Neural bases of space-specific distractor biases in visual working memory

Authors & Affiliations:

Deepak V Raya[1], Sanchit Gupta[1], and Devarajan Sridharan[1,2,*]

[1]Centre for Neuroscience, Indian Institute of Science, Bangalore 560012, India

[2]Computer Science and Automation, Indian Institute of Science, Bangalore 560012, India

***Correspondence to:**

Devarajan Sridharan,
GS Bhat Chair Associate Professor in Digital Health,
DST SwarnaJayanti Fellow,
Centre for Neuroscience,
Indian Institute of Science,
C. V. Raman Avenue,
Bangalore 560012, India
Phone: +91 802 293 3434
E-mail: sridhar@iisc.ac.in

Abstract

Information held in working memory (WM) is remarkably resilient to distraction. Yet, perceptual distractors that share mnemonic features can impact WM profoundly; the neural basis of this phenomenon remains unclear. With multivariate decoding of human electroencephalography recordings, we investigate how delay-period perceptual distractors bias WM. Participants memorized the orientations of cued and uncued grating memoranda that appeared in opposite hemifields. A grating distractor, flashed during the delay period, produces space-specific biases: memorized features are attracted towards or repelled away from the distractor's orientation depending, respectively, on when the distractor appeared in the same hemifield as the memorandum, or opposite to it. Neural prioritization in WM by cueing, and stronger memorandum maintenance mitigate this bias, whereas stronger distractor encoding enhances it. Lastly, a ring-attractor model with cross-hemifield inhibition mechanistically explains the origins of these spatially-antagonistic biases. Our results reveal how lateralized sensory buffers critically enable perceptual distractors to bias visual WM.

Keywords: mnemonic representations; perceptual distractors; attractive and repulsive biases; electroencephalography; multivariate decoding; ring attractor model.

Lay Summary

Working memory (WM) – the ability to momentarily store important items – is remarkably robust to distraction. Yet, when a salient object with features resembling the memorized item ("perceptual distractor") appears in the environment, it can alter WM appreciably. What neural mechanisms render WM resilient to distraction, and how can perceptual distractors affect it so profoundly? We address this question by presenting a salient distractor, at unpredictable times, during the WM delay-period. Surprisingly, the distractor's bias on WM depends on its proximity to the memorandum's original location. With state-of-the-art neural decoding and computational modeling, we identify mechanisms that mediate these space-specific distractor effects. The findings advance our understanding of WM's resilience to distraction and may inform cognitive therapies for treating WM deficits.

Introduction

Working memory (WM) is a remarkable cognitive ability that enables us to transiently hold and manipulate information that is no longer present in the environment. Mnemonic information in WM must be maintained in the absence of external sensory input, and it is increasingly evident that early sensory areas are actively involved in such maintenance[1–5]. As a result, mnemonic information may be susceptible to interference by new sensory inputs, such as salient "perceptual" distractors[5–7]. Yet, a significant body of literature suggests that information in WM is surprisingly resilient to distraction[5,6,8,9]. For example, behavioral performance in a WM task involving oriented gratings was robust, despite multiple task-irrelevant (face or place) distractors being presented during the delay period[6]. Similar results have been observed in many recent studies of visual WM[5,10].

Despite this robustness, recent evidence suggests that perceptual distractors may interfere systematically, with the content of WM[5,8,10,11]. In particular, some distractors are more potent than others at "biasing" the content of working memory, especially during the delay-period [8,9,11–16]. The extent to which a distractor's features match task-relevant features of the memorandum determines the strength of its bias, with greater feature similarity yielding greater bias in WM[8,15]. Typically, such biases are "attractive" in that the memorandum's report is biased towards the distractor's features[8,10,11,17]. For instance, viewing a task-irrelevant orientation during the WM delay period biases the remembered target's orientation towards the distractor's orientation[10,17]. Similarly, spatial distractors can bias the remembered location towards the distractor's location[11]. Such perceptual distractor biases have been reported even for complex features. For example, remembered face representations are more strongly biased by the presentation of face image distractors than by scene image distractors[15].

Identifying mechanisms that mediate distractor interference, including perceptual biases, in WM representations is a topic of active research[5,8–11]. Previous studies measured brain responses during the WM delay epoch, and have found evidence for a distributed representation of mnemonic information in diverse brain regions[18], including sensory (visual) cortex[1–5], posterior parietal cortex (PPC, [10,19,20]), the intraparietal sulcus (IPS, [6,10,21]), as well as in the prefrontal cortex (PFC, [7,21,22]). Converging evidence from these studies suggests that while mnemonic information in sensory cortex may be susceptible to distractor interference and bias, information in higher cortical areas, including PPC[6,10] and PFC[9,23], is much more robust to such interference. For example, when a distractor grating with a random orientation was presented during the delay period of a working memory task, it produced biases in neural representations in the early visual cortex (V1-V3), but not in the IPS[10]. A related study showed that the distractor's orientation affected neural representations in early visual cortex (V1-V4), but higher cortical regions, including the IPS and lateral occipital cortex, were unaffected[5]. In fact, such stability of representations in higher cortical areas is hypothesized to mediate the remarkable resilience of WM to sensory distractors [8,24].

Despite this emerging literature, a key question remains unanswered: what neural mechanisms govern how a perceptual distractor affects the contents of working memory during maintenance ("delay-period perceptual distraction"[25])? Previous studies on distractor effects overwhelmingly employed simple WM task designs: for example, most tasks involved a single, central distractor presented at fixed times during the delay epoch[5,8–12,14–16,26]. Moreover, many more previous studies investigating mnemonic bias employed functional magnetic resonance imaging (fMRI)[5,10,11] to study WM biases in the brain. The poor temporal resolution of fMRI precludes addressing fundamental questions about rapid, distractor timing effects: Because the memorandum and distractor occur within a few hundred milliseconds of each other, their evoked blood oxygenation level dependent (BOLD) responses evolve over the timescale of several

seconds and are challenging to disentangle. No previous study, to our knowledge, has explored distractor-induced bias effects by varying the distractor's location and its timing relative to the memorandum, systematically.

Here, we addressed this question with multivariate stimulus decoding of high-density human electroencephalography (EEG) recordings, which enabled measuring distractor's neural effects with millisecond precision. We characterized how perceptual distractors induce mnemonic biases in participants (n=24) performing a continuous response, visual WM task involving memorandum orientation recall (Fig. 1A). We show that a task-irrelevant distractor grating presented during the delay period produces either an attractive or a repulsive bias in features stored in WM depending on the distractor's proximity to the memorandum's location. With state-of-the-art neural decoding, we show that robust maintenance and input gating strongly shield mnemonic information from distractor interference. Lastly, we explain the mechanistic origin of these biases with a two-tier ring attractor model[27,28].

The results show how lateralized sensory buffers could provide a conduit by which task-irrelevant sensory input that is perceptually similar to remembered items could gain access to WM and bias its mnemonic content. Our study also provides a general computational framework to study the resilience of visual working memory to perceptual distractor effects.

Results

Distractors induce space-specific biases in visual working memory

We tested n=24 human participants on a visual WM task involving grating orientation report (Fig. 1A, Methods). Following fixation, two randomly oriented gratings ("memory array") were flashed briefly (100 ms) one in each visual hemifield. A brief interval (500 ms) after grating offset, a spatial retro-cue (filled white semi-circle) appeared, which indicated the hemifield of the erstwhile grating that would likely be probed for response (cue validity: 70%). A fixed epoch (2450 ms) after the memory array offset, a probe appeared (central filled circle) whose color (white or black) indicated the grating (cued or uncued, respectively) whose orientation the participants must reproduce (Methods).

To study the effect of perceptual distractors in visual WM, a singleton, distractor grating – randomly oriented but otherwise identical to the stimuli in the memory array (Methods) – was flashed briefly (100 ms) at a random time (500-1250 ms, exponentially distributed) during the delay period after retro-cue offset. The distractor was presented in 80% of the trials – interleaved pseudorandomly with 20% "no distractor" trials – and appeared either on the left or the right hemifield with equal probability. We refer to the retro-cued hemifield, the grating at this location or its representation in WM as the "cued" hemifield, grating, or representation, respectively; their counterparts in the opposite hemifield are referred to as the "uncued" hemifield, grating or representation, respectively.

First, we tested if participants were indeed utilizing the retro-cue for selection in visual WM, including only trials in which no distractor appeared (20% of all trials). The error in orientation reports, as measured by the circular standard deviation (CSD) of the orientation report error, was significantly lower on cued trials compared to uncued trials (CSD_{Cued}=23.4°±1.6° [mean±s.e.m.];

CSD_{Uncued}=35.2°±2.2°, p<0.001, Wilcoxon signed rank test, Bayes Factor (BF_{10}) >10^2; n=24 participants) (Fig. 1B), confirming robust selection of the retro-cued representation in visual WM.

Next, we analyzed the effect of the perceptual distractor on the cued and uncued gratings' orientation reports including all trials (80%) in which the distractor was presented. We tested for spatially lateralized effects; specifically, we hypothesized that the precision of orientation reports would be different depending on whether the distractor was presented in the same ("distractor-same" condition) or on the opposite hemifield ("distractor-opposite" condition) as the probed stimulus. We found robust evidence confirming this hypothesis: a 2-way ANOVA with the CSD of orientation errors as the dependent variable revealed a significant main effect of distractor location ($F_{(1,23)}$=11.06, p<0.001), but no significant interaction effect between cueing and distractor location ($F_{(1,23)}$=0.48, p=0.625). Post-hoc analyses revealed a significant increase in error for distractor-same compared to no-distractor trials (Fig. 1C, CSD_{NoDist}=29.3°±1.7°, $CSD_{DistSame}$=32.0°±1.9°, p=0.002, BF=40.34, signed rank), but not for distractor-opposite compared to no-distractor trials (Fig. 1C, CSD_{NoDist}=29.3°±1.7°, $CSD_{DistOpp}$=30.3°±1.7°, p=0.103, BF=1.16, signed rank); data pooled across cued and uncued locations).

Lastly, we analyzed the degradation in the precision of orientation reports across the cued and the uncued locations, induced by the distractor. For this, we pooled the orientation reports across all participants and fit a von Mises and uniform mixture model[29] to behavioral reports, separately for distractor-same and distractor-opposite trials; the degradation in the accuracy of orientation reports induced by the distractor was quantified with the difference in precision ($\Delta\kappa$) across these two conditions ($\kappa_{DistSame}-\kappa_{DistOpp}$) (Methods). $\Delta\kappa$ was significantly greater on the uncued hemifield ($\Delta\kappa_{uncued}$= -1.09±0.075 than the cued hemifield ($\Delta\kappa_{cued}$=-0.06±0.057, p=0.002, permutation test) (Fig. 1D). Essentially identical results were obtained using the modulation index of the precision, quantified as MI-κ = ($\kappa_{DistSame}-\kappa_{DistOpp}$)/ ($\kappa_{DistSame}+\kappa_{DistOpp}$) (p=0.003, permutation test).

In other words, the distractor affected the precision of orientation reports for the memorandum when it occurred, subsequently, in the same hemifield, and this effect was stronger for the uncued memorandum, as compared to the cued memorandum.

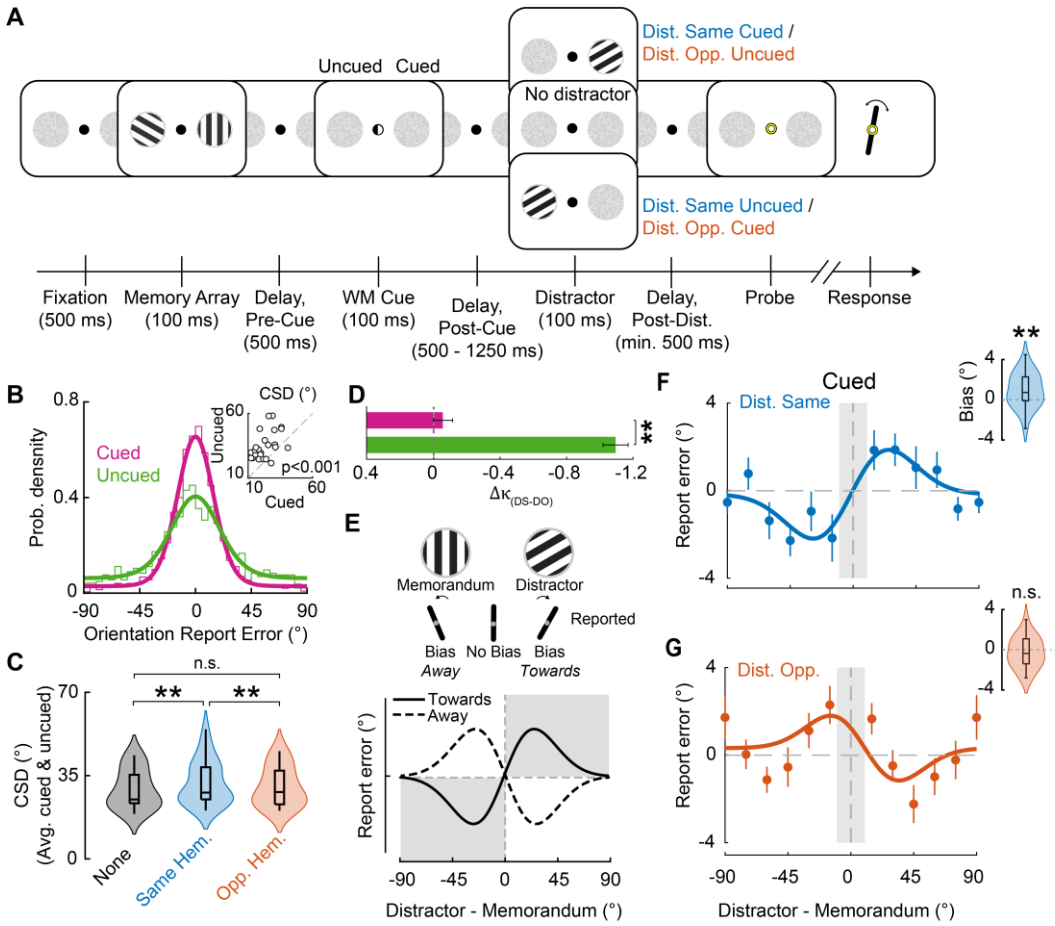


Figure 1. Perceptual distractor produces antagonistic biases during recall.

A. Visual working memory task with a distractor. Two randomly oriented gratings were flashed briefly, one in each hemifield. A retro-cue (white semicircle) indicated the location of the erstwhile grating that would likely be probed for response (70% validity). Following a variable delay-period, on 80% of the trials, a singleton perceptual distractor grating was flashed either on the right or left hemifield (upper and lower rows); on the remaining trials (interleaved) no distractor appeared (middle row). At the end of the trial, a central response probe (white or black) indicated the location of the grating (cued or uncued, respectively) to be recalled and reported.

