_{FEF}, has two minor consequences: (i) it allows area 8l and 8m to exhibit a higher level of persistent activity during distributed WM –as their hierarchical position and recurrent strength are not strong enough to sustain activity otherwise, as previously suggested in anatomical studies[19,21], and (ii) it slightly shifts the bifurcation point in cortical space (see Fig. 2f). Unless specified otherwise (and in Fig. 2d, e, where the limitation is not considered for a cleaner study of the effects of the FB bias towards inhibition), we consider that the SLN-driven modulation of FB projections to 8l and 8m is never larger than C_{FEF}=0.4.

<u>Data analysis: Overview</u>

We developed a numerical method to estimate the number of stable distributed WM attractors for a particular set of parameters values of our large-scale model. This method, which follows simplified density-based clustering principles, is used to obtain the results shown in Figs. 3 and 5. To allow for a cleaner estimation, we do not consider noise in the neural dynamics during these simulations.

Our large-scale cortical model has 30 areas, with each of them having two selective excitatory populations A and B. Simply assuming that each of the areas can reach one of three possible states (persistent activity in A, persistent activity in B, or spontaneous activity) means that our model can potentially display up to 3 to the power of 30 attractor combinations. This number can be even larger if we refine the firing rate reached by each area rather than simply its persistent/non-persistent activity status. Since it is not possible to fully explore this extremely large number of possible attractors, we devised a strategy based on the exploration of a sample of the input space of the model. The core idea is to stimulate the model with a certain input pattern (targeting randomized areas) and registering the fixed point that the dynamics of the model converges to. By repeating this process with a large number of input combinations and later counting the number of different attractors from the obtained pool of fixed points, we can obtain an estimate of the number of attractors for a particular set of parameter values.

Data analysis: Stimulation protocol

A given input pattern is defined as a current pulse of fixed strength (I_{pulse} =0.2) and duration (T_{pulse} =1 sec) which reaches a certain number P of cortical areas. Only one population (A or B, randomized) in each area receives the input, and the P cortical areas receiving the input are randomly selected across the top 16 areas of the spine count gradient. This decreases the amount of potential input combinations we have to deal with by acknowledging that areas with stronger recurrent connections (such as 9/46d) are more likely to be involved in distributed WM patterns than those with weaker connections (such as MT). P can take any value between one and P_{max} =16, and we run a certain number of trials (see below) for each of them. Different values of I_{pulse} and T_{pulse}, as well as setting the randomly selected areas at a high rate initial condition instead of providing an external input, have been also explored and lead to qualitatively similar results. Similar approaches regarding the use of random perturbations in high dimensional systems have been successfully used in the past[55].

It is also important to consider that not all values of P have the same number of input combinations. For example, P=1 allows for 16*2=32 different input combinations (if we discriminate between populations A and B), while P=2 allows for 16*(16-1)*2=480 input combinations, and so on. For a given value of P, the number of possible input combinations N_c is given by

$$N_c = 2^P \binom{P_{max}}{P} = 2^P \frac{P_{max}!}{(P_{max}-P)!\,P!} \qquad \text{(Eq. 22)}$$

By summing all values of N_c for P=1, …P_{max}, we obtain around 43 million input combinations, which are still too many trials to simulate for a single model configuration. To simplify this further, we consider a scaling factor F_c on top of N_c to bring down these numbers to reasonable

levels for simulations. We use F_c =0.0002 (or 0.02% of all possible combinations) for our calculations, which brings down the total number of simulated input combinations to around 9000. Other options, such as decreasing P_{max} and using a larger scaling factor (P_{max} =12, F_c =0.01 or 1% or all possible combinations) give also good results. Since the rescaling can have a strong impact for small P (yielding a number of trials smaller than one), we ensure at least one trial for these cases.

To guarantee the stability of the fixed points obtained during these simulations, we simulate the system during a time windows of 30 seconds (which is much larger than any other time scale in the system), and check that the firing rates have not fluctuated during the last 10 seconds before we register the final state of the system as a fixed point.

Data analysis: Estimating the number of attractors

The final step is to count how many different attractors have been reached by the system, by analyzing the pool of fixed points obtained from simulations. A simple way to do this is to consider that, for any fixed point, the state of each area can be classified as persistent activity in population A (i.e. mean firing rate above a certain threshold of 10 spikes/s), persistent activity in population B, or spontaneous activity (both A and B are below 10 spikes/s). This turns each fixed point into a vector of 30 discrete states, and the number of unique vectors among the pool of fixed points can be quickly obtained using standard numerical routines in Matlab (such as the 'unique' function).

A more refined way to count the number of attractors, which we use in this work, is to define the Euclidean distance to discriminate between an attractor candidate and any previously identified attractors. Once the first attractor (i.e. the first fixed point analyzed) is identified, we test whether the next fixed point is the same than the first one by computing the Euclidean distance E_d between them:

$$E_d = \frac{1}{n} \sum_{i=1}^{n} (r_i^{new} - r_i^{old})^2 \qquad \text{(Eq. 23)}$$

where n=30 is the total number of areas in the network (only one of the populations, A or B, needs to be considered here). If E_d is larger than a certain threshold distance ε, we consider it a new attractor. We choose ε=0.01, which grossly means that two fixed points are considered as different attractors if, for example, the activity of one of their cortical areas differs by 0.5 spikes/s and the activity on all other areas is the same for both. The particular value of ε does not have a strong impact on the results (aside from the fact that smaller values of ε gives us more resolution to find attractors). When several attractors are identified, each new candidate is compared to all of them using the same method.

Both the first and the second method to count attractors deliver qualitatively similar results (in terms of the dependence of the number of attractors with model parameters), although as expected the second method yields larger numbers due to its higher discriminability.

