Multiple component networks support working memory in prefrontal cortex

David A. Markowitz[a], Clayton E. Curtis[a,b], and Bijan Pesaran[a,1]

[a]Center for Neural Science, New York University, New York, NY 10003; and [b]Department of Psychology, New York University, New York, NY 10003

Edited by Michael E. Goldberg, Columbia University College of Physicians, New York, NY, and approved July 17, 2015 (received for review February 28, 2015)

Lateral prefrontal cortex (PFC) is regarded as the hub of the brain's working memory (WM) system, but it remains unclear whether WM is supported by a single distributed network or multiple specialized network components in this region. To investigate this problem, we recorded from neurons in PFC while monkeys made delayed eye movements guided by memory or vision. We show that neuronal responses during these tasks map to three anatomically specific modes of persistent activity. The first two modes encode early and late forms of information storage, whereas the third mode encodes response preparation. Neurons that reflect these modes are concentrated at different anatomical locations in PFC and exhibit distinct patterns of coordinated firing rates and spike timing during WM, consistent with distinct networks. These findings support multiple component models of WM and consequently predict distinct failures that could contribute to neurologic dysfunction.

working memory | prefrontal cortex | macaque | coherence

High-level cognition depends on the ability to translate stored information about recent experience into a behaviorally appropriate response, an ability known as working memory (WM). WM relies on a storage process that actively maintains information and a control process that manipulates stored information to support the selection and preparation of a contingent response (1–3). The neural mechanisms that support WM involve networks that are broadly distributed throughout the brain (4–7) and rely heavily on the prefrontal cortex (PFC) for normal operation (6–9). However, the degree to which WM is supported by a single distributed network or multiple specialized network components in PFC remains unclear (6, 10, 11), hindering progress in the search for neurocognitive therapies to treat disorders of cognition (12).

Persistent spiking activity is commonly thought to reflect the mechanistic basis of WM in PFC (13–16). This activity manifests in different ways, including time-varying neuronal responses that decay, ramp up, or are stable in time during memory delays. Although such a diversity of responses could reflect distinct modes of persistent activity, it has long been a standard practice to treat all persistently active neurons in PFC as representative of a single composite WM function that supports the maintenance and manipulation of information necessary for memory-guided behavior (14, 17–19). The implicit assumption that the representations of stored information and contingent responses overlap at the neural circuit level contrasts with an alternate view, which suggests that PFC primarily encodes the selection and preparation of responses (6, 10, 11). This difference highlights the need to directly investigate the circuit-level organization of storage and response preparation-related activity in PFC.

We address this problem here, using a simple manipulation of WM in concert with large-scale recordings from neurons across lateral PFC of macaque monkeys. By mapping neural activity during memory and visual delays of the same oculomotor delayed response (ODR) task, we show that WM is composed of three anatomically specific modes of persistent activity. The first two modes specifically encode early and late forms of memory storage, and the third mode predicts behavioral variability after the delay, consistent with response preparation. We then offer multiple convergent lines of evidence that the neural populations that support these three modes are organized with distinct spatiotemporal profiles in PFC. These results suggest that information storage and the preparation of contingent responses are supported by functionally specialized networks in PFC.

Results

To test whether storage and response preparation are supported by different modes of persistent activity during WM, we mapped neural activity with chronically implanted movable multielectrode arrays placed over the right prearcuate gyrus of lateral PFC in two monkeys (20). During each experimental session, each monkey performed an ODR task with randomly interleaved trials of memory-guided (mODR) and visually guided (vODR) saccades to one of eight isoeccentric targets (Fig. 1*A*) (21). Over the course of 41 sessions in monkey A and 43 sessions in monkey S, we slowly advanced the electrodes to isolate and record from 746 units (384 in monkey A, 362 in monkey S) at multiple depths extending 2.5 mm below the cortical surface. After the completion of all recordings, we registered the depths measured on each electrode to the cortical surface, using an iterative algorithm (see *Experimental Procedures*).

We hypothesized that interleaving memory-guided and visually guided trials should reveal multiple distinct storage and response preparation-related modes of persistent activity. Our first prediction is that storage modes should be revealed by differences in activity across the memory and visual delay conditions that arise from differences in how information is maintained during each task (Fig. 1*B*). A memory–visual comparison tests a critical assumption made by previous studies: that a memory delay is sufficient to identify neurons that reflect storage, and by extension, that a visual delay involves the same storage process as a memory delay. Note that this reasoning does not exclude the possibility that the visual delay also

Significance

Human cognition depends on the brain's ability to remember and manipulate information about recent experience. Although this working memory ability has been linked with the persistent spiking activity of neurons in prefrontal cortex (PFC), the population-scale organization of working memory in this region remains poorly understood. Here we provide the first population-scale map, to our knowledge, of persistent activity across multiple depths and topographic regions of PFC in two monkeys performing a delayed eye movement task. We show that working memory is supported by three functionally specialized networks with different anatomical localizations, and not by a functionally homogeneous network, as commonly believed. These findings reveal specific circuit-level targets for studying the causes of working memory dysfunction in many psychiatric disorders.

Author contributions: D.A.M., C.E.C., and B.P. designed research; D.A.M. and B.P. performed research; D.A.M. and B.P. analyzed data; and D.A.M., C.E.C., and B.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: bijan@nyu.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1504172112/-/DCSupplemental.