B. Distribution (thin histogram) of the orientation report errors for the cued (magenta) and uncued (green) trials (n=24, data pooled across participants). Thick lines: von Mises fits. (*Inset*) Circular standard deviation (CSD) of the report errors for "no distractor" trials, for cued (x-axis) versus uncued (y-axis) grating reports. Dashed line: line of equality. Data points: individual participants.

C. CSD for no distractor trials (black), when the distractor appeared in the same (blue), or opposite (orange), hemifield as the probed memorandum. Violin plots: kernel density estimates; center line: median; box limits: upper and lower quartiles; whiskers: 1.5x interquartile range.

D. Change in precision between the distractor-same and distractor-opposite trials ($\Delta\kappa=\kappa_{DistSame}-\kappa_{DistOpp}$) for the cued (magenta) and uncued (green) trials. Error bars: Jackknife.

E. Schematic of bias. (*top*) The distractor (left grating) may bias the memorandum's (right grating) recalled orientation away from its own (left bar), towards its own (right bar), or not at all (middle bar). (*bottom*) Error in the probed memorandum's orientation (signed) (y-axis) as a function of its orientation relative to the distractor (x-axis). Positive and negative values (both axes): clockwise and counter-clockwise relative orientations, respectively. Solid and dashed curves: attractive bias (toward), or repulsive bias (away from), the distractor's orientation, respectively.

F. Average behavioral bias (n=24) when the distractor appeared in the same hemifield as the cued memorandum. Solid line: Derivative of Gaussian fit (Methods). Data smoothed over three successive orientation bins for visualization purposes only. Error bars: s.e.m. Gray shaded rectangle: untested relative orientation values (±10°). (*Inset, top right*) Violin plot: bias strengths across participants; positive (negative) values denote attractive (repulsive) bias. Other conventions are the same as in panel C.

G. Same as in panel F, but for trials in which the distractor appeared in the hemifield opposite to the cued memorandum. Other conventions are the same as in panel F. (All panels) *: $p<0.05$; **: $p<0.01$; n.s.: not significant.

It is possible that the increase in orientation report error was merely a consequence of a generic disruption of WM maintenance by the perceptual distractor – a salient, flashed stimulus. Alternatively, it is possible that the increase in error arose from a specific bias in WM during the delay induced by the distractor's orientation. To distinguish between these possibilities, we tested whether the behavioral orientation report of the probed memorandum was biased by the distractor's orientation, by plotting the signed orientation report error for the probed grating as a function of the difference between the distractor and the probed gratings' orientation; bias was quantified with the signed area-under-the-curve metric[28] (Fig. 1E) (Methods).

Orientation reports for the cued memorandum were systematically biased towards the perceptual distractor's orientation ("attractive" bias) when the latter appeared in the same hemifield as the former (Fig. 1F) (Cued: $Bias_{DistSame}$=0.94°±0.35°, p=0.004, BF_{+0}=7.50). By contrast, when the distractor appeared in the opposite hemifield, orientation reports revealed a trend of being biased away ("repulsive" bias) from the distractor's orientation (Fig. 1G) (Cued: $Bias_{DistOpp}$=-0.3°±0.3°, p=0.177, BF_{-0}=0.53). Although the repulsive bias effect was not statistically significant when data from all trials were pooled, in subsequent sections we show that when subsets of trials are selected based on objective criteria (e.g., distractor encoding strength) the pattern of repulsive bias becomes palpable, and significant (see next; e.g., Fig. 3 C-D).

A similar trend in attractive bias was apparent for the uncued grating also when the distractor appeared on the same hemifield as the memorandum (SI Fig. S1). Yet, both bias trends on the uncued hemifield were not statistically significantly different from zero (Uncued: $Bias_{DistSame}$=1.76°±1.15°, p=0.067, BF_{+0}=1.12; $Bias_{DistOpp}$=0.98°±1.07°, p=0.818, BF_{-0}=0.115), possibly because of the comparatively fewer trials (30%) in which uncued gratings were probed. Given the substantially greater proportion of trials probed on the cued hemifield that led to robustly measurable bias effects, we limit subsequent analyses of distractor-induced behavioral biases to cued-hemifield probed trials.

To test for the possibility that the behavioral bias arose from the observers mis-reporting the distractor's orientation as that of the probed memorandum's (target) on a subset of trials, we fit a model (von Mises) comprising a mixture of target and distractor orientations, as well as uniform noise, to their responses ("target+distractor" model, Methods) (SI Fig. S2A). We then compared this model against a more parsimonious model that did not incorporate the distractor orientations in the mixture ("target alone" model) (SI Fig. S2B). Models were fit for each participant individually, and formal model comparison was performed with the corrected Akaike information criterion (AICc) and Bayesian Information Criterion (BIC) that balance model complexity with goodness-of-fit[31,32]; a lower value of each metric indicates the preferred (selected) model.

AICc was significantly higher for the "target+distractor" model than for the "target alone" model (AICc: target + distractor =764.98±34.78, mean±s.e.m. across participants, target alone = 760.91±34.94, $p<0.001$, Wilcoxon signed-rank test) (SI Fig. S2C). These results were exactly replicated also with the BIC metric (BIC: target + distractor = 789.48±34.785, target alone = 773.26.66±34.94, $p<0.001$) (SI Fig. S2C). Moreover, examining the mixture proportions revealed that the mixture probability associated with the distractor component was substantially and significantly smaller (~1/12*th*, $p<0.001$) than the target component in the "target + distractor" model (SI Fig. S2D). We repeated these analyses with a mixture model that did not include the random component in either the "target + distractor" or "target alone" models (von Mises alone, see Methods) and observed an essentially identical pattern of results.

In sum, a perceptual distractor presented during the delay period systematically disrupted working memory, with cued memoranda being better shielded from interference as compared to uncued memoranda. Moreover, the distractor induced antagonistic biases – attractive and repulsive – when it appeared in the same versus opposite hemifields, respectively, as the cued memorandum, during the delay-period. This space-specific pattern of biases suggests that maintenance in working memory is mediated, at least in part, by a spatially lateralized,

240 hemisphere-specific process, that are susceptible to interference by a perceptual distractor. To
241 further pinpoint neural mechanisms underlying this bias, we analyzed neural representations of
242 the memory array and distractor with multivariate feature decoding based on the
243 electrophysiological recordings.

244

245 **Fidelity of target and distractor neural representations govern mnemonic bias**

246 To better understand the neural basis of the perceptual distractor bias, we computed the
247 association between the neural representations of the target memoranda and the distractor, and
248 the behavioral bias. Specifically, we hypothesized that the strength of encoding and maintenance
249 of the memorandum would render it resilient to distractor bias whereas, conversely, the strength
250 of encoding of the distractor would render it more potent at biasing responses.

251 For this, first, we performed trial-wise decoding of the memoranda orientations from time-
252 resolved EEG activity using a multivariate decoder based on the pair-wise Mahalanobis
253 distance[29,30]. The orientation of both memory array gratings could be decoded robustly from the
254 (respective) contralateral electrodes shortly following their onset (Fig. 2A, 2B, left) ($p<0.001$,
255 cluster-forming threshold $p<0.05$; $n=23$ participants for whom EEG data was available, see
256 Methods). In addition, the orientations of both the cued and the uncued memoranda could be
257 decoded for a few hundred milliseconds following the memory array onset (Fig. 2B, right; earliest
258 distractor onset: 1.2 s), with the cued memorandum exhibiting significant decodability later into
259 the maintenance epoch as compared to the uncued memorandum (Fig. 2B, inset, violin plots).

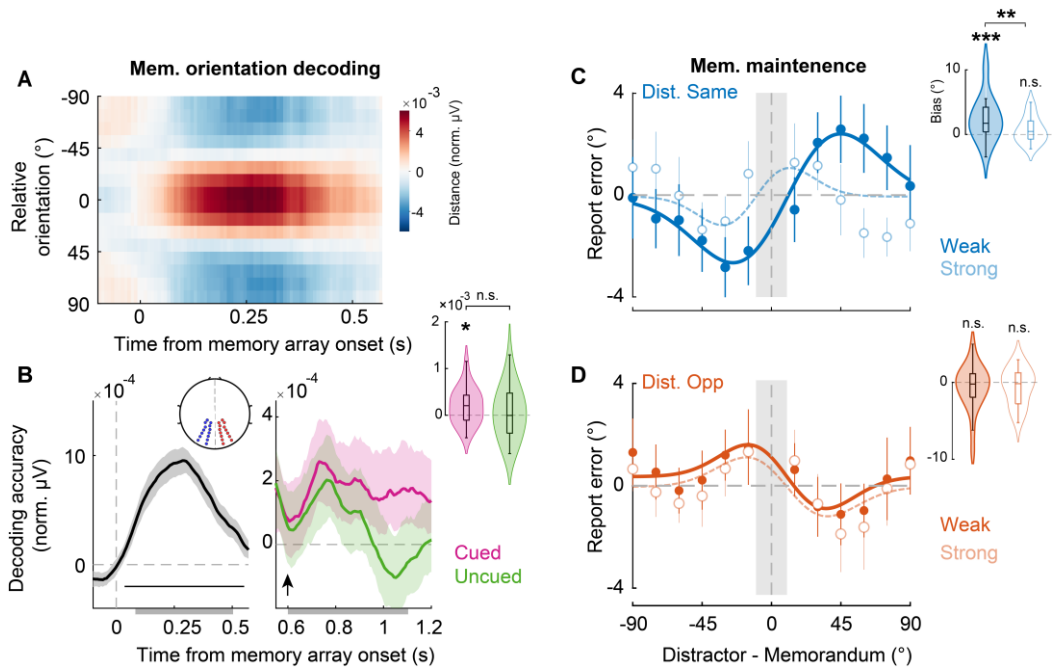


Figure 2. Stronger mnemonic neural representation mitigates distractor-induced bias

A. Orientation tuning map for memory array decoding for each time point relative to memory array onset. The tuning is computed as the similarity (sign reversed, mean-centered Mahalanobis distance) between the neural activity of the test trial and reference trials for stimuli at different relative orientations (y-axis). (All panels) Unless indicated otherwise, data averaged across participants (n=23, EEG) and for the left and the right gratings.

B. (*Left*) Memoranda decoding accuracy (dot product of tuning curve with cosine, normalized uV, Methods) for each time point locked to the memory array onset (dashed vertical bar). Horizontal black bar: significant decoding epochs (cluster-based permutation test, Methods). Shading: s.e.m. Profiles were smoothed with an 80 ms rectangular moving window for clarity of visualization. *(Inset)* Posterior electrodes used for the decoding analysis, contralateral to the respective grating. (*Right*) Same as in the left panel but showing decoding accuracy separately for cued (magenta) and uncued (green) gratings following retro-cue onset (arrowhead). (*Inset*) Violin plots showing the distribution of average decoding accuracy across participants (n=23) for cued and uncued gratings following the cue onset until the earliest distractor onset (1.2s).

C. Behavioral bias curves for the cued memorandum, but plotted based on a median split of memorandum maintenance strength – quantified with neural decoding accuracy during the early delay period (gray shaded bar, panel B, right) – when the distractor appeared on the same hemifield as the memorandum (n=23). Solid curve, filled symbols and filled violin plot (*inset*): Weaker maintenance trials. Dashed curve, open symbols and open violin plot (*inset*). Stronger maintenance trials. Other conventions are the same as in Figure 1F.

D. Same as in C, but when the distractor appeared on the hemifield opposite to the cued memorandum. Other conventions are the same as in panel C and Figure 1G. (All panels) *: $p<0.05$; **: $p<0.01$; ***:$p<0.001$; n.s.: not significant.

Next, to test our hypotheses regarding the association between the fidelity of these neural representations and the distractor bias, we performed two broad types of median split analyses.

First, we performed a median split of trials into weak and strong "maintenance" – based on the strength of cued memorandum's decoding accuracy (Methods) in a 500 ms window in the the delay period – beginning with cue onset, until 100 ms preceding the earliest distractor onset (Fig. 2B, right, gray shaded bar). We observed strong evidence for attractive bias – towards the distractor's orientation – on the trials with weak memorandum maintenance, when the distractor appeared on the same hemifield as the cued memorandum (weak maint.: $Bias_{DistSame}=2.40\pm0.71$, $p<0.001$, $BF_{+0}=29.5$). By contrast, trials with strong memorandum maintenance exhibited no statistically significant evidence for an attractive distractor-induced bias (strong maint.: $Bias_{DistSame}=0.69\pm0.43$, $p=0.062$, $BF_{+0}=1.26$). Moreover, the attractive bias for the weak memorandum maintenance trials was significantly higher as compared to that for the strong maintenance trials (weak vs. strong maint.: $p=0.009$, $BF_{+0}=5.09$, Fig . 2C). We also performed the same median split analysis for when the distractor appeared on the hemifield opposite to the cued memorandum, but did not find significant evidence for a repulsive bias modulation with maintenance strength ($p>0.05$ for all tests, Fig. 2D).

Additionally, we tested if the “encoding” strength of the cued memorandum also predicted a similar resilience to distractor bias. For this, we performed a median split analysis based on the decodability of the memorandum on the cued hemifield, but before the cue had appeared, in a window from 100 ms to 500 ms, following memory array onset (Fig. 2B, left, gray shaded bar). In contrast to maintenance strength, encoding strength did not modulate distractor-induced bias significantly. Specifically, we observed no statistically significant difference of cued memorandum encoding strength on bias either when the distractor appeared on the same hemifield as the cued memorandum (weak vs. strong enc.: $p=0.78$, $BF_{+0}=0.132$), or when it appeared on the opposite hemifield ($p=0.506$, $BF_{-0}=0.39$) (SI Fig. S3). Lastly, we tested if the distractor-induced bias for the cued memorandum would be influenced by the maintenance strength of the uncued memorandum; the latter was quantified, again, based on its neural decodability in a 500 ms window following cue onset (Fig. 2B, right). In this case also, we observed no significant effects of uncued memorandum maintenance strength on the bias either when the distractor appeared on the same hemifield (uncued-weak vs. strong maint.: $p=0.763$, $BF_{+0}=0.11$), or on the opposite hemifield ($p=0.052$, $BF_{-0}=1.52$), as the cued memorandum.

In other words, the robustness of maintenance, but not the strength of encoding, provided a clear predictor of susceptibility to distractor-induced bias, an effect that was strongly evidenced particularly when the distractor appeared on the same hemifield as the cued memorandum.

Second, we tested whether the strength of the perceptual distractor’s encoding influences the behavioral bias. For this, first, we decoded the orientation of the distractor from EEG traces locked to its onset. The distractor’s orientation could be decoded reliably for a few 100 ms, following its onset ($p<0.001$, cluster-forming threshold $p<0.05$) (Fig. 3A-B). Moreover, the strength of distractor encoding was not significantly different between strong and weak cued memorandum maintenance trials ($p>0.4$; Methods). As a result, any effects of distractor encoding strength on

behavioral bias cannot be trivially explained by the aforementioned effects of cued memorandum maintenance.