References:

51. Duan, H. *et al.* Age-related dendritic and spine changes in corticocortically projecting neurons in macaque monkeys. *Cereb. Cortex* **13**, 950–961 (2003).
52. Young, M. E., Ohm, D. T., Dumitriu, D., Rapp, P. R. & Morrison, J. H. Differential effects of aging on dendritic spines in visual cortex and prefrontal cortex of the rhesus monkey. *Neuroscience* **274**, 33–43 (2014).
53. Abbott, L. F. & Chance, F. S. Drivers and modulators from push-pull and balanced synaptic input. *Prog. Brain Res.* **149**, 147–155 (2005).
54. Markowitz, D. A., Curtis, C. E. & Pesaran, B. Multiple component networks support working memory in prefrontal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 11084–11089 (2015).
55. Sussillo, D. & Barak, O. Opening the Black Box: Low-Dimensional Dynamics in High-Dimensional Recurrent Neural Networks. *Neural Computation* **25**, 626–649 (2012).
56. Siegle, J. H. *et al.* A survey of spiking activity reveals a functional hierarchy of mouse corticothalamic visual areas. *bioRxiv* 805010 (2019) doi:10.1101/805010.
57. Schmidt, M. *et al.* A multi-scale layer-resolved spiking network model of resting-state dynamics in macaque visual cortical areas. *PLOS Computational Biology* **14**, e1006359 (2018).
58. Elston, G. N. & Rosa, M. G. The occipitoparietal pathway of the macaque monkey: comparison of pyramidal cell morphology in layer III of functionally related cortical visual areas. *Cereb. Cortex* **7**, 432–452 (1997).
59. Elston, G. N., Tweedale, R. & Rosa, M. G. Cortical integration in the visual system of the macaque monkey: large-scale morphological differences in the pyramidal neurons in the occipital, parietal and temporal lobes. *Proc. Biol. Sci.* **266**, 1367–1374 (1999).
60. Elston, G. N. & Rosa, M. G. Morphological variation of layer III pyramidal neurones in the occipitotemporal pathway of the macaque monkey visual cortex. *Cereb. Cortex* **8**, 278–294 (1998).
61. Elston, G. N. & Rosa, M. G. Complex dendritic fields of pyramidal cells in the frontal eye field of the macaque monkey: comparison with parietal areas 7a and LIP. *Neuroreport* **9**, 127–131 (1998).
62. Elston, G. N. & Rockland, K. S. The pyramidal cell of the sensorimotor cortex of the macaque monkey: phenotypic variation. *Cereb. Cortex* **12**, 1071–1078 (2002).
63. Elston, G. N., Benavides-Piccione, R., Elston, A., Manger, P. R. & De Felipe, J. Pyramidal cells in prefrontal cortex of primates: marked differences in neuronal structure among species. *Front Neuroanat* **5**, 2 (2011).

64. Elston, G. N. Interlaminar differences in the pyramidal cell phenotype in cortical areas 7 m and STP (the superior temporal polysensory area) of the macaque monkey. *Exp Brain Res* **138**, 141–152 (2001).

65. Elston, G. N., Benavides-Piccione, R. & De Felipe, J. A study of pyramidal cell structure in the cingulate cortex of the macaque monkey with comparative notes on inferotemporal and primary visual cortex. *Cereb. Cortex* **15**, 64–73 (2005).

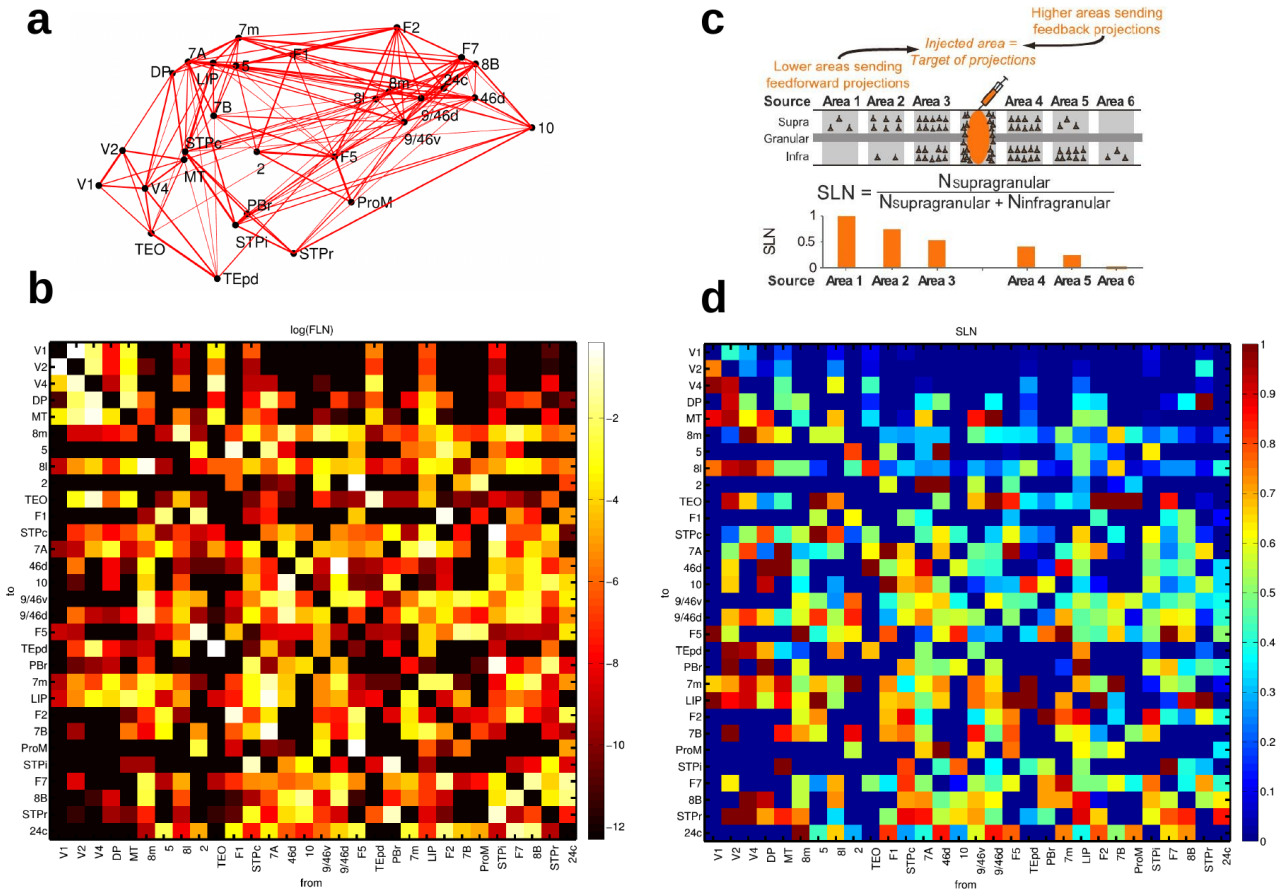
Acknowledgments: We thank Rishidev Chaudhuri, John Murray and Jorge Jaramillo for their support during the development of this work, and Henry Kennedy for providing the connectivity dataset.