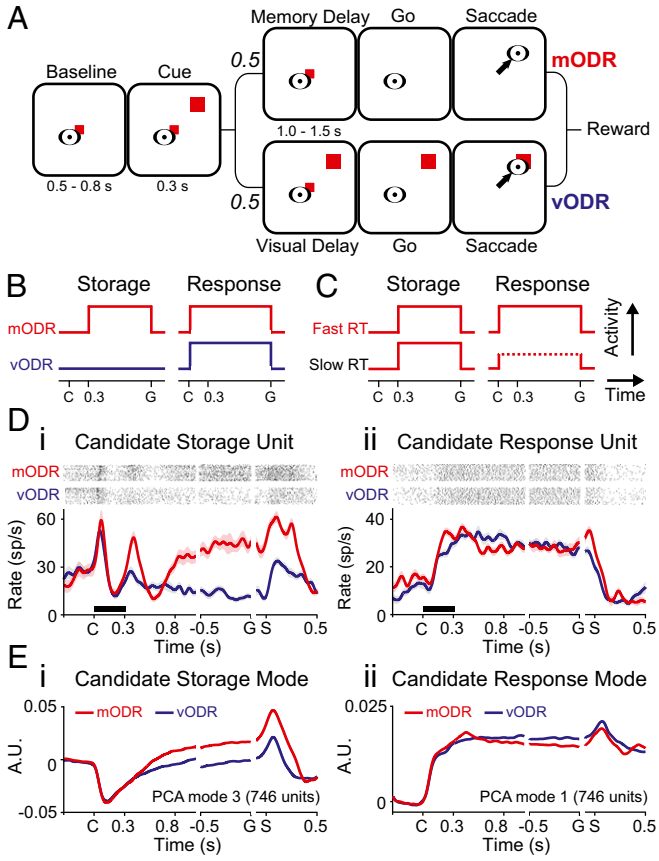

Fig. 1. Experimental design for testing the distinct mode hypothesis. (*A*) ODR paradigm with interleaved mODR and vODR trials. (*B* and *C*) Predicted responses of neurons that encode stored information and prepared responses as a function of (*B*) delay type and (*C*) reaction time after a memory delay. (*D*) Preferred target spike rasters and peri-stimulus time histograms (PSTHs) for two spatially tuned units that fire persistently during the memory delay. mODR traces in red, vODR traces in blue. Black bar denotes Cue interval. Event alignment labels: C (Cue), G (Go), S (Saccade). (*i*) Candidate storage unit. (*ii*) Candidate response unit. (*E*) Two modes of activity that were identified by principal component analysis in the full population of 746 recorded units. (*i*) Candidate storage mode. (*ii*) Candidate response mode. Axis labels as in *D*.

involves a storage process, and that this reasoning only predicts that storage differs across the memory and visual delays. In contrast, we reasoned that the modes reflecting the selection and preparation of the response should not differ across the memory–visual comparison. We predict that the activity of neurons related to the response will be elevated equally during both memory and visual delay periods, because the same eye movement is selected and prepared under both conditions. Furthermore, we predict that the storage and response processes should differ in their relationships to movement reaction time (RT) after a memory delay (Fig. 1*C*). Specifically, the activity of neurons that encode prepared responses will covary with future RT, whereas the activity of storage neurons will not, because only the response mode, which reflects the manipulation of information to generate the response, influences the timing of behavior after the delay. With these predictions in mind, we examined the neural responses.

We first examined individual neuron responses to test whether both storage and response activities are present in PFC. We found that many PFC neurons responded with different activity patterns during the memory and visual delays (Fig. 1 *D*, *i*), consistent with a role in a memory storage process. Others responded equally during the delay periods of both tasks (Fig. 1 *D*, *ii*), consistent with a role in preparing the response. Finally, a third population responded strongly during the extended presentation of preferred visual stimuli and subsequent visual delay, but only briefly during the memory delay, consistent with a role in a cue encoding and early storage process. These task-selective responses were present in each monkey individually, and typically emerged in response to preferred target locations only (*SI Appendix*, Fig. S1).

Although the presence of memory delay-specific responses suggests that storage and response preparation map to distinct modes of activity, individual examples are not sufficient to draw this conclusion. To more rigorously test whether distinct single-unit responses are representative of distinct modes at the population level, we studied population dynamics using a dimensionality reduction approach. We used principal component analysis to identify eigenmodes of population activity that compactly describe the responses of all 746 isolated neurons, including those without clear spatial tuning or task selectivity, across all eight targets during both tasks (see *Experimental Procedures*). This revealed modes with delay activity that qualitatively resembled the task-selective single unit responses described earlier, including putative storage-related (Fig. 1 *E*, *i*) and response-related (Fig. 1 *E*, *ii*) modes and a visually driven mode with decaying activity during the memory delay (*SI Appendix*, Fig. S2*A*).

To determine whether the putative storage- and response-related modes of population activity are associated with functionally distinct networks, we further examined the activity of individual PFC neurons. In general, single unit responses reflect a mixture of persistent modes across all phases of the task (*SI Appendix*, Fig. S2*B*), arguing against the existence of functionally specialized neurons. However, many responses are clearly dominated by a single task-selective mode during the late delay, just before the Go command, as illustrated in Fig. 1*D*. If different unit responses reflect activity in different WM modes, then we reasoned it should be possible to study and compare the properties of different modes by operationally classifying single-unit responses by their dominant mode during the late delay.