As before, for this analysis, we divided trials based on distractor's orientation decoding accuracy in a window from 100-400 ms following its onset (Fig. 3B, gray shaded bar); we term these the "weak" and "strong" distractor encoding trials. When the distractor appeared in the same hemifield as the cued memorandum, distractor-induced bias was significantly attractive for strong distractor encoding trials (strong dist. enc.: behavioral $Bias_{DistSame}$=1.79°±0.74°, p=0.006, BF_{+0}=4.70. By contrast, this bias was not significantly different from zero for weak distractor encoding trials (weak dist. enc.: behavioral $Bias_{DistSame}$=0.96°±0.75°, p=0.110, BF_{+0}=0.80) (Fig. 3C). Similarly, when the distractor appeared in the hemifield opposite to the cued memorandum the distractor-induced bias was significantly repulsive on strong distractor encoding trials ($Bias_{DistOpp}$=-1.1±0.53, p=0.022, BF_{+0}=2.54, but not significantly different from zero for weak distractor encoding trials ($Bias_{DistOpp}$=0.15±0.78, p=0.6, BF_{-0}=0.189) (Fig. 3D). Even though these effects were individually significant, we did not observer a significant difference between the strong and weak distractor encoding trials for either relative location of the distractors (same or opposite hemifield as the cued memorandum) (Fig. 3C-D, insets); in this case, bias estimates were perhaps rendered noisier with only half of the number of trials within each split. Nevertheless, distractor encoding strength clearly modulated the magnitude of the antagonistic biases observed in behavior.

Taken together, these results reveal a mechanistic neural basis of perceptual distractor biases in visual WM. Stronger maintenance of the cued memorandum, as quantified by neural decodability during the delay period, produced weaker distractor-induced biases in behavioral reports. Interestingly, this bias was unaffected by memorandum encoding strength. Conversely, stronger encoding of the distractor produced stronger, antagonistic biases. We sought to identify additional task and neural variables that controlled the magnitude of the distractor bias.

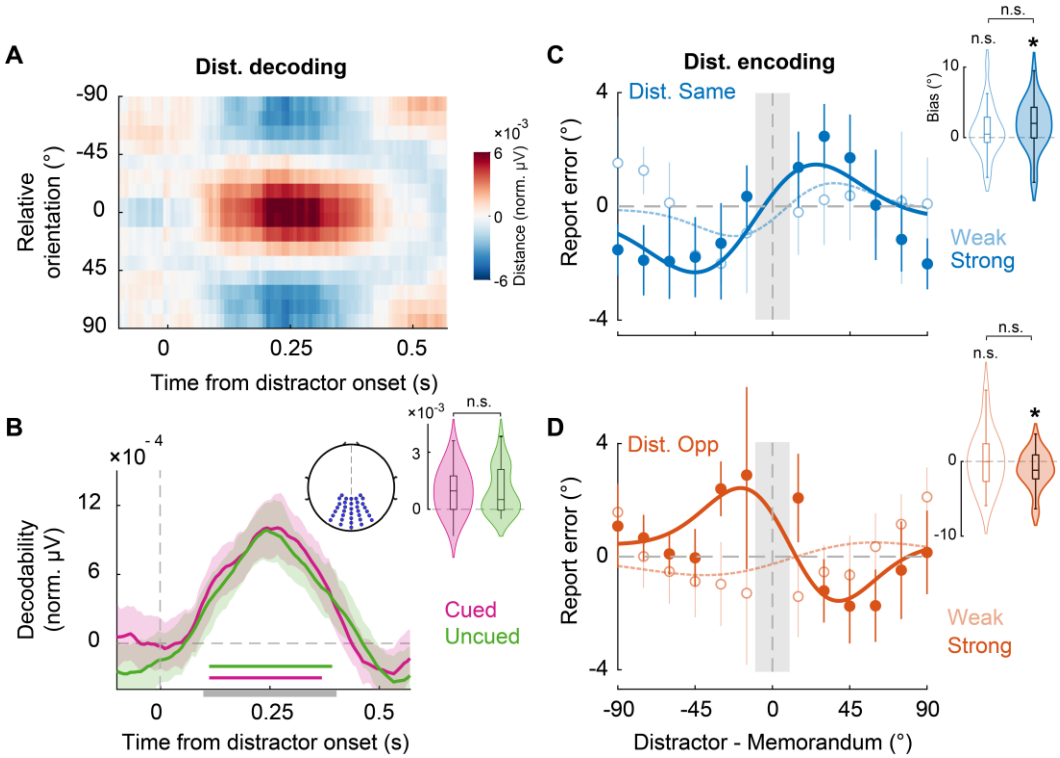


Figure 3. Stronger distractor encoding enhances distractor-induced bias.

A. Same as in Figure 2A, but showing the orientation tuning map for distractor decoding for each time point relative to distractor onset. Data averaged across participants (n=23, EEG). Other conventions are the same as Figure 2A.

B. Same as in Figure 2B (right), but showing distractor decoding accuracy for each time point locked to the distractor onset (dashed vertical bar). Magenta and green: Distractor decoding accuracy when it appeared on the cued or uncued hemifields, respectively. Horizontal bars: significant decoding epochs (cluster-based permutation test, Methods). Shading: s.e.m. *(Inset, top)* Posterior electrodes used for the decoding analysis.

(*Inset, right*) Violin plots of distractor decoding accuracy on the cued (magenta) and uncued (green) hemifields.

C. Same as in Figure 2C, but showing behavioral bias curves for the cued memorandum based on a median split of distractor encoding strength – quantified with neural decoding accuracy

following distractor presentation (gray shaded bar, panel B) – when the distractor appeared in the same hemifield as the memorandum (n=23). Solid curve and filled violin plot (*inset*): Stronger distractor encoding trials. Dashed curve and open violin plot (*inset*). Weaker distractor encoding trials. Other conventions are the same as in Figure 2C.

D. Same as in C, but when the distractor appeared in the hemifield opposite to the cued memorandum. Other conventions are the same as in panel C and Figure 2D. (All panels) *: $p<0.05$; **: $p<0.01$; ***:$p<0.001$; n.s.: not significant.

Distractor delay-period timing and gating control bias magnitude

To further understand the mechanistic basis of these distractor-induced biases, we investigated two specific factors that could affect the strength of these biases in WM. Based on recent literature[8], we hypothesized that i) distractor's timing of appearance – early versus late – during the delay period, and ii) its input gating – the extent to which the distractor is gated from access to WM – during the delay-period would modulate distractor-induced bias.

While distractor timing is readily quantified, to quantify distractor input gating we leveraged the finding that information on the prioritized (cued) hemifield is shielded more strongly against distractor interference (Fig. 2B, right)[56–58] . We explored neural signatures associated with the distractor that exhibited marked hemifield asymmetries between the cued and uncued hemifields. Interestingly, the strength of distractor encoding, as quantified by neural decodability over occipital cortex, was not significantly different regardless of whether the distractor appeared on the cued or the uncued hemifields (Fig. 3B) (decoding accuracy: Cued=$9.9\pm2.6\times10^{-4}$, Uncued=$9.8\pm2.60\times10^{-4}$, $p=0.98$, BF=0.22); thus, sensory decodability did not emerge as a reliable marker of distractor gating during the delay period.

We quantified distractor-evoked event-related potential (ERP) components, N1 and P2/P3a (Methods), which are known to index the allocation of spatial attention and working

memory, respectively[31–33]. Although the distractor-evoked contralateral N1 amplitude was not different between hemifields (Fig. 4A) (Cued=-0.85±0.22, Uncued=-0.84±0.21, p=0.73, BF=0.22) the contralateral P2/P3a amplitude was significantly higher for the cued than the uncued hemifield (Cued=0.91±0.19, Uncued=0.72±0.18, p=0.015, BF=4.135). Consequently, we quantified P2/P3a amplitude as a potential indicator of distractor input gating (see Discussion for literature relevant to this claim). We then tested whether the distractor's timing and input gating, as quantified by P2/P3 amplitude, would affect the distractor-induced bias in behavioral reports.

For the first analysis, we median split trials based on when the perceptual distractor appeared during the working memory epoch; we term these "early" and "late" distractor trials; as with all previous analyses, these also included only the trials in which the cued grating was probed. Because our task design incorporated a fixed delay between the memory array and response probe presentation, early (late) distractors appeared earlier (later) during the WM maintenance epoch, and were temporally more distal (proximal) to the response probe (Fig. 1A). Distractor-induced bias, when the distractor appeared on the same hemifield as the cued grating, was significantly positive for late distractor trials ("late": behavioral $Bias_{DistSame}$=1.6°±0.4°, p<0.001, $BF_{+0}>10^2$) (Fig. 4B). By contrast, this bias was not significantly different from zero for early distractor trials ("early": behavioral $Bias_{DistSame}$=-0.3°±0.8°, p=0.64, BF_{+0}=0.16) (Fig. 4B). Moreover, the bias was significantly stronger for late distractors compared to early distractor trials (p=0.014, BF=3.4).

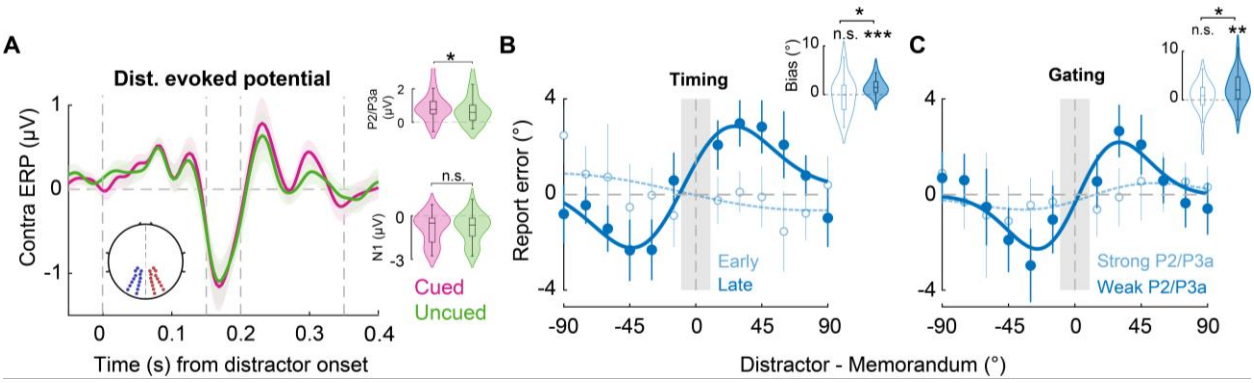


Figure 4. Distractor's timing and gating predict mnemonic biases.

A. Event-related potentials (ERP) locked to distractor onset (dashed vertical line at 0s) over posterior electrodes contralateral to distractor. Magenta and green traces: ERPs evoked by distractor in the cued and the uncued hemifields, respectively. Solid vertical lines at 0.15s-0.2s and 0.2s-0.35s demarcate the temporal windows for computing N1 and P2/P3a amplitudes, respectively. Shading: s.e.m. (*Insets*) Distribution of the P2/P3a (*top*) and N1 (*bottom*) amplitudes evoked by the distractor in the cued (magenta) and uncued (green) hemifields, respectively.

B. Same as in Figure 3C, but showing behavioral bias curves for the cued memorandum separately for trials in which the distractor appeared early (dashed, open symbols) or late (solid, filled symbols) during the delay period, when the distractor appeared on the same hemifield as the memorandum. (*Inset*) Violin plots show the distribution of bias strengths across participants (n=23) for early (open) and late (filled) distractor trials.

C. Same as in panel B, but showing behavioral bias curves for the cued memorandum separately for trials with strong (dashed, open symbols) or weak (solid, filled symbols) distractor-evoked P2/P3a amplitude, a putative marker of distractor gating (panel A). (*Inset*) Violin plots showing the distribution of bias strengths across participants (n=23) for strong (open) and weak (filled) distractor gating trials. (All panels) *: $p<0.05$; **: $p<0.01$; ***: $p<0.001$; n.s.: not significant.

For the second analysis, we divided trials based on the amplitude of the perceptual distractor-evoked contralateral P2/P3a component, for trials in which distractor appeared on the same hemifield as the cued grating. Distractor-induced bias was not significantly different from zero for high P2/P3a-amplitude trials ("strong": behavioral $Bias_{DistSame}$=0.63°±0.68°, p=0.20, BF_{+0}=0.52) (Fig. 4C). By contrast, this bias was significant, and attractive, for weak P2/P3a amplitude trials ("weak": behavioral $Bias_{DistSame}$=2.4°±0.66°, p<0.001, BF_{+0}=47.03) (Fig. 4C). Moreover, we observed a significant difference between the high and low P2/P3a amplitude trials, such that putatively weakly gated trials revealed a stronger bias than strongly gated trials (p=0.030, BF=1.90).

We repeated the same median split analyses for trials in which the distractor appeared on the hemifield opposite to the probed memorandum. In this case, while qualitative trends were apparent in the repulsive bias along expected lines – stronger biases for late and weakly gated distractor trials – none of these trends were statistically significant (SI Fig. S4A-B).

Finally, we report an omnibus analysis combining across all three perceptual distractor dependent factors: encoding strength, timing and gating. For this, we first tested and confirmed that these three factors were not mutually correlated with each other. We found no significant difference in the strength of distractor encoding between early and late distractor trials (p=0.153, BF=1.13). Similarly, there was no significant difference in distractor-evoked P2/P3a amplitude between early and late distractor trials (p=0.563, BF=0.22). Finally, there was also no significant difference in P2/P3a amplitude between trials with strong versus weak distractor encoding (p=0.362, BF=0.37). Given their trail-wise independence, we pooled together trials across functionally matched levels of all three factors (SI Fig. S4C-D) – strong encoding strength, late timing and weak gating ("strong interference") – as well as their converse counterparts – weak encoding strength, early timing and strong gating ("weak interference"). In the strong interference trials, we observed robust evidence for antagonistic biases: a strongly attractive bias

452 ($Bias_{DistSame}$=1.94°±0.35°, p<0.001, BF_{+0}> 10^4) or a repulsive bias ($Bias_{DistOpp=}$=-0.77°±0.37°,
453 p=0.02, BF_{-0}=1.96), when the distractor was on the same or opposite hemifield as the cued
454 memorandum, respectively. By contrast, no such trends were apparent in the weak interference
455 trials ($Bias_{DistSame}$=0.4°±0.43°, p=0.181, BF_{+0}=0.33; $Bias_{DistOpp}$=-0.011°±0.34°, p=0.486, BF_{-0}=0.14).