Funding: This work was supported by the NIH grant R01MH062349, ONR grant N00014-17-1-2041 and the Simons Foundation Collaboration on the Global Brain grant (to XJW).

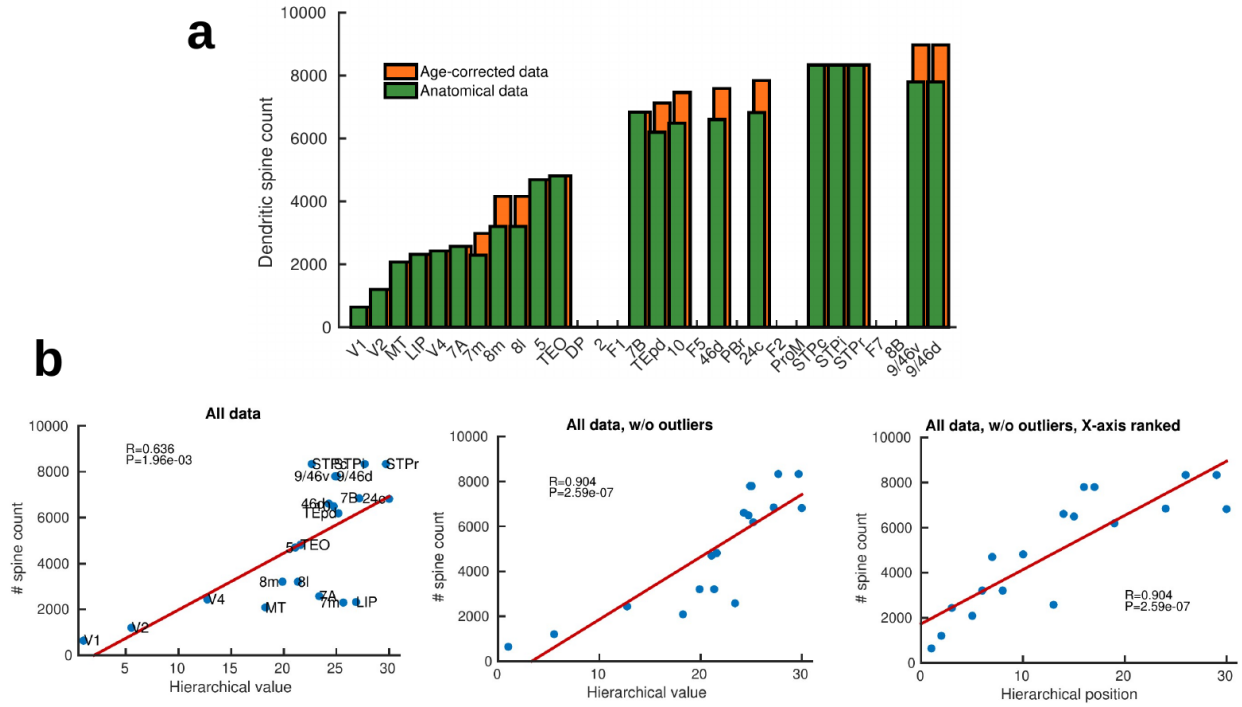
Author contributions: JFM and X-JW designed the study, JFM performed the research, JFM and X-JW discussed the results and wrote the manuscript.

Competing interests: Authors declare not competing interests.