To support this analysis, we identified units that persistently encoded spatial information during the late delay and then classified them by their task selectivity during this epoch. We focused on positively tuned regular-spiking units in this study (n = 365; *SI Appendix*, Fig. S3) because neurons of this type are believed to play a fundamental role in the maintenance of persistent activity states in PFC (15, 22). To classify neurons, first we quantified each unit's spatial tuning during the late memory delay and, separately, during the late visual delay with a tuning z-score (Z_{tun}) (23). We then quantified each unit's target tuning during the late memory or visual delay with a z-score (Z_{sel}) that compared the late memory and visual delay responses to a null distribution. We assigned each persistently active unit with significant late delay tuning ($Z_{tun} > 1.65$) to a memory-selective ($Z_{sel} > 1.65$; n = 93) or nontask-selective population ($|Z_{sel}| < 1.65$; n = 200) based on the task selectivity of its preferred target response during the late delay (see *Experimental Procedures*). The remaining 72 units all exhibited significant visual task selectivity during the late delay ($Z_{sel} < -1.65$) and likely reflected a mixture of purely visual and storage-related units. We controlled for the presence of purely visual responses by requiring all units in this group to exhibit significant spatial tuning during the early memory delay (0–300 ms after Cue offset), leaving 59 units in this population.

Trial-averaged responses for each classified population of WM units are shown in Fig. 2. Activity in the visually selective population decayed slowly during the memory delay, but not the visual delay, consistent with a cue encoding and early storage process (Fig. 2 *A*, *i*). In support of this interpretation, the number of spatially tuned units in this population decreased from 59 to 30 during the transition from the early to late memory delay, suggesting a weakening role in storage over time. In contrast, activity in the memory-selective population ramped up during the memory delay, but not the visual delay, consistent with a late

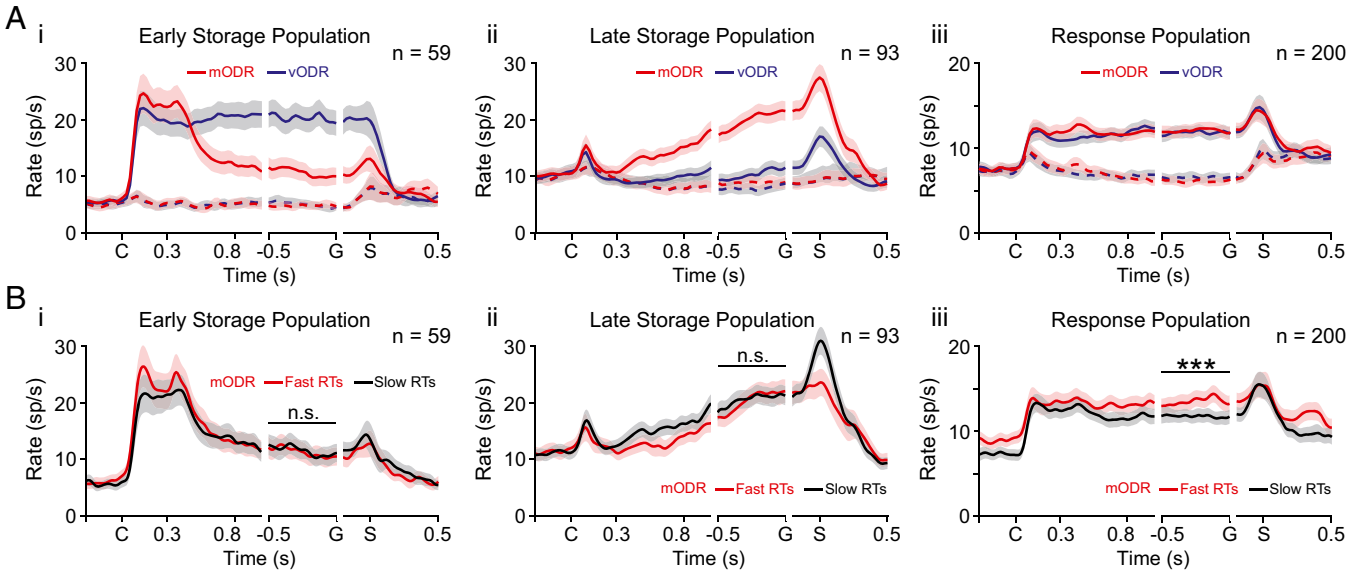

Fig. 2. Persistent neural activity reflects distinct storage and response modes. (*A*) Mean firing rate response of classified (*i*) early storage, (*ii*) late storage, and (*iii*) response populations during mODR (red) and vODR (blue) trials. Solid and dotted lines indicate responses to preferred and null targets, respectively. (*B*) Mean firing rate response of each population to preferred stimuli during memory trials, grouped by the fastest 50% (red) and slowest 50% (black) of reaction times after the Go command. In all panels, horizontal bars denote a permutation test over the difference in firing rates across conditions during the last 500 ms before the Go command. n.s., $P > 0.05$; ***$P < 0.001$ (permutation test).

storage process (Fig. 2 *A*, *ii*). The number of spatially tuned units in this population increased from 67 to 93 during the transition from the early to late memory delay, suggesting a strengthening role in storage over time. We confirmed that these two populations capture distinct storage-related modes of activity at the population level by performing a targeted dimensionality reduction analysis of all 746 recorded units (see *Experimental Procedures*). Targeted dimensionality reduction is a constrained extension of principal component analysis that seeks to discover the dimensions of variability in data that are most closely linked with task variables (in this case, target location and whether or not memory storage was required during the delay). This analysis revealed two modes of task-related persistent activity that exhibited slowly decaying and slowly ramping activity during the memory delay, consistent with early and late storage modes (*SI Appendix*, Fig. S4). Finally, neural activity persisted in the nontask-selective population during both delays, consistent with response preparation (Fig. 2 *A*, *iii*). These results demonstrate that persistently active units can be assigned to populations that broadly reflect three distinct modes of persistent activity in PFC, which may be linked to storage and response preparation.