457 Overall, these results show that perceptual distractors that appeared later during working
458 memory maintenance, produced a stronger bias in orientation reports of the cued memorandum.
459 Moreover, the effectiveness of distractor input gating – as quantified operationally by the
460 distractor-evoked P2/P3a amplitude – reliably predicted the magnitude of bias. A combination of
461 these factors – stronger distractor encoding, late timing and weak gating – produced robust,
462 antagonistic patterns of distractor-induced bias in behavioral responses. We synthesized these
463 links between neural and behavioral effects with a computational attractor model of visual WM.

464

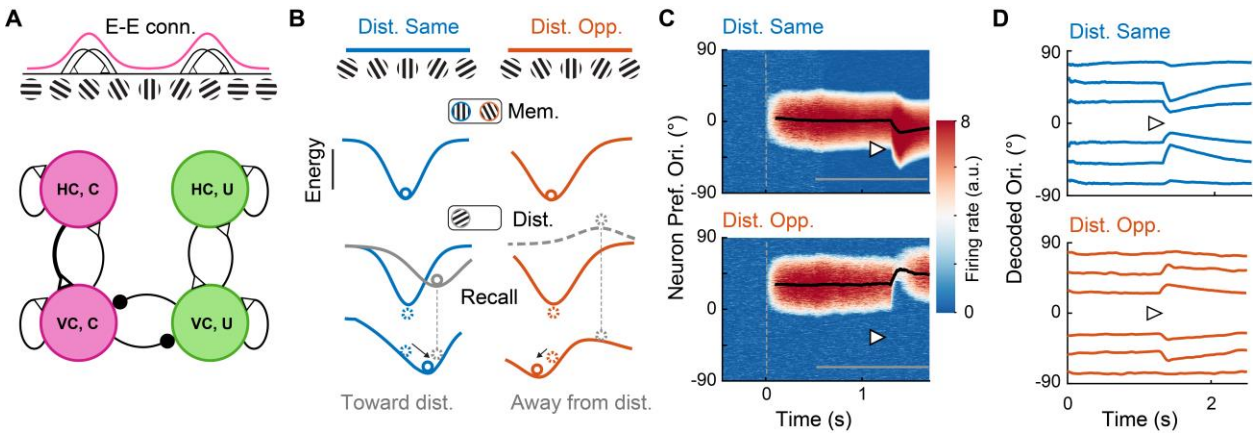
465 **An attractor model explains distractor-induced biases in visual WM**

466 To provide a parsimonious, mechanistic explanation of these perceptual distractor-
467 induced biases during the delay period, we designed a two-tier bump attractor model. Based on
468 extensive recent literature[3–5,9,13,18,34–38], working memory maintenance in our model involves both
469 the sensory (visual) cortex (VC), and higher cortex (HC), such as the prefrontal cortex (PFC) or
470 the parietal cortex (PPC). In fact, the latter regions both are reported to be involved in robust
471 maintenance in the presence of distractors[9,23,24,39]. Moreover, a key element of novelty in our
472 model is as follows: while distinct attractors independently encode stimuli in the two hemifields,
473 they interact with each other via competitive, cross-hemifield inhibition, based on biological
474 observations[40–44]. We provide a brief description of the model below; a detailed description is
475 provided in the Methods.

476 The VC and the HC were modeled with separate ring attractors; we also employed distinct
477 attractors for encoding the memorandum in each (cued, uncued) hemifield, yielding a total of 4
478 attractor networks (Fig. 5A). Each attractor network comprised both excitatory (E) and inhibitory
479 (I) neurons (512 each) with dynamics described by rate-based models (Methods, equations 1-2).
480 Recurrent connections among E neurons within an attractor were local and topographic (Fig. 5A,
481 top), whereas all other connection types (E-I, I-E, I-I) were global (see Methods). VC neurons
482 provided input to HC neurons via topographic, excitatory feedforward connections and received
483 input from HC via reciprocal, topographic feedback connections (Methods, equations 3-4). Based
484 on our experimental observation, we expected cued information in the VC to be more robust to
485 distractors. Therefore, HC→VC feedback connections were 20% stronger for the cued, than the
486 uncued, hemifield (Fig. 5A, thick HC-VC connection). In addition, previous studies suggest that
487 mnemonic information in higher order areas (e.g., PFC, PPC) is relatively robust to
488 distractors[8,9,13,23,39]. Consequently, we transiently blocked feedforward connections from the VC
489 to the HC during the maintenance epoch, thereby rendering the HC mnemonic information
490 immune to disruption by the distractor upon its presentation; this input was selectively reinstated
491 for the weak distractor input gating experiments, described below. Finally, based on previous
492 reports of competitive visual inhibition across the hemifield[40–44], we also modeled cross-
493 hemispheric inhibitory connections between the VC attractors in each hemisphere (Fig. 5A, black
494 circles; equation 3); these topographic connections played a critical role in explaining the
495 antagonistic pattern of biases (see next).

496 With this model, we sought to replicate the attractive and repulsive biases that occurred
497 across space. We explain the expected sequence of events, first, with an energy landscape model
498 (Fig. 5B); detailed simulations are presented subsequently. The energy landscape provides
499 schematic representation activity and attractors in the network[45–47]: Minima (valleys) in the energy
500 landscape reflect stable fixed points that cause stored patterns to drift towards their location

501 whereas, conversely, maxima (peaks) reflect unstable fixed points that cause stored patterns to
502 drift away from them.
503



504

505 **Figure 5. Two-tier attractor model explains spatially antagonistic, mnemonic biases.**

506 **A.** Two-tier ring attractor model. The visual cortex (VC, bottom) and higher cortex (HC, top) were
507 modeled with separate ring attractors. Distinct attractors also encoded stimuli in the cued (C, left,
508 magenta) and the uncued (U, right, green) hemifields. Top-down input from HC to VC was
509 stronger for the cued (thick black line) compared to the uncued hemifield. Mutual inhibition
510 between the VC,C and VC,U attractors was mediated by topographic, cross-hemispheric
511 inhibitory connections (black circles with horizontal arcs). (*Inset, top*) Recurrent excitatory (E-E)
512 connectivity schematic showing local and topographic connectivity profile (see text for details).

513 **B.** An energy landscape intuition for spatially antagonistic distractor-induced biases. (*Top row*)
514 The memorandum's orientation (blue or orange circle) is maintained at the lowest energy state
515 (highest firing rate). (*Middle row, left*) When the distractor (gray circle) appears in the same
516 hemifield as the memorandum, it causes a dip (energy minimum; solid gray line) in the energy
517 landscape. (*Middle row, right*) By contrast when the distractor appears in the opposite hemifield,
518 it causes a peak (energy maximum; dashed gray line) in the energy landscape. (*Bottom row*) As
519 a result, the final energy landscape, obtained by summing the memorandum and distractor

energies, attracts the memorandum's representation towards the distractor's orientation when the latter appears in the same hemifield (blue circle), whereas it repels the memorandum's representation away from the distractor's orientation when the latter appears in the opposite hemifield (orange circle).

C. Firing rate heat maps of the neural activity; hotter colors indicate higher activity. y-axis: E neurons arranged in order of orientation preference on the ring; x-axis: Time. (*Top)* In this example, the onset (dashed vertical line) of the memorandum gratings (0° orientation) causes a localized bump in firing rate at the respective location on the ring, which persists throughout the delay period (gray horizontal bar). Thick black curve: orientation decoded from the population activity vector. When the distractor grating (black triangle; -30° orientation) is presented transiently, and in the same hemifield as the memorandum, the latter's representation is biased toward the distractor's orientation. *(Bottom)* Same as in the top panel, but in this case the distractor grating (-30° orientation) is presented transiently, and in the opposite hemifield as the memorandum (+30° orientation). Here, the memorandum's representation is biased away from the distractor's orientation.

D. Decoded neural representations across time for six different orientations of the memoranda (curves, distinct simulations). Distractor grating (black triangle; 0° orientation) presented in the same (*top,* blue curves) or the opposite (*bottom,* red curves) hemifield as the respective memoranda.

The presentation of a memorandum produces a valley in the VC attractor's energy landscape at its orientation, which persists during the maintenance epoch due to recurrent excitation (Fig. 5B, top). Next, the distractor appearance in the same hemifield as the memorandum (DS) causes another dip in the VC attractor's energy landscape at the distractor's orientation (Fig. 5B, left column, middle row, gray trace). If the dip induced by the distractor is large enough, it perturbs the minimum in the landscape causing the minimum to shift, and the

545 maintained representation to drift, towards the distractor's orientation (Fig. 5B, left column, bottom
546 row, blue trace); this yields an attractive bias in the distractor's hemifield (Fig. 5B, left column,
547 bottom row, blue circle). By contrast, the distractor also creates a peak at its own orientation in
548 the VC attractor's energy landscape in the opposite hemifield (DO); this energy peak arises as a
549 result of the topographic, cross-hemispheric inhibitory connections (Fig. 5B, right column, middle
550 row, gray trace). As schematized, this peak repels the minimum in the landscape, and causes the
551 maintained representation of the memorandum to drift away from the distractor's orientation (Fig.
552 5B, right column, bottom row, orange trace); this yields a repulsive bias in the hemifield opposite
553 the distractor (Fig. 5B, right column, bottom row, orange circle).

554 Next, we describe the results of simulating the model (parameters in SI Table S1). We
555 modeled the initial memory array presentation with an additive input current to the VC-E neurons
556 of each attractor (cued, uncued) using a von Mises spatial profile centered on the orientation of
557 the respective grating (κ_M=2.7°, equation 5). This initiated localized bumps of activity in VC
558 attractors, at locations corresponding to the orientations of the encoded gratings (Fig. 4C top and
559 bottom, activity at 0° and +30°, respectively). Upon disappearance of the memory array, during
560 the delay period (Fig. 5C, gray underbar), the bump of activity was maintained through recurrent
561 excitation in each attractor[26,27,48] (Fig. 5C, red patches). A distractor was presented at t=1300 ms
562 during the delay period (Fig. 5C, triangle); distractor presentation was modeled, also with a von
563 Mises input profile (κ_D=2.3°, equation 6) centered on the distractor's orientation (Fig. 5C;
564 distractor at −30° for both panels).

565 Upon its presentation the distractor induced antagonistic mnemonic bias across hemifields
566 (Fig. 5C). When the distractor was presented in the same hemifield as the memorandum the
567 bump of activity shifted systematically toward the distractor's orientation (Fig. 5C, top, heat map).
568 By contrast, when it was presented in the opposite hemifield as the memorandum the bump of
569 activity shifted systematically away from the distractor's orientation (Fig. 5C, bottom, heat map).
570 We repeated these simulations with six different memoranda of different orientations, with a

distractor always appearing at 0° orientation. In each case, the distractor biased the decoded neural representations towards (Fig. 5D, top) or away from (Fig. 5D, bottom) itself, albeit to varying degrees (see next).

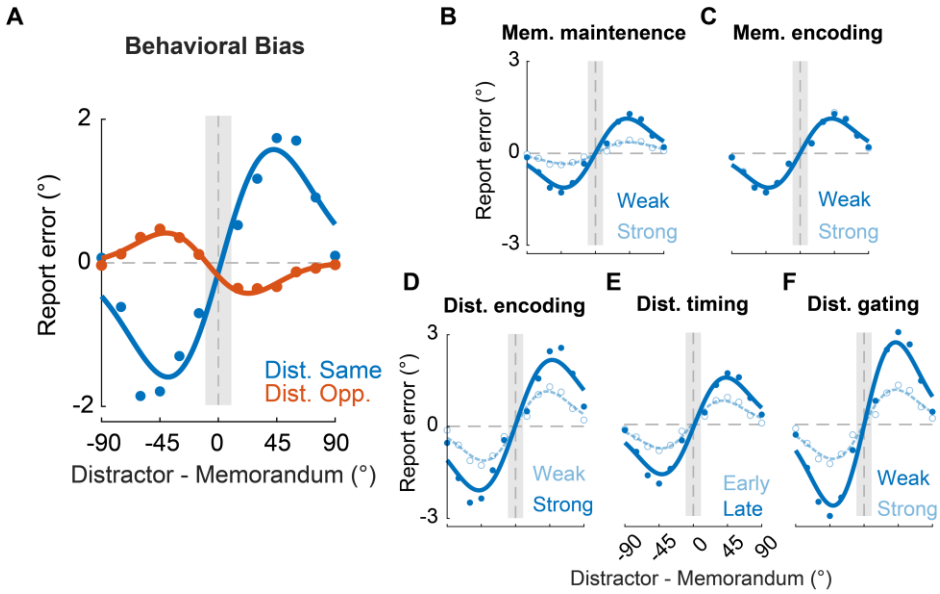


Figure 6. Model simulations qualitatively replicate distractor-induced mnemonic biases.

A. The model's simulated behavioral bias when the distractor appeared in the same (blue), or opposite (orange), hemifield as the cued memorandum. Compare with Figures 1F-G (experimental data). Other conventions are the same as in Figures 1F-G.

B. Same as in panel A, but showing behavioral bias for the cued memorandum plotted separately for simulated trials with weak (solid curve and filled symbols) or strong (dashed curve and open symbols) memorandum maintenance trials (compare with experimental Figure 2C).

C. Same as in panel B, but for simulated trials with weak (solid curve and filled symbols) or strong (dashed curve and open symbols) memorandum encoding.

D. Same as in panel B, but for simulated trials with weak (dashed curve and open symbols) or strong (solid curve and filled symbols) distractor encoding (compare with experimental Figure 3C).

586 **E.** Same as in panel B, but for simulated trials with early (dashed curve and open symbols) or late
587 (solid curve and filled symbols) distractor presentation in the delay-period (compare with
588 experimental Figure 4B).

589 **F.** Same as in panel B, but for simulated trials with weak (solid curve and filled symbols) or strong
590 (dashed curve and open symbols) distractor gating (compare with experimental Figure 4C) (B-F)
591 Other conventions are the same as in Figure 2B. See text for details of how each simulation
592 condition was controlled.

593

594 Next, we sought to replicate key experimental observations regarding the effects of
595 perceptual distractors on behavioral bias magnitude (Methods). Behavioral estimates of
596 orientation were obtained by averaging the memorandum's orientation decoded from both VC
597 and HC neural activity at the end of the trial; this averaging rendered behavioral estimates more
598 robust to noise (see SI Results Section for a detailed analysis of error-correcting dynamics).

599 First, we studied the effect of the distractor's location on simulated behavioral biases. The
600 distractor produced an attractive bias when it appeared in the same hemifield (Fig. 6A) and a
601 repulsive bias when it appeared the opposite hemifield as the memorandum. As in the
602 experiments, the magnitude of this bias varied parametrically with the angular difference between
603 the distractor and the memoranda: the strongest bias appearing for intermediate angle
604 differences, and the weakest appeared for extreme differences (0° or ±90°) (Fig. 1G-H). Neural
605 bias – with orientation decoded based on the peak of the activity bump in the VC attractor
606 (Methods) – revealed an identical trend (SI Fig. S5A).