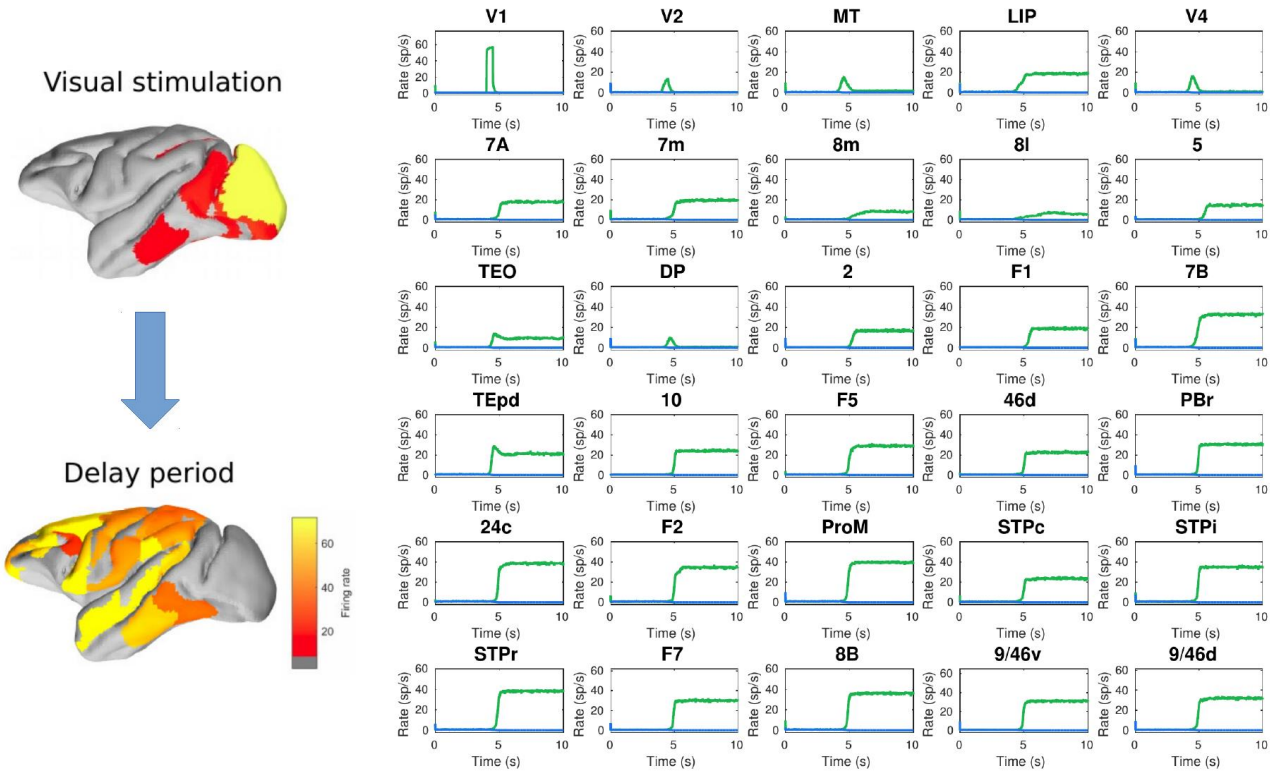
Data and materials availability: All information needed to reproduce the results of this manuscript are in the main text and supplementary material, and the code used will be made available upon publication of this work.



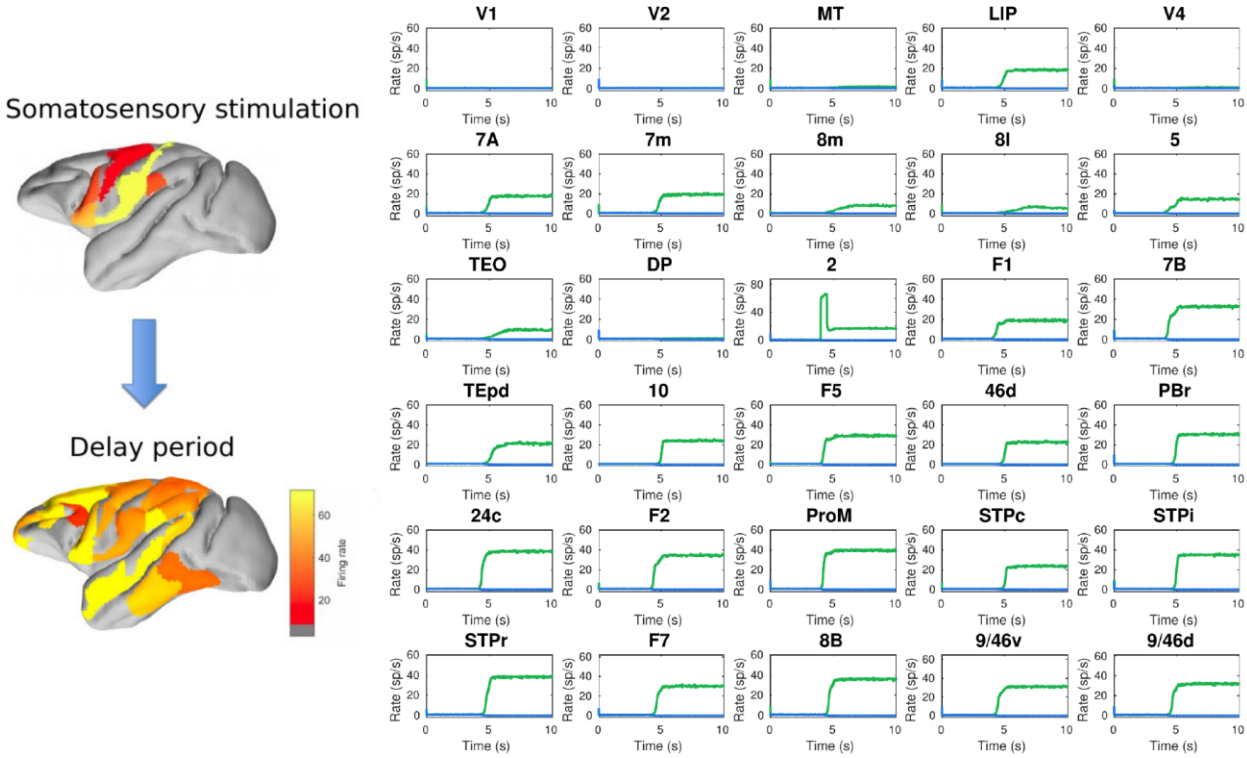
Extended Data Fig. 1: Anatomical connectivity data of the macaque cortex. Data from anatomical studies[19–21]. (a) Connectivity of the 30 areas (positioned in 3D space following injection coordinates of experiments). Width of the lines denote two-way averaged FLN values (i.e. average strength of the projection). (b) Map of FLN values for all connections considered. (c) The proportion of supragranular vs infragranular neurons projecting to the injection site allowed to define an anatomical hierarchy. (d) Map of SLN values for all connections considered.