If stored information and response preparation map to different modes of population activity, the second prediction made earlier is that the mode that encodes responses should encode the metrics of a prepared movement, whereas the two storage modes should not. If this is correct, response-related activity immediately before the Go command will covary with RT, whereas storage activity will not, as predicted by the distinct mode hypothesis. To test this prediction, we grouped memory trials by the fastest 50% and slowest 50% of saccades made by each monkey and then compared the preferred target responses of each neural population across RT conditions during the last 500 ms of the memory delay. We reasoned that activity during this presaccadic interval was most likely to reveal whether a population played a role in determining subsequent RTs. Spike rates did not differ across RT conditions in the early storage (Fig. 2 *B*, *i*) or late storage (Fig. 2 *B*, *ii*) populations ($P > 0.05$ permutation test). In contrast, delay activity in the response population was significantly lower during slow RT trials than during fast trials (Fig. 2 *B*, *iii*; $P < 0.001$ permutation test). These results hold for each animal individually. A further comparison of delay activity across correct trials and infrequent saccade error trials (2.6% of trials; 15.4° median angle between cue location and saccade endpoint) revealed that only the response population showed a significant drop in activity before inaccurate saccades (*SI Appendix*, Fig. S5). Therefore, information encoded in the response network directly influences the timing and accuracy of planned behavior. Together, these results demonstrate that storing information and generating a response involve distinct modes of persistent activity in PFC.

The identification of storage and response preparation with different groups of neurons raises the possibility that these networks could map to distinct anatomical regions. Therefore, we investigated whether these circuits were randomly organized in PFC or whether they were functionally localized with distinct spatial signatures. Specifically, we used our database of recordings from chronically implanted movable electrode arrays to map the spatial organization of persistently active neurons within a 3D cortical volume. Although neurons from each WM network were broadly distributed throughout the sampled volume, the resulting brain maps revealed localized functions with substantial differences in the topography of the late storage and response networks (Fig. 3 *A* and *B*). Neurons in the late storage network were concentrated in posterior PFC, proximal to area 8, whereas the response network was concentrated in anterior PFC, proximal to area 46. These unit classes occurred with significantly different likelihoods in the anterior half of the array (0.68 ± 0.04 SEM for response neurons vs. 0.25 ± 0.03 SEM for late storage neurons; $P < 0.05$, binomial exact test). Furthermore, late storage neurons were most likely to be recorded from superficial depths (Fig. 3 *C* and *D*; 0.37 ± 0.06 SEM at 0.1 mm vs. 0.22 ± 0.05 SEM at 1.0 mm; $P < 0.05$, binomial exact test), whereas response neurons were highly biased toward deeper sites (0.73 ± 0.07 SEM at 1.3 mm vs. 0.50 ± 0.06 at 0.1 mm; $P < 0.05$, binomial exact test). The structure of these anatomical maps does not change after excluding data from electrodes that were suspected to have descended down sulcal banks. Together, these findings demonstrate that the storage and response preparation modes are spatially organized and not randomly dispersed in PFC. We

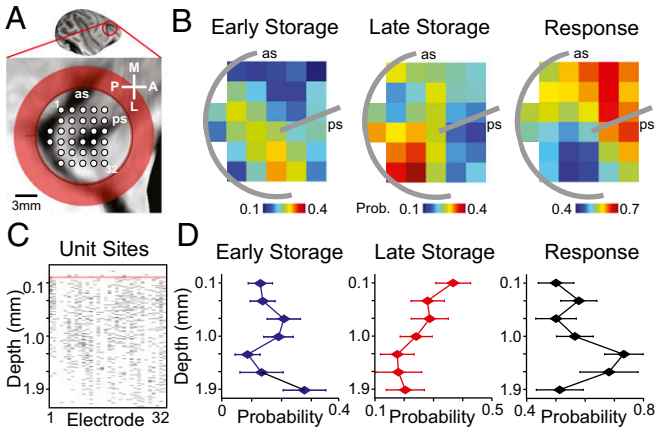

Fig. 3. Working memory modes are anatomically patterned. (*A*) MRI-guided stereotaxic chamber placement over PFC of monkey A. Red circle indicates the implanted chamber location (±1 mm positional error). White dots indicate electrode penetration sites. Sulcal landmarks: *as*, arcuate sulcus; *ps*, principal sulcus. Compass labels: *A*, anterior; *P*, posterior; *M*, medial; *L*, lateral. Channel numbers are indexed 1–32 from left to right. (*B*) Topographic distribution of units in each population, reported as a fraction of all classified units on each electrode, independent of depth. Histograms were smoothed with a 2D Gaussian kernel ($\sigma^2 = 1.5$ mm). All data pooled across two monkeys. (*C*) Map of recording depths at which spiking activity was observed on each electrode, pooled across two monkeys. Red line indicates registered zero cortical depth. (*D*) Depth distribution of units in each population, reported as a fraction of all classified WM units within each 300-μm depth window.

thus conclude that WM is supported by multiple anatomically specific modes of persistent activity.