607 Second, we studied the effect of memorandum maintenance strength on distractor-
608 induced biases (Fig. 6B). We modeled the strength of memorandum maintenance by modulating
609 feedback input from the HC to VC by 2x; we hypothesized that stronger feedback would render
610 mnemonic representations in the VC more robust to interference via error-correcting

611 dynamics[8,9,24]. Indeed, stronger top-down feedback from the HC to VC enabled more accurate
612 maintenance of the memorandum features during the delay period (see SI Results section on
613 *Error-correcting dynamics in a two-tier attractor model* and SI Fig. S6). As a result, the magnitude
614 of distractor-induced bias was systematically higher for the weakly maintained memoranda (Fig.
615 6B, solid), compared to the strongly maintained ones (Fig. 6B, dashed), consistent with
616 experimental results. Interestingly, increasing the memorandum's encoding strength – by
617 affording a 60% higher input current amplitude to the memorandum during the encoding phase –
618 by itself, did not modulate distractor bias (Fig. 6C), precisely in line with experimental
619 observations.

620 Third, we studied the effect of distractor encoding on bias (Fig. 6D). Here, we modulated
621 distractor encoding strength by affording a 60% higher input current amplitude to strongly, as
622 compared to weakly, encoded distractors. The stronger perturbation of the energy landscape
623 induced by the strongly encoded distractors yielded faster drifts in the memorandum's
624 representation thereby yielding a higher bias for strongly (Fig. 6E, solid) compared to weakly (Fig.
625 6E, dashed) encoded distractors; again, these results matched with the experimental
626 observations.

627 Fourth, we found that late distractors produced stronger mnemonic biases (Fig. 6F). We
628 propose that this occurred because of the timing of the distractor vis-a-vis the error-correcting
629 dynamics via top-down feedback from HC to the VC. Specifically, late distractors appear later
630 during the WM maintenance epoch and are temporally more proximal to the response probe than
631 early distractors. Consequently, the former received ameliorative top-down feedback for a shorter
632 duration than the latter. This produced a higher (residual) bias for late (Fig.6F, solid) compared to
633 early (Fig. 6F, dashed) distractor trials at the end of the delay epoch, again, matching
634 experimental findings.

635 Fifth, we studied the effect of distractor input gating on simulated biases (Fig. 6F). We
636 modulated the extent of distractor gating by introducing feedforward input from the VC to HC

637 (strength: 0.01) uniformly during the delay period. We hypothesized that stronger feedforward
638 input into the HC ("weak input gating") would render mnemonic representations in the HC more
639 prone to interference. Indeed, the magnitude of distractor-induced bias was systematically higher
640 for the weakly-gated distractors (Fig. 6F, solid), compared to the strongly gated distractors (Fig.
641 6F, dashed), consistent with experimental results (see also SI Fig. S5).

642 Finally, distractor effects were stronger (magnitude higher) when the distractor appeared
643 in the same hemifield as the memorandum, as compared to when it appeared in the opposite
644 hemifield, uniformly across all simulations (Fig. 6A, SI Fig. S5B-F). Notably, this disparity in bias
645 magnitudes occurred also in the experimental data across hemifields (Fig. 1G-H). In our model
646 this can be explained by the fact that the distractor's influence in the hemifield opposite the
647 memorandum, mediated by the indirect, cross-hemifield inhibition, decayed comparatively faster
648 than its effect in the same hemifield, which was mediated by direct, local excitation.

649 Taken together, our results uncover space-specific, antagonistic patterns of induced by
650 perceptual distractors in visual working memory. Behavioral reports were biased towards (away)
651 the distractor's orientation when it appeared in the same (opposite) hemifield as the
652 memorandum. Cued memoranda were more strongly shielded from distractor interference as
653 compared to uncued memoranda. A two-tier ring attractor model provided a mechanistic
654 explanation for the origins of these biases. Moreover, the model comprehensively explained
655 diverse factors that mediate distractor-induced bias, including cued memorandum maintenance,
656 distractor encoding, timing in the delay period, and input gating. Overall, the results identify
657 lateralized sensory buffers as the putative neural underpinnings of perceptual distractor-induced
658 biases in visual working memory.

Discussion

We discovered that a salient perceptual distractor that shares task-relevant features with the memoranda produces space-specific, antagonistic biases in visual working memory. Memorandum reports were biased towards the distractor's orientation when the latter occurred in the same hemifield as the former, but were biased away when these stimuli appeared in opposite hemifields[22,49–52]. Over time, mnemonic information became more labile and prone to interference by distractors that appeared later in the maintenance epoch. Yet, information associated with the selected (retro-cued) hemifield was held more robustly and shielded against distractor interference, possibly by "error-correcting" top-down control mechanisms[8,9,13,19,24]. Nevertheless, we show that lateralized sensory buffers could enable task-irrelevant perceptual input – sharing features with the memoranda – to gain access to mnemonic information, and thereby, influence the contents of WM profoundly.

Several lines of evidence show that the perceptual distractor biased the information maintained in WM, rather than merely inducing a bias in the behavioral report (response bias). First, had the distractor merely produced a response bias, we would have expected a similar bias to occur regardless of the distractor's location and its orientatation. In our task, even though the response probe and bar were always presented centrally, we observed distinct patterns of biases depending on where the distractor occurred relative to the probed memorandum, and in a manner that systematically varied with their relative orientations (Fig. 1G-H). Second, the distractor was flashed briefly (100 ms) and was followed immediately by a noise mask (pedestal) to diminish its persistent representation. Moreover, participants were aware that the distractor was task irrelevant, and it appeared at least 500 ms before the probe, rendering biases arising from persistence in iconic memory unlikely. Third, we performed a control analysis by fitting the participants' responses based on a combination of target and distractor orientations (von Mises mixture model), to model the scenario where participants reported the distractor's orientation on

684 a subset of trials in which, perhaps, their memory of the memorandum orientation was weak.
685 Formal model selection analysis, based on AICc and BIC, unequivocally ruled out this scenario.

686 Our results strongly support the sensory recruitment hypothesis of visual WM, which posits
687 that brain regions involved in processing the memorandum during encoding also contribute to its
688 maintenance in WM[1–5]. We were able to successfully decode the orientation of the memoranda
689 during the delay epoch from occipital electrodes, and also showed that contralateral decoding
690 strength was a key determinant of distractor-induced biases in behavioral reports (Fig. 2A-B). Yet,
691 due to the low spatial resolution of EEG we cannot readily disambiguate whether the neural
692 signatures that we observed originated entirely from the early visual cortex or also included
693 contributions from other areas, like the parietal or prefrontal cortex[49]. Lastly, distractor-induced
694 bias effects were hemifield-specific; this is consistent with previous studies which suggest that
695 storage of mnemonic information involves a lateralized buffer including, possibly, also the sensory
696 cortex [22,50–53]. Such contralateral buffers may offer a conduit to perceptual distractors to bias the
697 contents of working memory, especially when such information is labile to interference, such as
698 during the delay-period.

699 Nevertheless, prioritization in working memory promoted distractor resilience: cued
700 memoranda were better shielded from distractor interference than uncued memoranda. Taken
701 together with previous studies[24,54], our results show that prioritized information is more robust to
702 interference, possibly due to feedback-mediated stabilization of the cued memorandum's
703 representation[24,54]. Indeed, when simulated in our computational model, top-down feedback from
704 higher cortex stabilized WM representation in the visual cortex on the cued hemifield (Fig. 6B).
705 On the other hand, it is also possible that such stabilization of the cued representation arises from
706 a mixture of persistent activity and activity-silent states[7,29,48,55]; the latter are arguably less prone
707 to interference from sensory distractors[7]. Future experiments can disambiguate these hypotheses

708 by using a “pinging” protocol that activates latent representations and studying the effect of
709 distractor interference on these activated representations.

710 Among the multiple electrophysiological indices we tested, we propose the P2/P3a ERP
711 amplitude as a putative marker for distractor gating into working memory. Previous literature has
712 often linked the P2 component to the gating (or suppression) of distractor information during
713 working memory tasks[56–58.] For instance, higher distractor-evoked P2 amplitude, indicative of
714 greater distractor gating, correlated positively with performance in an auditory WM task[31].
715 Similarly, the short peak latency P3 component, referred to as the P3a, possibly originates in the
716 frontal lobe[57] and indexes top-down inhibition of distractors[56,57.] In fact, distractor-evoked P3a
717 amplitudes were correlated with frontal lobe grey matter volume across individuals[59], suggesting
718 its involvement in the distractor-related processing. Our study extends these findings and shows
719 that the P2/P3a amplitude contralateral to the distractor was higher at the cued compared to the
720 uncued hemifield; in other words, the P2/P3a may index a stronger input gating of cued mnemonic
721 information stored in a lateralized mnemonic buffer (Fig. 4A). Moreover, the P2/P3a amplitude
722 directly predicted the magnitude of the distractor-induced bias, providing further confirmation for
723 this “gating hypothesis”.

724 We explained the mechanistic origins of perceptual distractor biases with a two-tier ring
725 attractor WM model. Attractor models have been frequently used for explaining diverse
726 phenomena in WM, including variable precision[27], serial biases[48], categorical biases[60] and the
727 like[28]. For instance,-a recent study employed an attractor model to explain serial biases in spatial
728 WM[48]. Other studies employing attractor models have been able to explain the mechanisms
729 behind trial to trial variability and the presence of categorical biases in visual WM[26,27,48]. In our
730 study, a ring attractor model was a natural candidate to explain our results. Recent research has
731 shown that maintenance in visual working memory (WM) involves multiple regions across the
732 brain, including both sensory and higher-order cortex[18,35,37]. We, therefore, employed a two-tier

733 attractor model that includes both sensory and higher order brain regions to simulate our
734 experimental findings.

735 In our model, a higher-order cortical region (HC) stored information relevant to the
736 memorandum and provided top-down influence to the visual cortex (VC) during the delay period
737 to mitigate distractor interference in WM. Candidate regions for the HC include parietal cortex or
738 prefrontal cortex both of which are known to contain robust WM representations, especially in the
739 presence of distractors[9,13,23,24,39]. In our model, the representation in HC was rendered immune to
740 distractor interference by transiently blocking its bottom-up input from VC during the delay period,
741 and enabling this input only for weakly gated distractors. Neurophysiological mechanisms that
742 could mediate such dynamic connectivity include oscillatory coherence[61] or neuromodulatory
743 gating (e.g., acetylcholine[62]). Yet, some recent studies have shown that disruption in visual WM
744 maintenance occurs in both lower (V1, V4) and higher level cortex[11]. In this case, the coordinated
745 activity across these regions could help improve the robustness of maintenance in WM, even in
746 the presence of distractors[18,35–37,63]. Consistent with these results, our simulations showed that
747 combining VC and HC activity yielded a more accurate behavioral readout of the memorandum's
748 orientation, as compared to either VC or HC activity alone (SI Fig. S5A). Future work, employing
749 fMRI, perhaps in combination with EEG, could help identify the precise brain regions, and their
750 dynamics, that mediate distractor biases in WM.

751 Our model accurately reproduced the effect of distractor timing on mnemonic bias:
752 distractors presented later during the delay epoch induced a stronger bias, and vice versa. This
753 experimental result is in line with literature suggesting that neural WM representations are not
754 static but evolve over time, as dynamic processes[8,30,64]. We leveraged these findings in our model
755 to explain distractor timing effects. In our model, the bias due to error-correcting dynamics in
756 visual WM decays steadily; these dynamics are, in turn, mediated by top-down feedback from
757 higher cortex (e.g., PFC[9,13,23,24,39]). Due to this effect, early distractor trials provided more time for

758 the error-correcting mechanism to operate thereby yielding a weaker behavioral bias, as
759 compared to late distractor trials. While model simulations accurately recapitulated experimental
760 data, future experiments can test this hypothesis by stimulating neurons in higher brain regions
761 during WM maintenance and studying its effects on the extent of error-correcting dynamics
762 observed in lower order brain regions and in behavior.

763 In conclusion, our study offers putative mechanistic insights into neural processes by
764 which perceptual distractors induce space-specific biases in working memory. Complementing
765 these experimental results, our simulations suggest key computational hypotheses that by which
766 distractors bias in WM representations. These hypotheses may be tested causally with non-
767 invasive brain stimulation (e.g., transcranial magnetic stimulation) by focally perturbing activity in
768 specific brain regions (e.g., visual cortex, prefrontal cortex), and at specific times during the delay
769 epoch. Understanding the neural basis of distractor interference could have critical implications
770 for improving the robustness of working memory in patients with attention deficit disorder (ADD)
771 or mild cognitive impairment (MCI[65]).

Materials and Methods

Participants

24 participants (7 female; age range: 23-38 yrs; median age: 28 yrs), all right-handed, participated in the experiments. All participants had normal, or corrected-to-normal, vision and no known history of neurological disorders. Participants provided written and informed consent to take part in the experiments. Study protocols were approved by the Institute Human Ethics Committee at the Indian Institute of Science, Bangalore.

Sample size estimation. Sample sizes were not estimated with power analysis. The sample size in this study (n=24) is largely similar to that of several earlier studies on working memory and EEG-based grating orientation decoding[12,29,30].

Behavioral data acquisition

Task and stimuli. Participants completed the experimental tasks in a dimly lit room while seated, with their heads secured by a chin rest, with the eyes located 60 cm from a stimulus display screen. Stimuli were displayed on a contrast calibrated (i1 Spectrophotometer, X-Rite Inc.) 24-inch LED monitor (XL2411Z, BenQ corp.) with a resolution of 1920x1080 pixels and a refresh rate of 100 Hz. Stimuli were programmed in Matlab 2015b (Mathworks Inc.) with Psychtoolbox version 3.0.15. During the experiment, participants used an optical mouse (M-U0026, Logitech) to provide responses. Participants' gaze position was tracked throughout the experiment using an infrared eyetracker (500 Hz, HiSpeed, iView X, SensoMotoric Instruments Inc.).

Each trial began with a fixation dot (full black, radius: 0.3 dva) presented at the center of the screen. Concurrently with this, circular placeholders (uniform white noise patches) were presented symmetrically in either hemifield along the azimuth (eccentricity: ±8.0 dva, radius: 4.0 dva); both the fixation dot and the placeholders remained on the screen throughout the trial. 500

ms after fixation dot onset, the memory array, comprising two sine-wave gratings (50% contrast, radius: 3.5 dva, spatial frequency: 0.9 cpd), appeared for a brief interval (100 ms), each embedded (50% alpha blended) concentrically inside one of the placeholders in each hemifield. 500 ms after memory array offset a retro-cue (white filled semicircle) appeared for 100 ms which indicated the location of the grating that would likely be probed for response (cue validity: 70%). Following a random delay (500-1250 ms, drawn from an exponential distribution), a singleton distractor grating (100% contrast) appeared for 100 ms in one of the two hemifields, pseudorandomly, with equal probability. This "perceptual" distractor's size and spatial frequency was identical to that of the memory array gratings, but its orientation was drawn randomly from a uniform distribution across all angles at least ±10° away from either grating's orientation in the memory array. The distractor was irrelevant to the task and the participants were explicitly instructed about the same. No distractor grating appeared in 20% of the trials; these trials interleaved among the distractor presentation trials.