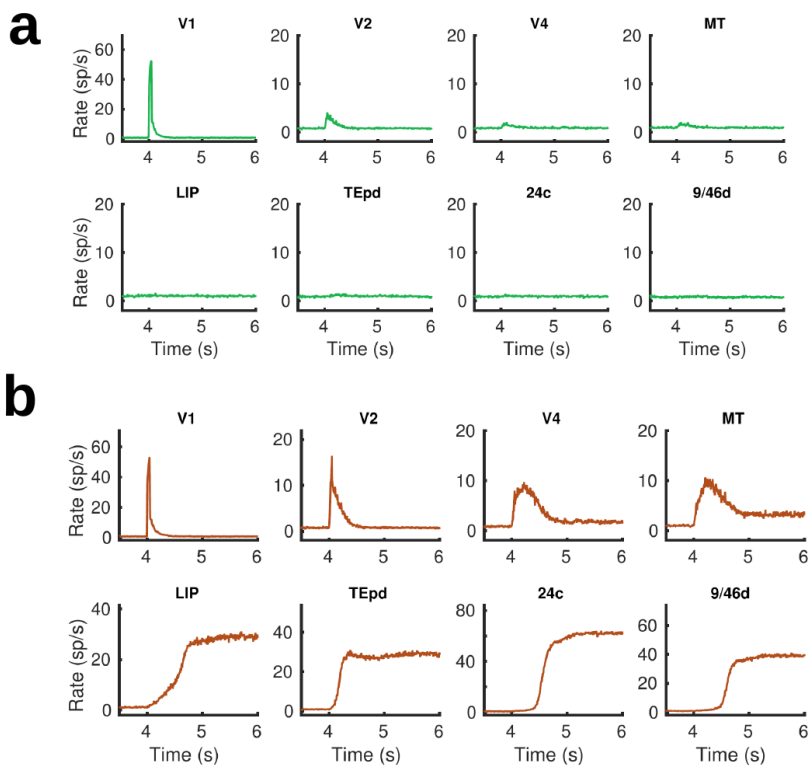
Extended Data Fig. 2: Spine count data used to constrain connectivity strength. Data from anatomical studies[26], see also Extended Data Table 1. (a) Spine count of the basal dendrites of layer 2/3 neurons across cortical areas of young (2~3 years old) macaques. When data from older macaques had to be considered, an age correction was introduced (orange bars). (b) Correlation between the spine count data and the hierarchical value for all areas (left), the hierarchical value for all areas except outliers (7m, LIP, and STPc), and the hierarchical position (rank) for all areas except outliers. Pearson correlation and corresponding P-value are shown in each panel. In the model, the connectivity strength of areas for which spine count data was not available was estimated using their hierarchical value.



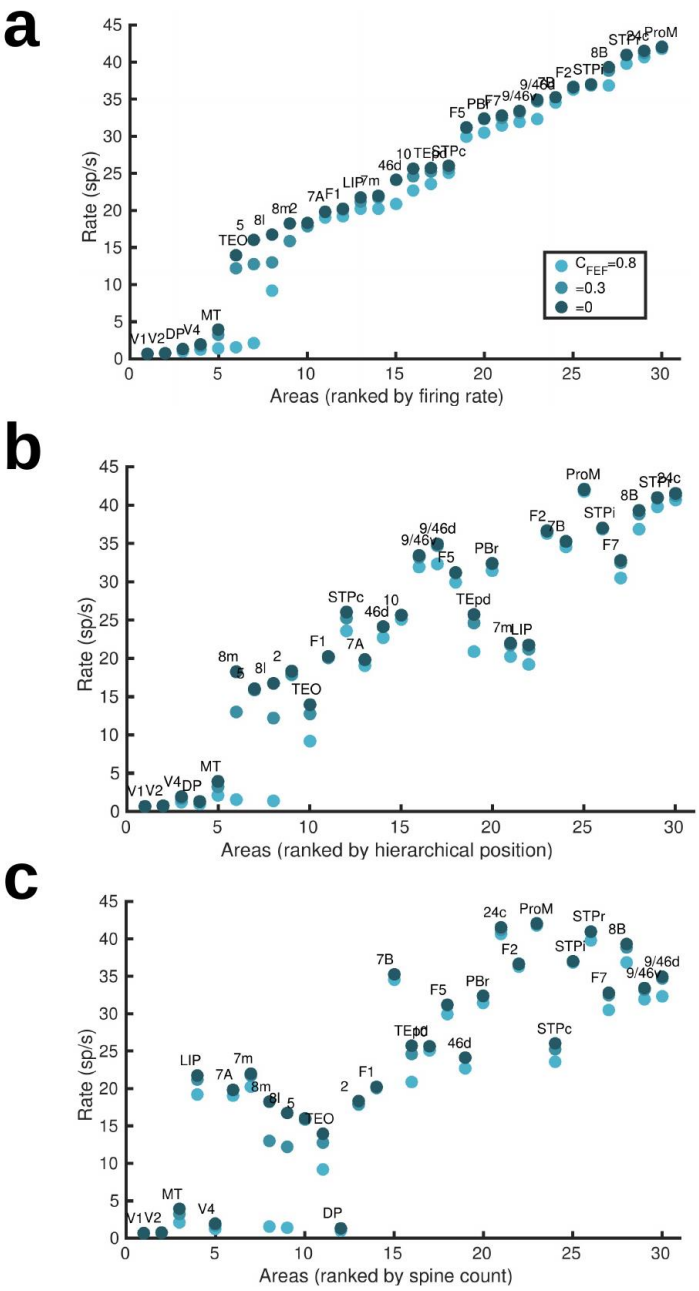
Extended Data Fig. 3: Behavior of all areas in the network during the visual WM task. Left: spatial maps of the simulated macaque brain during stimulation and delay period, with activity color coded. Right: evolution of the firing rate of all areas in the network. Stimulation occurs in V1 at t=4 seconds and has a duration of 500 ms. Parameters as in Fig. 2c.