The clear functional and anatomical differences between the storage and response populations suggest they make up distinct networks. However, these populations could also reflect different aspects of a single, functionally integrated, distributed network that uses long-range interactions to coordinate three anatomically specific modes of activity. We tested this hypothesis by looking for evidence that persistent activity is coordinated across columns separated by ≥1.5 mm within each of the three populations, as expected for components of a large network with globally coordinated activity. We quantified "coordination" as the trial-to-trial correlated variability ("noise correlation") between the firing rates of similarly tuned neurons that were recorded on different electrodes (see *Experimental Procedures*). We found that similarly tuned pairs within the early and late storage populations exhibited positive noise correlations during the memory delay and zero noise correlation during the visual delay (Fig. 4*A*), consistent with memory delay-specific coordination across columns. In contrast, similarly tuned neuron pairs in the response network exhibited zero noise correlation during both memory and visual delays, indicating that distributed coordination is unlikely to support persistence within this population. All correlations were independent of firing rate and were zero during the memory delay in both storage and response pairs with large spatial tuning differences (Fig. 4*B*). These results indicate that coordinated cross-columnar interactions can support persistence in the two storage populations, but cannot support persistence in the response population. Therefore, the storage and response populations reflect anatomically distinct networks that coordinate their activity over different spatial scales.

Although the generation of a contingent response from a stored memory is clearly required for the mODR task, the lack of a clear link between the firing rates of the storage and response networks makes it unclear whether or how the storage mode of population activity influences the memory-guided response. For example, if the response mode changes its functional significance across the visual and memory delays, such that the response network supports all maintenance and manipulation functions, then the late storage mode may not be involved in the generation of a memory-guided response. To study this problem, we tested the hypothesis that the timing of action potentials in all three populations is coordinated in relation to the timing of a prepared response, which is one possible signature of networks that collectively support memory-guided behavior (24). To quantify "coordination" across each population, we estimated spike-field coherence between spikes and local field potentials recorded on separate electrodes within each network during the last 1 s of the memory delay on preferred stimulus trials, and then compared spike-field coherence across fast and slow RT trials (see *Experimental Procedures*). In support of the temporal coordination hypothesis, spike-field coherence in the 14–30-Hz band was strongest during slow RT trials among all three populations of early storage (Fig. 4*C*), late storage (Fig. 4*D*), and response neurons (Fig. 4*E*) ($P < 0.05$, permutation test). This coherence

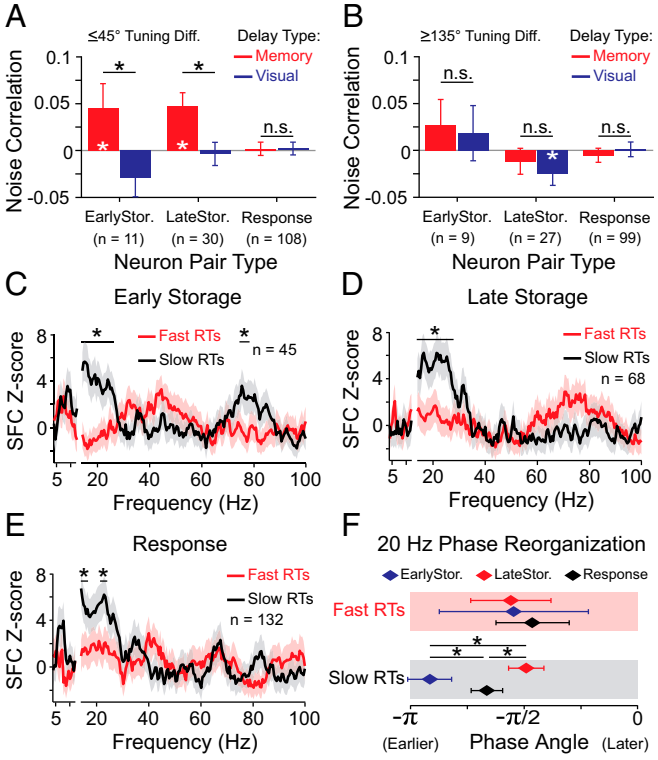

Fig. 4. Working memory networks exhibit distinct patterns of firing rate and spike timing coordination. (*A*) Correlated variability of firing rates across trials during the last 300 ms of the memory (red) and visual (blue) delay. Unit pairs were drawn from within the same population and had ≤45° difference in preferred target location. (*B*) Unit pairs with ≥135° difference in preferred target location. White asterisks indicate a significant difference from zero ($P < 0.05$; permutation test), and black asterisks indicate a significant difference across conditions ($P < 0.05$; permutation test). (*C–E*) Coherence between units from each network and fields recorded on a different electrode during the last 1 s of the memory delay on preferred target trials. Trials were divided by the fastest 50% (red) and slowest 50% (black) of reaction times after the Go command. The analysis used 10 Hz bandwidth at frequencies above 13 Hz and 4 Hz bandwidth at lower frequencies. This change in bandwidth is delineated by the gap along the frequency axis in each panel. *Z*-scores reflect deviations in raw coherence from a null distribution estimated after shuffling trials. (*C*) Early storage coherence. (*D*) Late storage coherence. (*E*) Response coherence. Asterisks indicate significant differences across conditions ($P < 0.05$, permutation test). (*F*) Coherence phase angle at 20 Hz for the preceding analyses during fast (*Top*) and slow (*Bottom*) RT trials. Asterisks indicate significant differences between networks ($P < 0.05$, permutation test).