A fixed interval (1950 ms) after retro cue onset, a response probe (white or black filled circle, 0.3 dva, with a yellow ring around it) appeared on the screen indicating the grating relevant for response. A white (black) probe indicated that the participants must reproduce the orientation of the cued (uncued) grating. A randomly oriented bar (length: 7 dva, width 0.3 dva) appeared at the center of the screen upon response initiation. Participants reproduced the orientation of the grating on the probed hemifield from memory by rotating the bar with the mouse using their right (dominant) hand, and pressed the left mouse button to register their response. Although there was no constraint on the time to initiate the response after probe onset, participants had to complete their response, once initiated, within 2 seconds. The next trial began after an inter-trial interval of 2 seconds.

<u>Training and testing.</u> Before the main experiment, participants completed between 2-4 blocks (100 trials each) of the task, typically, without EEG recordings. To familiarize and train participants

821 on the task, feedback was provided at the end of each trial. The feedback consisted of an oriented,
822 magenta bar whose orientation precisely matched that of the probed grating shown along with
823 the reported orientation (black bar); the feedback was shown for 200 ms at the end of the
824 response period. Following this, they also received textual feedback, displayed on the screen,
825 which indicated whether their response was “precise” (error <= 15°) or “imprecise” (error > 15°).
826 Training data were not used for further analyses.

827 Following a brief break (typically 20 min), participants performed the main task, without
828 behavioral feedback, but with concurrent EEG recordings (see next). Each participant completed
829 6 blocks of the task (100 trials each), with periodic breaks at the end of every set of 50 trials.
830 Thus, in total, we acquired and analyzed 14,400 trials of behavioral data from 24 participants.

831

832 Eye-tracking and gaze-based trial exclusion. For all participants, gaze fixation position was
833 monitored binocularly at 500 Hz, and stored offline for analysis. In *post hoc* analyses, any trial in
834 which the subject’s gaze deviated by more than 1 dva from the fixation dot along the azimuthal
835 direction was flagged, and excluded from further analysis. Given that we were primarily interested
836 in the effect of distractors on visual working memory, this criterion for gaze-based exclusion was
837 applied only in a window around (100 ms before to 400 ms after) distractor onset during the delay
838 epoch. We did not reject trials based on gaze deviation during other times in the delay epoch,
839 because, barring the fixation dot and placeholders, there were no other stimuli on screen during
840 this epoch. We also did not reject trials based on gaze deviation during the memory array
841 presentation, because there was no prior information (or pre-cue) that could have systematically
842 biased gaze toward either grating. In addition, no gaze-based rejection was performed for trials
843 in which the distractor did not appear. Finally, for 6/24 participants gaze fixation could not be
844 reliably recorded, largely because these participants wore eyeglasses with strongly reflective
845 lenses; trial rejection rates for these participants were >15%. To avoid major biases in the

statistical distribution of task parameters, for these 6 participants, we included all trials in the analyses; the results were similar even with these participants excluded from the analyses. For the remaining participants (n=18/24), the median eye-tracking rejection rate was 5.2%±0.8% (mean±s.e.m.).

Behavioral data analyses

Estimating precision and bias in behavior. Participants reported their best estimates of grating orientation. Simple behavioral metrics, computed directly from the data, revealed the effects of the distractor on precision and bias in participants' reports[66–68], without the need for fitting complex psychophysical models[69].

We computed the effect of the distractor on the precision of each participant's orientation report using the circular standard deviation (CSD)[68], a circular measure of dispersion of the orientation reports about its true value. A smaller CSD indicates higher average precision in orientation reports. Data were pooled from trials across all blocks for this analysis. We computed and compared the CSD separately for the trials in which the distractor appeared on the same hemifield, versus the opposite hemifield, of the probed stimulus; this analysis was performed separately for the cued and uncued grating reports. In addition, we fit a mixture model comprising von Mises and uniform distributions to estimate precision from the pooled responses, separately for the distractor-same and distractor-opposite trials. The difference ($\Delta\kappa$) in or modulation index (MI-κ; defined in the Results) of precision between the distractor-same and distractor-opposite trials was quantified separately for the cued and the uncued memoranda, and compared with permutation tests (see Methods section on *Statistical Tests*).

We also computed the effect of the distractor in biasing orientation reports for the probed grating, following a standard protocol[28,67]. First, we plotted the median orientation report error for

the probed grating (reported – actual probed orientation) as a function of the angle difference between the distractor and the probed grating (distractor – actual probed orientation); this computation was performed with trials pooled in overlapping bins (size: ±15°, shift: 1°; range: –90° to 90°), based on the difference between the distractor and the probed grating orientation. Note that, in Figure 1E, if orientation reports were systematically biased towards the distractor orientation, then the y-axis (orientation report error) would be positive for positive values of the x-axis (relative distractor orientation), and vice versa for bias away from the distractor orientation. We quantified this bias by computing the signed area under the curve, with a procedure similar to that employed in previous studies (e.g., Myers et al[28]). Briefly, the area under the curve in quadrants II and IV were summed and subtracted from the summed area under the curve in quadrants I and III (Fig. 1E) – relative orientation values from -10° to 10° were excluded in this analysis because the distractor's orientation was at least 10° away from either of the memory array grating orientations (see Methods, Behavioral Data Acquisition section). This analysis was performed both with data from all trials, and also separately for trials in which cued and uncued gratings were respectively probed. For visualization purposes, the bias curves were fit with a Gaussian derivative function[67].

<u>Estimating the contribution of distractor orientation reports.</u> To quantify the contribution of distractor orientations reports to the observed behavioral responses, we fit, for each participant, a three-component mixture model – comprising two von Mises distributions and a uniform distribution [29]:

$$p(\hat{t}) = \alpha\, f(\hat{t}; t, \kappa_t) + \beta\, f(\hat{t}; d, \kappa_d) + \gamma \frac{1}{2\pi}; \qquad \alpha + \beta + \gamma = 1$$

$\hat{t}$ denotes the reported orientation, $p(\hat{t})$ the density (fit) of orientation reports, $f(\hat{t}; t, \kappa_t)$ is a von Mises distribution centered on the target orientation *t*, with concentration parameter κ_t, $f(\hat{t}; d, \kappa_d)$

894 is a von Mises distribution centered on the distractor orientation d, with concentration parameter
895 κ_d, the uniform distribution models random guesses, and (α, β, γ) represent, respectively, the
896 mixture weights associated with the target distribution, distractor distribution, and the uniform
897 weight; these are constrained to sum to 1. We estimated mixture weights and concentration
898 parameters $(\kappa_t\ , \kappa_d)$ using expectation–maximization[29] for the response distribution associated
899 with the cued memorandum, and separately for the distractor-same and distractor-opposite trials.
900 We call this the "target + distractor" model; the "target" refers to the cued memorandum, when it
901 was probed for recall.

902 Next, to evaluate the contribution of distractor orientation reports, we fit a simpler mixture
903 model excluding the distractor von Mises component; we term this the "target alone" model:

904 $$p'(\hat{t}) = \alpha\, f(\hat{t}; t, \kappa_t) + \beta \frac{1}{2\pi}; \qquad \alpha + \beta = 1$$

905 The two models – "target+distractor" and "target alone" were compared using the corrected
906 Akaike Information Criterion (AICc), as well as the Bayesian Information Criterion (BIC) [31,32].
907 Briefly, the two metrics trade off goodness-of-fit – quantified with the log-likelihood – against
908 model complexity – based on the number of model parameters – with BIC imposing a more
909 stringent penalty on model complexity. AICc and BIC values were computed separately for the
910 distractor-same and distractor-opposite conditions, and then summed. Finally, we also repeated
911 these same model comparison analyses by fitting both "target+distractor" and "target alone"
912 models omitting the uniform component, and obtained results similar to those with this term
913 included (see Results).

914

915 **EEG data acquisition and preprocessing**

916 <u>EEG acquisition.</u> High density EEG data was acquired with the Active Two recording system
917 (Biosemi, Amsterdam, Netherlands), with 128 Ag-AgCl active electrodes (Biosemi Inc.). During

recording, the data were referenced online to the Common Mode Sense (CMS) active electrode. The data was recorded continuously throughout the task with a sampling rate of 4096 Hz and stored for offline analysis. EEG data for one participant could not be retrieved due to a critical hardware error. EEG analyses were performed with the remaining 23/24 participants.

EEG Preprocessing. EEG data was preprocessed in MATLAB using the EEGLAB[70], ERPLAB[71] and Noise Tools (http://audition.ens.fr/adc/NoiseTools/) toolboxes. The data was first resampled at 256 Hz (*pop_resample* function, EEGLAB), bandpass filtered from 0.05 to 40 Hz using Butterworth filter of order-2 (*pop_basicfilter*, ERPLAB), and cleaned of line noise at 50 Hz (*pop_cleanline*, EEGLAB). Next, noisy electrodes were identified as those having more than 33% of their values >4 times the median absolute value across all electrodes and time points (*nt_find_bad_channels,* Noise Tools); this yielded a channel removal rate of 4.1% ± 0.8% [mean ± s.e.m.] on average, per participant. The removed electrodes were then interpolated using the neighboring electrodes' data with spherical spline interpolation (*pop_interp,* EEGLAB). Subsequently, the trials were epoched from -1600 to 2300 ms relative to the cue onset and median re-referenced by subtracting the median value across all electrodes[76]. The epoched data were then baseline normalized by subtracting the mean activity from -200 to 0 ms relative to cue onset from all time points, separately for all electrodes and trials (*pop_rmbase,* EEGLAB). Next, we applied independent component analysis (ICA) using the infomax algorithm (*pop_runica*, EEGLAB) to identify artifacts arising from physiological and other, non-neural sources. We used ICLabel[72], an EEGLAB extension that probabilistically identifies ICs corresponding to neural sources versus those arising from eye movements, muscle artifacts, channel noise and other noise sources. Any component belonging to a non-neural source with high probability (>70%) was flagged, and removed after visual inspection; 3.2 ± 0.6 [mean ± s.e.m.] components were removed on average, per participant. In a final pass, we identified noisy trials using the SCADS algorithm[73].

943 Briefly, SCADS identifies outlier signal values based on statistical measures like the standard
944 deviation, maximum amplitude, and the like. Trials in which SCADS flagged >20 percent of
945 electrodes (outlier criterion, λ=5.0) were removed from the analysis. In other trials, noisy
946 electrodes were linearly interpolated based on the nearest 4 neighbors; this yielded a median trial
947 removal rate of ~3% trials per participant.

948

949 **EEG data analyses**

950 Decoding stimulus orientation. We decoded the orientation of memory array and distractor
951 gratings based on multivariate pattern analysis of EEG responses. Whereas the memory array
952 grating orientations were decoded both during memory array presentation (Fig. 1A) and the delay
953 epoch (Fig. 1A), distractor grating's orientation was decoded only during its presentation epoch
954 (Fig. 1A). The orientations of the memoranda were decoded with 14 posterior electrodes over the
955 hemisphere contralateral to the respective stimulus (Fig. 2B, inset) pooled across the cueing
956 conditions (see also control analysis, next, in which occipital electrodes were used bilaterally for
957 decoding). The distractor's orientation was decoded with all[39] posterior electrodes over occipital
958 cortex (e.g., Fig. 3A-B). We employed a multivariate decoder based on the Mahalanobis-distance,
959 following a procedure established in literature[12,29,30]. Briefly, EEG data were downsampled to 128
960 Hz by taking an average of each set of two successive samples (original sampling rate: 256 Hz).
961 To improve the decoding accuracy, we used a sliding window (size: 300 ms, shift: 8 ms) and
962 concatenated all the data points in that time window from all electrodes into a one-dimensional
963 vector (38 time points in each time window x number of electrodes). We utilized the pattern of
964 temporal variability of the EEG response within each time window for decoding[30.] Therefore, we
965 subtracted the mean activity across the time window for each channel, and applied principal
966 component analysis (PCA) to reduce the number of feature dimensions; the top-k PCA

967 dimensions that explained 95% of the variance across trials were retained for decoding ("encoding
968 vector").

969 For training and evaluation, we employed a leave-one-trial-out cross-validation approach.
970 For each participant, the model was trained on all but one trial, with orientations binned relative
971 to the orientation in the left-out trial into one of 16 different bins (bin centers: 11.25° apart, bin
972 width: ±30°). Then the Mahalanobis distance was computed between the test trial's encoding
973 vector and the mean encoding vector of each training bin. The computed distances were de-
974 meaned and the sign of the distances was reversed for ease of visualization. To obtain a smooth
975 decoding estimate, this procedure was repeated 8 times by sliding the bin centers by 1.4° each
976 time, and the resultant decoding profiles were averaged across repetitions (e.g., Fig. 2A). To
977 estimate decoding accuracy, the sign-reversed distances were multiplied elementwise by a cosine
978 kernel, symmetric around 0° (period: 180°), and averaged. This procedure was performed for
979 each participant, memory array location (left/right), distractor location (left/right), and each time
980 window, separately.

981

982 Distractor-evoked ERP estimation. The event related potential (ERP) was measured by signal-
983 averaging across the same set of 14 contralateral (or ipsilateral) electrodes as employed in the
984 decoding analyses. First, the data were high-pass filtered with a lower cut-off at 0.5 Hz using an
985 FIR filter (N=500). Subsequently, the data were epoched from 800 ms before to 700 ms after
986 distractor onset and then averaged across channels and trials to obtain the ERP waveform. This
987 was done separately for each of the four distractor conditions (cued/uncued x left/right hemifield).
988 Click or tap here to enter text. The N1 amplitude was defined as the mean signal value in a window
989 extending from 150 to 200 ms following distractor onset and P2/P3a amplitude as the mean signal
990 value in a ±40 ms window around the respective P2 and P3a peaks [31,33].

991

992 Median-split analyses based on memorandum maintenance, distractor encoding strength, timing
993 and gating. We tested the association of diverse neural metrics and task variables to distractor-
994 induced behavioral bias. For example, we measured the association between the strength of the
995 cued memorandum's maintenance and behavioral bias. For this analysis, the strength of the cued
996 memorandum's maintenance was quantified based on the average decoding accuracy (cosine
997 similarity) in a 500-ms window in the delay period – beginning with cue onset, until 100 ms
998 preceding the earliest distractor onset. Then, we divided trials into "weak" and "strong"
999 memorandum maintenance trials based on a median split of the cosine similarity metric; the
1000 median split was performed for each participant and for each distractor condition (cued/uncued x
1001 left/right hemifield) separately. Weak and strong memorandum maintenance trials were then
1002 aggregated across the conditions, and the behavioral bias in the cued orientation report was
1003 computed separately for the two groups of trials with a procedure identical to that described
1004 previously (see Methods section on *Estimating precision and bias in behavior*).