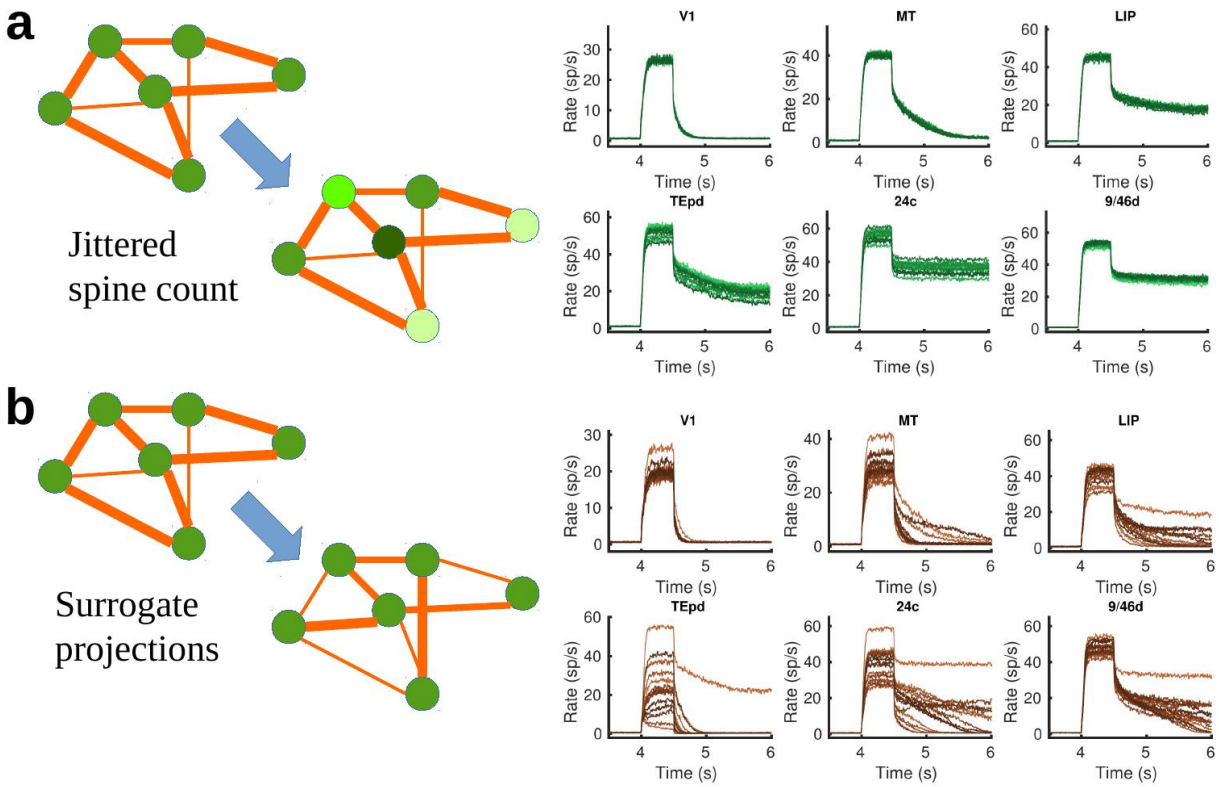
Extended Data Fig. 4: Behavior of all areas in the network during the somatosensory WM task. Left: spatial maps of the simulated macaque brain during stimulation and delay period, with activity color coded. Right: evolution of the firing rate of all areas in the network. Stimulation occurs in area 2 (primary somatosensory area) at t=4 seconds, and has a duration of 500ms. Other parameters as in Fig. 2c.



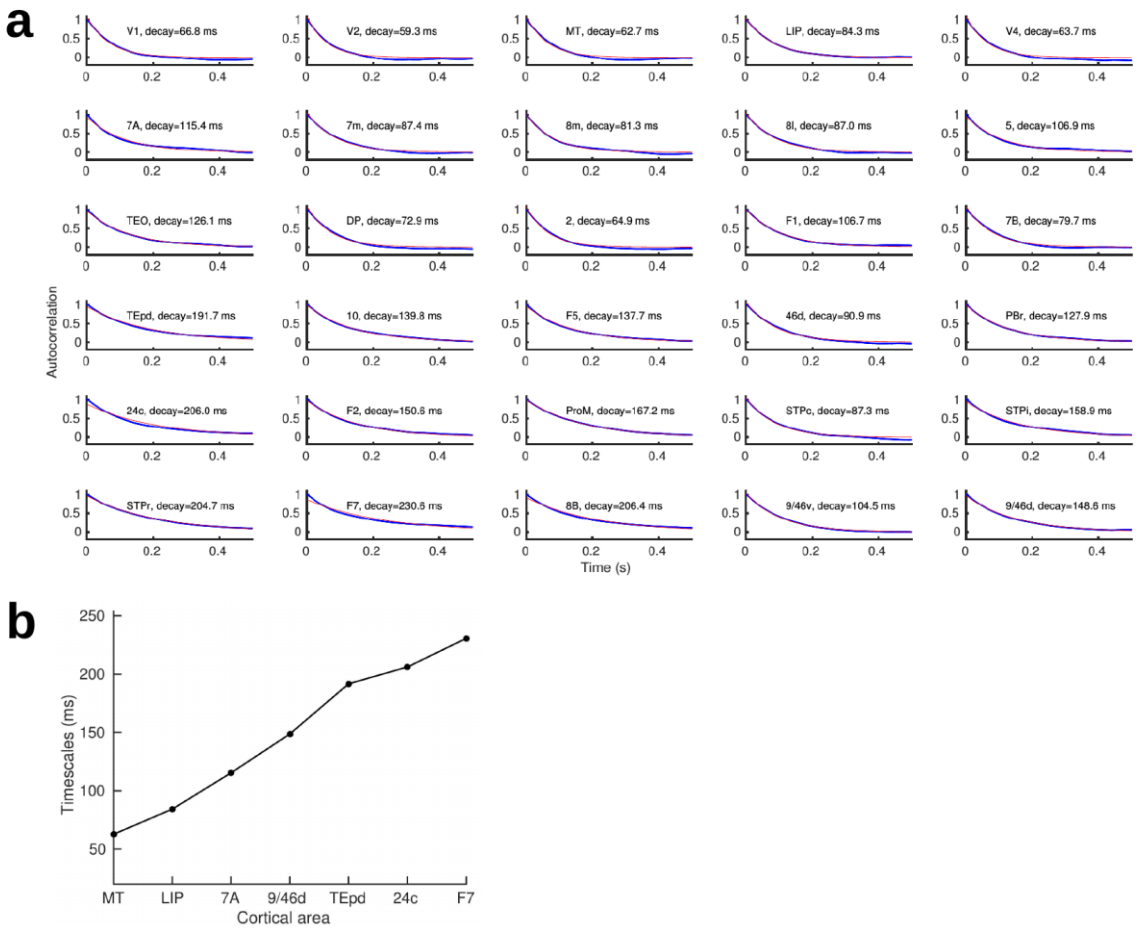
Extended Data Fig. 5: Firing rates for selected areas during a visual WM task with a short (50ms) stimulus duration. (a) Only NMDA and GABA synapses are considered in the network, the stimulus does not reach frontal areas. (b) By introducing simplified AMPA-like synapses (proportional to the firing rates) on feedforward excitatory projections, we obtain distributed WM patterns for this type of short inputs. The results suggests that it is convenient to consider AMPA/NMDA asymmetry along the FF/FB projections in the hierarchy for brief stimulation patterns. Other parameters as in Fig. 2c.