NEUROSCIENCE

persisted throughout the delay (*SI Appendix*, Fig. S6) and was significantly weaker during fast RT trials, indicating a potential role for coherent activity in response selection. In further support of the hypothesis that multiple networks are coordinated during memory-guided behavior, neurons in all three populations displayed distinct patterns of temporal coordination during memory delays on slow RT trials. The preferred spike phases at 20 Hz were precisely organized (Fig. 4*F*), such that spikes occurred in a sequence of phases that were earliest in the early storage population and latest in the late storage population. This relative phase code establishes the involvement of the storage and response networks in memory-guided behavior and reveals that PFC maintains a strict temporal separation between the neural codes for stored information and prepared responses.

Discussion

Here we present several convergent lines of evidence to support models of WM that posit distinct component networks (1, 3, 9). Using a simple behavioral manipulation, we reveal a previously unknown distinction between storage- and response-related modes of persistent activity in the brain. The storage and response populations shown here display distinct functions, are patterned differently in the anatomical domain, and exhibit distinct patterns of firing rate and spike timing coordination during WM. Together, these findings demonstrate that anatomically specific modes of persistent activity correspond to distinct storage and response networks in PFC.

We identify the response network with a preparatory process because response population delay activity is linked with reaction time and saccade accuracy after the Go command. Both forms of behavioral variability are internally directed by the monkey in our experiments and are not instructed by external cues, making them good indicators of the monkey's internal state of preparation. This method of identifying internally directed responses during a prosaccade task is complementary to previous task designs that explicitly separated cue location and the direction of the impending response (25–27). Those studies pooled activity across the entire delay (without distinguishing between early or late delay epochs) and found that only 13–30% of persistently tuned units encode a prepared response. However, subsequent work has shown that the vector of PFC population activity gradually rotates during the delay period from the cue direction to the saccade direction (28), indicating that a single statistic may be insufficient to capture nonstationary responses in this area. Our work supports this view by demonstrating that a majority of units with late delay tuning are linked with response preparation (200/365 units, or 55%). It should be noted, however, that the response population identified here might not exclusively mediate response preparation and may also be involved in response selection. The response network could also mediate an additional component of the storage process that is indistinguishable from response-related activity, using our task design. Therefore, although we have shown that the early and late storage populations are distinct from the response network, the possibility remains that other aspects of the storage process are supported by the response network in PFC.

Furthermore, although we ascribe a memory storage process to the late storage network, in principle, this activity could also reflect a spatial attention process that in turn supports memory. In some cases, spatial WM may depend on the sustained allocation of attention to the spatial location of the memorized cue (29, 30). Although human fMRI studies disagree on whether attention mechanisms account for activity in PFC during WM maintenance (30, 31), at the cellular level, neurons in monkey PFC can encode stored locations and attended locations, and some even encode both (32). Although our study was not designed to distinguish between the storage and attention-related cognitive mechanisms, it firmly establishes a mode of population activity in PFC that encodes locations only when visual cues have been extinguished during the memory delay. Simple attention cannot explain this finding because one would expect that during the delay period, attention would be directed toward both the visible and memorized target. In contrast, theories of WM have always incorporated attention in one way or another (2). For instance, the storage of WM representations may be maintained through sustained attention to internal representations of the memoranda. Future work in this area is needed to clarify the role attention may play in WM processes.

In the anatomical domain, the storage and response networks are most prevalent at different topographic locations and depths of PFC. This functional topography is striking, in part because a previous study using the same task design and acute recording techniques did not find evidence for a specialized storage network in primate PFC (33). That study focused on recording sites that were mostly anterior to where we have observed the late storage network, however, indicating that a difference of only a few millimeters in recording location was critical to finding this network in PFC. Our results are also striking because the depth dependencies of the late storage and response networks are broadly consistent with tracing studies that have observed both highly recurrent cross-columnar interactions in superficial PFC (34), suggestive of a storage function, and projections from deep layers of PFC to subcortical areas involved in oculomotor control (35, 36), suggestive of a response function. Furthermore, the segregation of storage and response networks along the anteroposterior axis provides important physiologic support for previous human imaging results that suggest anterior regions of frontal cortex may be specialized for the preparation of goal-directed responses (10). A synthesis of these findings with previous studies of oculomotor control suggests that after the Go command is received, control of saccadic eye movements likely shifts from anterior to posterior locations in the frontal eye field (FEF), through the activity of classically identified "visuomovement" and "movement" neurons (37). We were unable to test this hypothesis in our data because we targeted gyral sites that were mostly anterior to FEF, and we did not perform microstimulation to identify FEF neurons to avoid damage to the recording electrodes and changes in how the measured responses varied with depth. Therefore, understanding how neurons in the distributed PFC storage and response preparation networks overlap with FEF is another important goal for future work.