1005 The association between other metrics like the cued memorandum encoding strength,
1006 uncued memorandum maintenance strength, distractor encoding strength, distractor timing or the
1007 strength of gating and behavioral bias was computed in an identical manner, except that the
1008 median split was performed based on the respective metric. For the cued memorandum encoding
1009 strength, a median split was performed based on cued memorandum decoding accuracy values
1010 in a window from 100 ms to 500 ms following memory array onset. For the uncued memorandum
1011 strength, a median split was performed based on uncued memorandum decoding accuracy
1012 values in a 500 ms delay-period window, identical to that for the cued memorandum. For distractor
1013 encoding strength, a median split was performed based on distractor decoding accuracy (cosine
1014 similarity) values computed in a window from 100-400 ms after distractor onset. For distractor
1015 timing (early versus late), a median split was performed based on the times of distractor onset

1016 relative to the cue. For distractor input gating, the median split was based on the amplitudes of
1017 the distractor-evoked P2/P3a component over the electrodes contralateral to the distractor.
1018 Statistical tests were performed both to assess whether the bias for each half of the median split
1019 was, individually, significantly different from zero, and also to assess for within-participant
1020 (pairwise) differences across the median split halves. For each split, we computed one-tailed
1021 Bayes Factors (see next), which enabled us to quantitatively evaluate evidence for or against the
1022 effect of each factor on distractor-induced bias.

1023 Prior to this analysis, we first tested and confirmed that these factors were independent of
1024 each other and of the distractor's encoding strength. First, trials were divided into high and low
1025 cued memorandum maintenance groups based on a median split, and we tested whether
1026 distractor encoding strength differed significantly between these two groups. Second, trials were
1027 split based on the timing of distractor appearance (early vs. late), and we assessed whether this
1028 timing influenced distractor encoding strength or distractor gating. Finally, trials were divided
1029 based on distractor encoding strength (strong vs. weak), and we examined whether distractor
1030 gating strength differed between these groups. These control analyses confirmed that the primary
1031 factors of interest were not confounded with one another, providing a basis for interpreting
1032 subsequent analyses.

1033 **Computational model of distractor biases in WM**

1034 i) Network architecture. We synthesized our behavioral and neural findings by modifying and
1035 extending a bump attractor model of working memory[12]. The visual cortex (VC) and higher cortex
1036 (HC), for each of the cued and uncued visual hemifields, were modeled with distinct ring attractors
1037 (total of 4 ring attractor models). Each attractor model comprised 1024 neurons in total out of
1038 which 512 were excitatory (E) and 512 were inhibitory (I). Each neuron was described by a
1039 standard rate model and represented stimulus orientation topographically, depending on its
1040 location along the ring (Fig. 5 A, top; θ ranging from -90° to +90°).

For ease of illustration, we start by describing the dynamical equation governing the firing rate r of the cued visual cortex attractor (VC,C) as a function of time t[27,28].

$$\tau_E \frac{dr_E^{VC,C}}{dt} = -r_E^{VC,C} + \phi\left(W_{EE}^{VC,C} r_E^{VC,C} - W_{IE}^{VC,C} r_I^{VC,C} + I_E^{VC,C} + C_E^{VC,C}\right) + \sigma_E^{VC,C}\xi(t) \qquad (1)$$

$$\tau_I \frac{dr_I^{VC,C}}{dt} = -r_I^{VC,C} + \phi\left(W_{EI}^{VC,C} r_E^{VC,C} - W_{II}^{VC,C} r_I^{VC,C} + C_I^{VC,C}\right) + \sigma_I^{VC,C}\xi(t) \qquad (2)$$

Where, $r_E^{VC,C}$ and $r_I^{VC,C}$ denote the firing rate of the excitatory and inhibitory neurons of the visual cortex, cued attractor (VC,C) respectively. T_E and τ_I are neuronal (integration) time constants for E and I neurons respectively, $W_{EE}^{VC,C}, W_{IE}^{VC,C}, W_{EI}^{VC,C}, W_{II}^{VC,C}$ are the connection strengths from E to E, I to E, E to I and I to I neurons respectively for the VC,C attractor. $C_E^{VC,C}$ and $C_I^{VC,C}$ are the constant input bias currents to E and I neurons respectively of VC,C attractor. $\Xi(t)$ represents Gaussian white noise with unit variance, scaled by parameters $\sigma_E^{VC,C}$ and $\sigma_I^{VC,C}$ for E and I neurons of VC,C attractor respectively. $I_{\mathrm{E}}^{VC,C}$ represents the input from other sources (explained in equations 3 and 4, see next). ϕ represents the function that maps the input current to the output firing rate adopted from Wimmer et al., 2014[27]

$$\phi(r) = \begin{cases} 0, & \forall\, r < 0 \\ r^2, & \forall\, 0 < r < 1 \\ sqrt(4r-3), & \forall\, r \geq 1 \end{cases}$$

Three of the four sets of recurrent connections – W_{IE}, W_{EI}, W_{II} – were uniform and all-to-all (i.e., $w^{ij} = w_0\ \forall\, i, j$). On the other hand, recurrent excitatory connections w_{EE}^{ij} were local, and followed a von Mises profile: $exp(\kappa cos(\theta_i - \theta_j))$ where κ is the concentration parameter and θ_i and θ_j are the orientation preference of excitatory neurons i and j respectively.

The firing rates of three other attractors, namely, the visual cortex uncued attractor (VC,U), higher cortex cued attractor (HC,C) and higher cortex uncued attractor (HC,U) also follow identical

dynamical equations. The values of the parameters for the 4 different attractor models are shown in SI Table S1.

Inputs to the cued visual cortex attractor ($I_{\mathrm{E}}^{VC,C}$, equation 1) comprised 3 contributions:

$$I_{\mathrm{E}}^{VC,C} = U^{VC,C} + W_E^{HC,C\rightarrow VC,C}\, r_E^{HC,C} + W_I^{VC,U\rightarrow VC,C}\, r_E^{VC,U} \qquad (3)$$

The first term, $U^{VC,C}$ represents the bottom-up (feedforward) inputs due to the external stimulus that become active during the presence of the memory array and distractor gratings (successively). The second term, $W_E^{HC,C\rightarrow VC,C} r_E^{HC,C}$ represents top-down (feedback) inputs from the HC to the VC attractor, organized topographically (one-to-one); here, $W_E^{HC,C\rightarrow VC,C}$ denotes the top-down excitatory connection weights from the HC,C to VC,C attractor. The third term, $W_I^{VC,U\rightarrow VC,C}\, r_E^{VC,U}$ represents long-range, lateral inhibitory inputs received from the VC attractor in the opposite hemisphere (cross-hemispheric inhibition); here, $W_I^{VC,U\rightarrow VC,C}$ represents the lateral inhibitory connection weights from the VC,U to VC,C attractor. Note that the magnitude this inhibition modulates with the firing rate of the excitatory neurons (r_E) in the opposite hemifield; an intervening inhibitory population that converts this long-range excitatory projection into functional inhibition (via weights, W_I) was not explicitly modeled. The inputs $I_E^{VC,U}$ to the uncued visual cortex attractor also followed an identical equation except for minor modifications to the respective bottom-up inputs ($U^{VC,U}$), top-down inputs ($W_E^{HC,U\rightarrow VC,U} r_E^{HC,U}$) and long-range, lateral inhibitory inputs ($W_I^{VC,C\rightarrow VC,U}\, r_E^{VC,C}$).

Inputs to the cued higher cortex attractor ($I_{\mathrm{E}}^{VC,C}$, equation 1) comprised only one contribution corresponding to bottom-up connections from the VC,C attractor, also organized topographically (one-to-one):

$$I_E^{HC,C} = W_E^{VC,C\rightarrow HC,C} r_E^{VC,C} \qquad (4)$$

Where $W_E^{VC,C\rightarrow HC,C}$ represents the topographic bottom-up connection strength from the E neurons of the VC,C to the E neurons of the HC,C attractor ($W_E^{VC,C\rightarrow HC,C}$). The inputs $I_{\mathrm{E}}^{HC,U}$ to the uncued higher cortex attractor also followed an identical equation except for its bottom-up inputs ($W_E^{VC,U\rightarrow HC,U}$).

<u>ii) Modeling the distinct task epochs.</u>

<u>Encoding:</u> 500 ms after the trial onset the memorandum was presented for a duration of 500 ms. The visual cortex (VC) attractor receives sensory input during this period (through $U^{VC,C}$ and $U^{VC,U}$ terms in equation 3). We modeled this by changing the spatial profile of the input current to the excitatory neurons of the VC attractor according to:

$$U_i^{VC,C} = A_M * exp(\kappa_M cos(\theta_i - \Theta_M^C)) \qquad (5)$$

Where, $U_i^{VC,C}$ is the input current to the excitatory neuron *I* of VC,C attractor, A_M is the amplitude of the input current, κ_M is the concentration parameter determining the spread, θ_i is the orientation preference of excitatory neuron i and Θ_M^C is the memory array grating's orientation on the cued hemifield. The spatial profile of the input current was centered on the E neuron representing the respective grating's orientation. This instantiated localized bumps of activity in VC, at locations corresponding to the orientations of the encoded gratings. A similar equation was used to model the input current for the VC,U attractor. The PFC attractor did not receive any direct inputs other than the bottom-up input received from the VC attractor during encoding (equation 4).

<u>Delay:</u> The delay period commenced immediately after memory array offset. During the delay period, until the end of the trial, we disconnected the bottom-up connections from the VC to the HC attractor ($W_E^{VC,C\rightarrow HC,C} = 0$, $W_E^{VC,U\rightarrow HC,U} = 0$ in equation 4), rendering the HC attractor immune

1106 to subsequent sensory input, thereby modeling "strong" distractor gating. These connection
1107 values were set to 0.01 only for the specific case of modeling "weak" distractor gating (Fig. 6E, SI
1108 Fig. S5F). 600 ms after the memory array onset, the cue appeared. To mimic prioritization of cued
1109 information in WM[24], we reduced the noise terms for both the cued attractors
1110 ($\sigma_E^{VC,C}, \sigma_I^{VC,C}, \sigma_E^{HC,C}$ and $\sigma_I^{HC,C}$ in equations 1 and 2) by 60% and increased top-down feedback to
1111 cued VC attractor ($W_E^{HC,C \to VC,C}$ in equation 4) by 20%, following cue onset.

1112

1113 *Distractor presentation:* The distractor was presented for a duration of 0.1 seconds, at a random
1114 time (distributed exponentially) between 1200 ms and 1950 ms after the memory array onset,
1115 mimicking the behavioral paradigm. The distractor was presented either in the cued or the uncued
1116 hemifield with equal probability and was encoded by the visual cortex (VC) attractor representing
1117 the corresponding visual hemifield. Distractor presentation was modeled by adjusting the spatial
1118 profile of the input current to the excitatory neurons of the VC attractor as follows:

1119 $$U_i^{VC,C} = A_D * exp(\kappa_D cos(\theta_i - \Theta_D)) \qquad (6)$$

1120 Where $U_i^{VC,C}$ is the input current to the excitatory neuron i of VC,C attractor, A_D is the amplitude
1121 of the input current, κ_D is the concentration parameter determining the spread, θ_i is the
1122 orientation preference of excitatory neuron i and Θ_D is the distractor's orientation. In addition, we
1123 enhanced the cross-hemifield inhibitory weights ($W_I^{VC,U \to VC,C}$ or $W_I^{VC,C \to VC,U}$ in equation 3)
1124 multiplicatively (by ~30x) during distractor presentation; this enabled modeling the bias away from
1125 the distractor's orientation observed in the distractor-opposite condition.

1126

1127 iii) Quantifying simulated behavioral and neural metrics. We ran a total of 10,000 simulations with
1128 a different random seed each time; all results represent an average of these simulations. During

each simulation, we randomly sampled the memory array (cued and uncued) and distractor orientation from a circular uniform random distribution. The "neural" orientation maintained by the network was decoded using the population vector representation in the visual cortex attractor given by:

$$\hat{\theta} = \arg\left(\sum_{j=1}^{N} r_{E,j}^{VC,C} \mathrm{e}^{i\theta_j}\right) \qquad (7)$$

where, $\hat{\theta}$ is the decoded orientation from the VC,C attractor. N is the number of E neurons. $r_{E,j}^{VC,C}$ and θ_j is the firing rate and preferred orientation respectively of neuron j in the VC,C attractor. Orientations from uncued visual cortex attractor were also decoded in an identical manner, using its respective population vector. The model's neural orientation decoding was computed with the VC attractors alone to mimic the decoding of neural orientation from only the occipital electrodes in the EEG data. By contrast, drawing inspiration from recent work that has highlighted the involvement of both sensory (visual) cortex (VC) and higher brain areas (HC) in robust working memory maintenance[3–5,9,13,18,41], the decoded orientation from the VC and HC attractors at the end of the delay period were averaged to generate the model's behavioral orientation estimate (see SI Results Section); orientation from the higher cortex attractor were decoded in a manner identical to that in equation 7, using their respective population vectors.

Statistical Analyses

We employed a non-parametric Wilcoxon signed rank test for comparing median errors in orientation reports (Fig. 1B, C). Permutation tests (10000 shuffles, unless otherwise indicated) were performed for behavioral bias across conditions; one-sided p-values are reported under the *apriori* hypothesis that the attractive (repulsive) biases occur when the distractor appeared on the same (opposite) hemifield as the memorandum (Fig. 1G-H; 2C-D; 3C-D, 4B-C). All the other tests

1152 were two-sided, unless otherwise stated. To study the effect of distractor condition and cueing on
1153 precision in visual WM (Fig. 1C), we performed a 2-way ANOVA with the circular standard
1154 deviation as the dependent variable and distractor condition (distractor-same, distractor-opposite
1155 and no-distractor) and cueing (cued and uncued) as predictors; participants was treated as
1156 random effects. *Post-hoc* analyses were performed by averaging the values across the cue
1157 conditions and comparing CSD values for distractor-same versus distractor-opposite conditions
1158 with Wilcoxon signed rank test. Permutation tests by shuffling (1000 shuffles) distractor-same and
1159 distractor-opposite labels, were used to compare precision degradation induced by the distractor
1160 across cued and uncued memoranda. To find significant time points in decoding accuracies
1161 plotted as a function of time (Fig. 2B), we used the two-sided cluster-based permutation test[79]
1162 with a cluster-forming threshold of $p<0.05$ and with 10,000 permutations of the data. The
1163 distribution of decoding accuracies across cued and uncued conditions was compared using
1164 permutation tests (two-sided) both for memorandum decoding and distractor decoding analysis
1165 (Fig. 2B, 3B). The distractor related ERPs across cued and uncued conditions were compared
1166 using the signed-rank test (Fig. 4A).