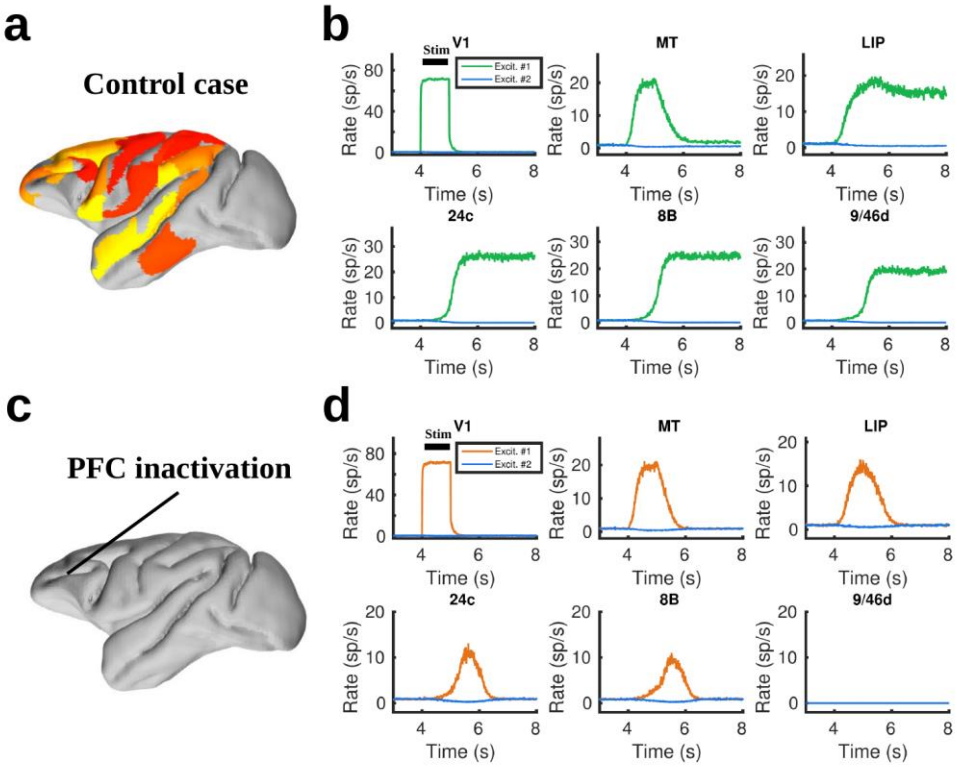
Extended Data Fig. 6: Firing rate of ranked cortical areas reveals a robust bifurcation in space with different ranking systems. (a) Areas ranked by displayed firing rate. (b) Areas ranked following their hierarchical (as obtained from SLN data) position. (c) Areas ranked by spine count. All panels show data for three different values of the cap in the modulation of feedback projections to inhibitory populations, C_{FEF}. Parameters as in Fig. 2f.