In the temporal domain, the precisely organized phase lag between action potentials from the storage and response networks could implement a multiplexed code for memoranda and their contingent responses. The selective implementation of this code during trials with slow RTs suggests spike timing may directly influence the timing of response selection during the delay. More generally, this code demonstrates that the brain uses coherent activity to maintain a temporal separation between the neural codes for stored information and prepared responses during memory-guided behavior. This separation may be necessary when behavior depends on the relationship between both a past stimulus and a rule-based response, such as is required for memory-guided antisaccades. In this regard, phase coding in PFC may reflect a basic mechanism of flexible behavioral control during WM. Moreover, such a mechanism could be applied generically to organize distinct information streams in support of other "top-down" processes that evoke coherent activity, such as attention (38), motor coordination (39), and decision making (40).

The finding that memory-guided saccade errors are linked with a failure of response preparation and not storage (*SI Appendix*, Fig. S5) supports a long-suspected but previously unproven mode of WM failure in PFC (14, 17, 25). Our results build on the recent finding that saccade accuracy is parametrically linked with the level of persistent activity by spatially tuned neurons in PFC (19), through

our demonstration that inaccurate saccades are specifically linked with a change in response-related activity only. Furthermore, these results inform disease models by indicating that impairments of memory-guided behavior in neuropsychiatric disorders (41, 42) may arise from disrupted response preparation, rather than from disrupted storage, as commonly believed. This revised hypothesis is consistent with the physiological effects of antidopaminergic agents that are commonly used in the treatment of schizophrenia, because D1 antagonist administration evokes persistent activity in two classes of neurons: those that exhibit ramping responses, consistent with the late storage network, and those that exhibit plateau responses, consistent with the response network (43). Therefore, it is plausible that pharmacologic agents used in the treatment of WM disorders act primarily by counteracting degraded activity in the response network, while simultaneously enhancing activity in an otherwise functional late storage network. Future work will need to test the link between dopamine receptor type and WM component-specific activity in PFC.

Experimental Procedures

We obtained neuronal recordings from across the depths of the arcuate gyrus of two macaque monkeys (*Macaca mulatta*) while the animals performed either a memory-guided or a visually guided ODR task on randomly interleaved trials. We then analyzed the recordings to determine whether networks recruited by WM contain components with different functions. All surgical and animal care procedures were approved by the New York University Animal Care and Use Committee and were performed in accordance with the National Institutes of Health guidelines for care and use of laboratory animals (44). Detailed methods can be found in the *SI Appendix, Experimental Procedures*.

ACKNOWLEDGMENTS. We thank Gerardo Moreno for surgical assistance; Roch Comeau, Stephen Frey, and Brian Hynes for custom modifications to the BrainSight system; and members of the B.P. laboratory for helpful feedback. This work was supported in part by NIH Ruth L. Kirschstein National Service Award F32-MH100884 from the National Institute of Mental Health (NIMH) (to D.A.M.), a Swartz Fellowship in Theoretical Neurobiology (to D.A.M.), NIH Training Grant T32-EY007158 (to D.A.M.), National Eye Institute (NEI) R01-EY016407 (to C.E.C.), NIMH R03-MH097206 (to C.E.C.), National Science Foundation Faculty Early Career Development (CAREER) Award BCS-0955701 (to B.P.), and NEI R01-EY024067 (to B.P.).

1. Baddeley AD (1986) *Working Memory* (Clarendon, Oxford).
2. Miyake A, Shah P (1999) *Models of Working Memory*, eds Miyake A, Shah P (Cambridge University Press, New York).
3. Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* 24:167–202.
4. Cohen JD, et al. (1997) Temporal dynamics of brain activation during a working memory task. *Nature* 386(6625):604–608.
5. Courtney SM, Ungerleider LG, Keil K, Haxby JV (1997) Transient and sustained activity in a distributed neural system for human working memory. *Nature* 386(6625): 608–611.
6. Rowe JB, Toni I, Josephs O, Frackowiak RS, Passingham RE (2000) The prefrontal cortex: Response selection or maintenance within working memory? *Science* 288(5471):1656–1660.
7. Curtis CE, Rao VY, D'Esposito M (2004) Maintenance of spatial and motor codes during oculomotor delayed response tasks. *J Neurosci* 24(16):3944–3952.
8. Smith EE, Jonides J (1999) Storage and executive processes in the frontal lobes. *Science* 283(5408):1657–1661.
9. Curtis CE, D'Esposito M (2003) Persistent activity in the prefrontal cortex during working memory. *Trends Cogn Sci* 7(9):415–423.
10. Rowe JB, Passingham RE (2001) Working memory for location and time: Activity in prefrontal area 46 relates to selection rather than maintenance in memory. *Neuroimage* 14(1 Pt 1):77–86.
11. Volle E, et al. (2005) Specific cerebral networks for maintenance and response organization within working memory as evidenced by the 'double delay/double response' paradigm. *Cereb Cortex* 15(7):1064–1074.
12. Fernando ABP, Robbins TW (2011) Animal models of neuropsychiatric disorders. *Annu Rev Clin Psychol* 7:39–61.
13. Fuster JM, Alexander GE (1971) Neuron activity related to short-term memory. *Science* 173(3997):652–654.
14. Funahashi S, Bruce CJ, Goldman-Rakic PS (1989) Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *J Neurophysiol* 61(2):331–349.
15. Goldman-Rakic PS (1995) Cellular basis of working memory. *Neuron* 14(3):477–485.
16. Major G, Tank D (2004) Persistent neural activity: Prevalence and mechanisms. *Curr Opin Neurobiol* 14(6):675–684.
17. Fuster JM (1973) Unit activity in prefrontal cortex during delayed-response performance: Neuronal correlates of transient memory. *J Neurophysiol* 36(1):61–78.
18. Kennerley SW, Wallis JD (2009) Reward-dependent modulation of working memory in lateral prefrontal cortex. *J Neurosci* 29(10):3259–3270.
19. Wimmer K, Nykamp DQ, Constantinidis C, Compte A (2014) Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nat Neurosci* 17(3):431–439.
20. Markowitz DA, Wong YT, Gray CM, Pesaran B (2011) Optimizing the decoding of movement goals from local field potentials in macaque cortex. *J Neurosci* 31(50): 18412–18422.
21. Funahashi S, Bruce CJ, Goldman-Rakic PS (1993) Dorsolateral prefrontal lesions and oculomotor delayed-response performance: Evidence for mnemonic "scotomas". *J Neurosci* 13(4):1479–1497.
22. Compte A, Brunel N, Goldman-Rakic PS, Wang XJ (2000) Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. *Cereb Cortex* 10(9):910–923.
23. Crammond DJ, Kalaska JF (1996) Differential relation of discharge in primary motor cortex and premotor cortex to movements versus actively maintained postures during a reaching task. *Exp Brain Res* 108(1):45–61.
24. Pesaran B, Pezaris JS, Sahani M, Mitra PP, Andersen RA (2002) Temporal structure in neuronal activity during working memory in macaque parietal cortex. *Nat Neurosci* 5(8):805–811.
25. Niki H, Watanabe M (1976) Prefrontal unit activity and delayed response: Relation to cue location versus direction of response. *Brain Res* 105(1):79–88.
26. Funahashi S, Chafee MV, Goldman-Rakic PS (1993) Prefrontal neuronal activity in rhesus monkeys performing a delayed anti-saccade task. *Nature* 365(6448):753–756.
27. Takeda K, Funahashi S (2002) Prefrontal task-related activity representing visual cue location or saccade direction in spatial working memory tasks. *J Neurophysiol* 87(1): 567–588.
28. Takeda K, Funahashi S (2004) Population vector analysis of primate prefrontal activity during spatial working memory. *Cereb Cortex* 14(12):1328–1339.
29. Awh E, Jonides J (2001) Overlapping mechanisms of attention and spatial working memory. *Trends Cogn Sci* 5(3):119–126.
30. Jerde TA, Merriam EP, Riggall AC, Hedges JH, Curtis CE (2012) Prioritized maps of space in human frontoparietal cortex. *J Neurosci* 32(48):17382–17390.
31. Postle BR, Awh E, Jonides J, Smith EE, D'Esposito M (2004) The where and how of attention-based rehearsal in spatial working memory. *Brain Res Cogn Brain Res* 20(2): 194–205.
32. Messinger A, Lebedev MA, Kralik JD, Wise SP (2009) Multitasking of attention and memory functions in the primate prefrontal cortex. *J Neurosci* 29(17):5640–5653.
33. Tsujimoto S, Sawaguchi T (2004) Properties of delay-period neuronal activity in the primate prefrontal cortex during memory- and sensory-guided saccade tasks. *Eur J Neurosci* 19(2):447–457.
34. Kritzer MF, Goldman-Rakic PS (1995) Intrinsic circuit organization of the major layers and sublayers of the dorsolateral prefrontal cortex in the rhesus monkey. *J Comp Neurol* 359(1):131–143.
35. Giguere M, Goldman-Rakic PS (1988) Mediodorsal nucleus: Areal, laminar, and tangential distribution of afferents and efferents in the frontal lobe of rhesus monkeys. *J Comp Neurol* 277(2):195–213.
36. Stanton GB, Goldberg ME, Bruce CJ (1988) Frontal eye field efferents in the macaque monkey: I. Subcortical pathways and topography of striatal and thalamic terminal fields. *J Comp Neurol* 271(4):473–492.
37. Bruce CJ, Goldberg ME (1985) Primate frontal eye fields. I. Single neurons discharging before saccades. *J Neurophysiol* 53(3):603–635.
38. Buschman TJ, Miller EK (2007) Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. *Science* 315(5820):1860–1862.
39. Dean HL, Hagan MA, Pesaran B (2012) Only coherent spiking in posterior parietal cortex coordinates looking and reaching. *Neuron* 73(4):829–841.
40. Pesaran B, Nelson MJ, Andersen RA (2008) Free choice activates a decision circuit between frontal and parietal cortex. *Nature* 453(7193):406–409.
41. Park S, Holzman PS (1992) Schizophrenics show spatial working memory deficits. *Arch Gen Psychiatry* 49(12):975–982.
42. Dowson JH, et al. (2004) Impaired spatial working memory in adults with attention-deficit/hyperactivity disorder: Comparisons with performance in adults with borderline personality disorder and in control subjects. *Acta Psychiatr Scand* 110(1):45–54.
43. Williams GV, Goldman-Rakic PS (1995) Modulation of memory fields by dopamine D1 receptors in prefrontal cortex. *Nature* 376(6541):572–575.
44. Committee on Care and Use of Laboratory Animals (1996) *Guide for the Care and Use of Laboratory Animals* (Natl Inst Health, Bethesda), DHHS Publ No (NIH) 85-23.