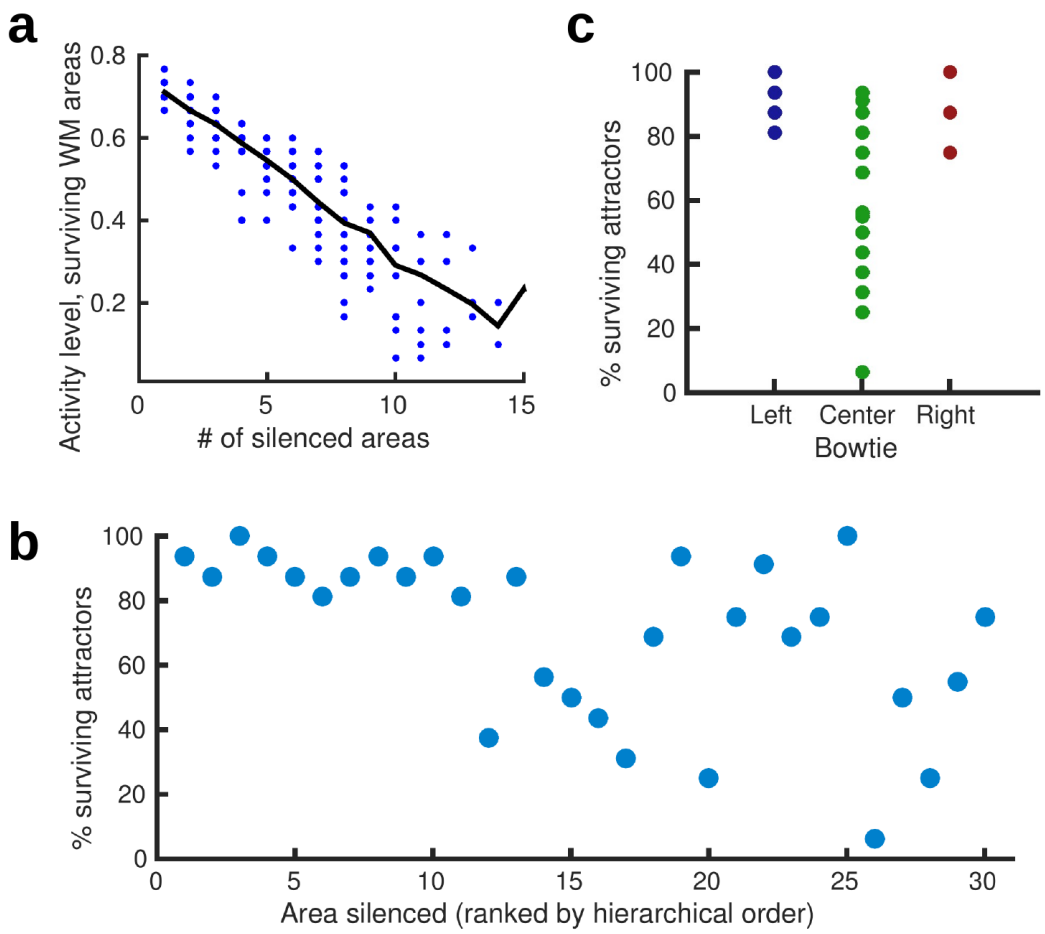
Extended Data Fig. 7: Effect of jitter on dendritic spine values and surrogate networks. (a) Networks with jittered values of the area-specific dendritic spine numbers, and therefore jittered local and long-range synaptic strengths (left), show little differences with respect to each other (right; lines represent activity for different configurations), with the only effect being a different persistent activity levels for some areas such as 24c. The jitter considered was 15% of the original value. (b) On the contrary, surrogate networks in which projection strengths (i.e. FLN values) were shuffled (left) showed marked differences form each other, with the ability of many areas like 24c and 9/46d to support distributed persistent activity being strongly affected (right; lines represent activities for different surrogate networks). Parameters as in Fig. 2c.



Extended Data Fig. 8: A gradient of time scales emerges in the network. (a) Autocorrelation function of the activity (low-pass-filtered to simulate an LFP recording) of each area in the network, fitted with a decaying exponential function. (b) The values for the local decay time constant increase with the cortical hierarchy, going from ~60 ms in early visual areas to ~300 ms in some frontal areas. This range is similar to the spike-count autocorrelations found experimentally across different cortical areas[33,56], and improves predictions from previous models[23] (but see additional considerations[57]). Parameters as in Fig. 2c.



Extended Data Fig. 9: Effect of silencing frontal areas on the emergence of distributed attractors. (a, b) Visual stimulation of a network with maximum local and long-range connectivity strength of J_{max}=0.25 and G=0.46, displaying a distributed working memory pattern. (c, d) Upon inactivating prefrontal areas 9/46d and 9/46v, the network is unable to sustain the visually evoked distributed persistent activity. Therefore, prefrontal areas still play a major role in working memory maintenance for networks with moderate values of J_{max}, which would explain classical prefrontal lesion studies[35]. Other parameters as in Fig. 4b.



Extended Data Fig. 10: Effect of inactivating areas on the stability of attractors. (a) The activity level, or firing rate, or areas involved in a visually evoked distributed WM attractor decreases linearly with the number of silenced areas. (b) The percentage of attractors surviving inactivations is specific of the area silenced, with inactivating areas higher in the hierarchy having a stronger impact. (c) Inactivating areas at the center of the anatomical bowtie hub has a stronger impact than inactivating the 'side areas'. Both (b) and (c) were obtained during the numerical exploration of the stable distributed attractors in the network. Parameters as in Fig. 5.

Extended Data Table 1.

Rank in SLN hierarchy	Area name	Measured spine count	Age correction factor	Source
1	V1	643	1	[58,59]
2	V2	1201	1	[58]
3	V4	2429	1	[60]
4	DP	-	-	
5	MT	2077	1	[59]
6	8m	3200	1.30	[61]
7	5	4689	1	[62]
8	8l	3200	1.30	[61]
9	2	-	-	
10	TEO	4812	1	[60]
11	F1	-	-	
12	STPc	8337	1	[59]
13	7a	2572	1	[58]
14	46d	6600	1.15	Estimated from [26]; (*47*)
15	10	6488	1.15	[63]
16	9/46v	7800	1.15	Estimated from [26]
17	9/46d	7800	1.15	Estimated from [26]
18	F5	-	-	
19	TEpd	7260	1	[59]
20	PBr	-	-	
21	7m	2294	1.30	[64]
22	LIP	2316	1	[58]
23	F2	-	-	
24	7B	6841	1	[62]
25	ProM	-	-	
26	STPi	8337	1	[59]
27	F7	-	-	
28	8B	-	-	
29	STPr	8337	1	[59]
30	24c	6825	1.15	[65]

Extended Data Table 1: Spine count data from basal dendrites of layer 2/3 pyramidal neurons in young (~2 y. o.) macaque, acquired from the specified literature. See also Extended Data Fig. 2.