1167 In addition, we also report the Bayes factor (BF[80]), a statistical measure that compares the
1168 likelihood of the alternate hypothesis to that of the null hypothesis under the observed data. BFs
1169 were computed with a JZS prior, using the Bayes Factor toolbox[81]. One-tailed BFs were computed
1170 specifically when comparing the effect of distractor features on behavioral biases; other BFs were
1171 two-tailed, unless otherwise stated. Moreover, one-tailed BFs, and associated p-values, were also
1172 used to compare the strength of biases across the median splits, because we had clear, *apriori*
1173 hypotheses regarding the effect of each memorandum (decodability) or distractor characteristic
1174 (timing, decodability, gating) on bias magnitude. BFs computed based on a two-tailed versus one-
1175 tailed tests are reported as BF_{10} versus BF_{+0}/BF_{-0}, respectively; the subscripts +0 and -0 denote
1176 the directionality of the hypothesis tested (greater or lesser than zero, respectively).

Data and code availability

Data and custom Matlab code to replicate the experimental findings as well as the computational modeling results are available at the following link:

https://osf.io/dqr7m/?view_only=71feeb114f744f8c987d6a22f230905f

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the second author used ChatGPT and Open AI as a language editing tool. After using this tool/service, the second author reviewed and edited the content as needed and takes full responsibility for the content of the parts of the publication that was co-written by them.

Acknowledgements

We thank Rishabh Bajpai and Mirudhula Mukundan for their help in piloting the task design. This research was supported by a Wellcome Trust-Department of Biotechnology India Alliance Intermediate fellowship, DST Swarna-Jayanti fellowship, a Pratiksha Trust Intramural grant, an India-Trento Programme for Advanced Research (ITPAR) grant and a Department of Biotechnology-Indian Institute of Science Partnership Program grant (all to DS).

Author contributions

D.S. and S.G. designed the study; S.G. conducted the experiment; D.R and S.G. analyzed the data and performed model simulations; D.R, S.G. and D.S. wrote the paper.

References

1. Pasternak, T. & Greenlee, M. W. Working memory in primate sensory systems. *Nat Rev Neurosci* **6**, 97–107 (2005).

2. Emrich, S. M., Riggall, A. C., La Rocque, J. J. & Postle, B. R. Distributed patterns of activity in sensory cortex reflect the precision of multiple items maintained in visual short-term memory. *J Neurosci* **33**, 6516–6523 (2013).

3. Gayet, S., Paffen, C. L. E. & Van der Stigchel, S. Visual Working Memory Storage Recruits Sensory Processing Areas. *Trends Cogn Sci* **22**, 189–190 (2018).

4. Scimeca, J. M., Kiyonaga, A. & D'Esposito, M. Reaffirming the Sensory Recruitment Account of Working Memory. *Trends Cogn Sci* **22**, 190–192 (2018).

5. Rademaker, R. L., Chunharas, C. & Serences, J. T. Coexisting representations of sensory and mnemonic information in human visual cortex. *Nature Neuroscience 2019 22:8* **22**, 1336–1344 (2019).

6. Bettencourt, K. C. & Xu, Y. Decoding the content of visual short-term memory under distraction in occipital and parietal areas. *Nat Neurosci* **19**, (2015).

7. Stokes, M. G. 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn Sci* **19**, 394–405 (2015).

8. Lorenc, E. S., Mallett, R. & Lewis-Peacock, J. A. Distraction in Visual Working Memory: Resistance is Not Futile. *Trends Cogn Sci* **25**, 228–239 (2021).

9. Xu, Y. Reevaluating the Sensory Account of Visual Working Memory Storage. *Trends Cogn Sci* **21**, 794–815 (2017).

10. Lorenc, E. S., Sreenivasan, K. K., Nee, D. E., Vandenbroucke, A. R. E. & D'Esposito, M. Flexible coding of visual working memory representations during distraction. *Journal of Neuroscience* **38**, 5267–5276 (2018).

11. Hallenbeck, G. E., Sprague, T. C., Rahmati, M., Sreenivasan, K. K. & Curtis, C. E. Working memory representations in visual cortex mediate distraction effects. *Nature Communications 2021 12:1* **12**, 1–18 (2021).

12. Van Ede, F., Chekroud, S. R., Stokes, M. G. & Nobre, A. C. Decoding the influence of anticipatory states on visual perception in the presence of temporal distractors. *Nature Communications 2018 9:1* **9**, 1–12 (2018).

13. Xu, Y. Revisit once more the sensory storage account of visual working memory. *Vis cogn* **28**, 433–446 (2020).

14. Magnussen, S., Greenlee, M. W., Asplund, R. & Dyrnes, S. Stimulus-specific mechanisms of visual short-term memory. *Vision Res* **31**, 1213–1219 (1991).

15. Yoon, J. H., Curtis, C. E. & D'Esposito, M. Differential effects of distraction during working memory on delay-period activity in the prefrontal cortex and the visual association cortex. *Neuroimage* **29**, 1117–1126 (2006).

16. van Moorselaar, D., Gunseli, E., Theeuwes, J. & Olivers, C. N. L. The time course of protecting a visual memory representation from perceptual interference. *Front Hum Neurosci* **8**, (2015).

17. Rademaker, R. L., Bloem, I. M., De Weerd, P. & Sack, A. T. The impact of interference on short-term memory for visual orientation. *J Exp Psychol Hum Percept Perform* **41**, 1650–1665 (2015).

18. Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R. & Haynes, J. D. The Distributed Nature of Working Memory. *Trends Cogn Sci* **21**, 111–124 (2017).

19. Suzuki, M. & Gottlieb, J. Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nature Neuroscience 2012 16:1* **16**, 98–104 (2012).

20. Berryhill, M. E., Chein, J. & Olson, I. R. At the intersection of attention and memory: The mechanistic role of the posterior parietal lobe in working memory. *Neuropsychologia* **49**, (2011).

21. Christophel, T. B., Iamshchinina, P., Yan, C., Allefeld, C. & Haynes, J. D. Cortical specialization for attended versus unattended working memory. *Nature Neuroscience 2018 21:4* **21**, 494–496 (2018).

22. Brincat, S. L. *et al.* Interhemispheric transfer of working memories. *Neuron* **109**, 1055-1066.e4 (2021).

23. Parthasarathy, A. *et al.* Mixed selectivity morphs population codes in prefrontal cortex. *Nature Neuroscience 2017 20:12* **20**, 1770–1779 (2017).

24. Myers, N. E., Stokes, M. G. & Nobre, A. C. Prioritizing Information during Working Memory: Beyond Sustained Internal Attention. *Trends Cogn Sci* **21**, 449–461 (2017).

25. Lorenc, E. S., Mallett, R. & Lewis-Peacock, J. A. Distraction in Visual Working Memory: Resistance is Not Futile. *Trends Cogn Sci* **25**, 228–239 (2021).

26. Feldmann-Wüstefeld, T. & Vogel, E. K. Neural Evidence for the Contribution of Active Suppression During Working Memory Filtering. *Cerebral Cortex* **29**, (2019).

27. Wimmer, K., Nykamp, D. Q., Constantinidis, C. & Compte, A. Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience 2014 17:3* **17**, 431–439 (2014).

28. 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. *Cerebral Cortex* **10**, (2000).

29. Bays, P. M., Catalao, R. F. G. & Husain, M. The precision of visual working memory is set by allocation of a shared resource. *J Vis* **9**, (2009).

30. Myers, N. E., Chekroud, S. R., Stokes, M. G. & Nobre, A. C. Benefits of Flexible Prioritization in Working Memory Can Arise Without Costs. *J Exp Psychol Hum Percept Perform* **44**, 398 (2018).

31. Akaike, H. A New Look at the Statistical Model Identification. *IEEE Trans Automat Contr* **19**, 716–723 (1974).

32. Schwarz, G. Estimating the Dimension of a Model. *https://doi.org/10.1214/aos/1176344136* **6**, 461–464 (1978).

33. Wolff, M. J., Jochim, J., Akyürek, E. G. & Stokes, M. G. Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience 2017 20:6* **20**, 864–871 (2017).

34. Wolff, M. J., Jochim, J., Akyürek, E. G., Buschman, T. J. & Stokes, M. G. Drifting codes within a stable coding scheme for working memory. *PLoS Biol* **18**, e3000625 (2020).

35. Polich, J. Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology* vol. 118 Preprint at https://doi.org/10.1016/j.clinph.2007.04.019 (2007).

36. Huang, W. J., Chen, W. W. & Zhang, X. The neurophysiology of P 300 - An integrated review. *Eur Rev Med Pharmacol Sci* **19**, 1480–1488 (2015).

37. Berti, S. & Schröger, E. A comparison of auditory and visual distraction effects: Behavioral and event-related indices. *Cognitive Brain Research* **10**, (2001).

38. Lijffijt, M. *et al.* P50, N100, and P200 sensory gating: Relationships with behavioral inhibition, attention, and working memory. *Psychophysiology* **46**, 1059 (2009).

39. Wongupparaj, P., Sumich, A., Wickens, M., Kumari, V. & Morris, R. G. Individual differences in working memory and general intelligence indexed by P200 and P300: A latent variable model. *Biol Psychol* **139**, 96–105 (2018).

40. Hillyard, S. A. & Anllo-Vento, L. Event-related brain potentials in the study of visual selective attention. *Proc Natl Acad Sci U S A* **95**, 781–787 (1998).

41. Pasternak, T. & Greenlee, M. W. Working memory in primate sensory systems. *Nat Rev Neurosci* **6**, 97–107 (2005).

42. Sreenivasan, K. K., Vytlacil, J. & D'Esposito, M. Distributed and dynamic storage of working memory stimulus information in extrastriate cortex. *J Cogn Neurosci* **26**, 1141–1153 (2014).

43. Mejías, J. F. & Wang, X. J. Mechanisms of distributed working memory in a large-scale network of macaque neocortex. *Elife* **11**, (2022).

44. Voitov, I. & Mrsic-Flogel, T. D. Cortical feedback loops bind distributed representations of working memory. *Nature 2022 608:7922* **608**, 381–389 (2022).

45. Khona, M. & Fiete, I. R. Attractor and integrator networks in the brain. *Nature Reviews Neuroscience* vol. 23 Preprint at https://doi.org/10.1038/s41583-022-00642-0 (2022).

46. Sakai, K., Rowe, J. B. & Passingham, R. E. Active maintenance in prefrontal area 46 creates distractor-resistant memory. *Nat. Neurosci.* **5**, 479–484 (2002).

47. Wyatte, D., Herd, S., Mingus, B. & O'Reilly, R. The Role of Competitive Inhibition and Top-Down Feedback in Binding during Object Recognition. *Front Psychol* **3**, (2012).

48. Dakos, A. S., Walker, E. M., Jiang, H., Stein, B. E. & Rowland, B. A. Interhemispheric visual competition after multisensory reversal of hemianopia. *European Journal of Neuroscience* **50**, 3702–3712 (2019).

49. Szczepanski, S. M. & Kastner, S. Shifting Attentional Priorities: Control of Spatial Attention through Hemispheric Competition. *Journal of Neuroscience* **33**, 5411–5421 (2013).

50. Reynolds, J. H., Chelazzi, L. & Desimone, R. Competitive Mechanisms Subserve Attention in Macaque Areas V2 and V4. *Journal of Neuroscience* **19**, 1736–1753 (1999).

51. Chinchani, A. M. *et al.* Tracking momentary fluctuations in human attention with a cognitive brain-machine interface. *Communications Biology 2022 5:1* **5**, 1–17 (2022).

52. Stokes, M. G. 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences* vol. 19 Preprint at https://doi.org/10.1016/j.tics.2015.05.004 (2015).

53. Panichello, M. F., DePasquale, B., Pillow, J. W. & Buschman, T. J. Error-correcting dynamics in visual working memory. *Nat Commun* **10**, (2019).

54. Renart, A., Song, P. & Wang, X. J. Robust spatial working memory through homeostatic synaptic scaling in heterogeneous cortical networks. *Neuron* **38**, (2003).

55. Barbosa, J. *et al.* Interplay between persistent activity and activity-silent dynamics in the prefrontal cortex underlies serial biases in working memory. *Nat Neurosci* **23**, 1016–1024 (2020).

56. Michel, C. M. & Brunet, D. EEG source imaging: A practical review of the analysis steps. *Front Neurol* **10**, (2019).

57. Delvenne, J. F. The capacity of visual short-term memory within and between hemifields. *Cognition* **96**, B79–B88 (2005).

58. Umemoto, A., Drew, T., Ester, E. F. & Awh, E. A bilateral advantage for storage in visual working memory. *Cognition* **117**, 69–79 (2010).

59. Adam, K. C. S., Robison, M. K. & Vogel, E. K. Contralateral Delay Activity Tracks Fluctuations in Working Memory Performance. *J Cogn Neurosci* **30**, 1229 (2018).

60. Luria, R., Balaban, H., Awh, E. & Vogel, E. K. The contralateral delay activity as a neural measure of visual working memory. *Neurosci Biobehav Rev* **62**, 100 (2016).

61. Souza, A. S. & Oberauer, K. In search of the focus of attention in working memory: 13 years of the retro-cue effect. *Atten Percept Psychophys* **78**, (2016).

62. Mongillo, G., Barak, O. & Tsodyks, M. SynaptiC Theory of Working Memory. *Science (1979)* **319**, 1543–1546 (2008).

63. Ford, J. M. *et al.* The relationship between P300 amplitude and regional gray matter volumes depends upon the attentional system engaged. *Electroencephalogr Clin Neurophysiol* **90**, (1994).

64. Eissa, T. L. & Kilpatrick, Z. P. Learning efficient representations of environmental priors in working memory. *bioRxiv* 2022.07.05.498889 (2022) doi:10.1101/2022.07.05.498889.

65. de Vries, I. E. J., Slagter, H. A. & Olivers, C. N. L. Oscillatory Control over Representational States in Working Memory. *Trends in Cognitive Sciences* vol. 24 Preprint at https://doi.org/10.1016/j.tics.2019.11.006 (2020).

66. Vijayraghavan, S. & Everling, S. Neuromodulation of Persistent Activity and Working Memory Circuitry in Primate Prefrontal Cortex by Muscarinic Receptors. *Frontiers in Neural Circuits* vol. 15 Preprint at https://doi.org/10.3389/fncir.2021.648624 (2021).

67. D'Esposito, M. & Postle, B. R. The cognitive neuroscience of working memory. *Annu Rev Psychol* **66**, 115–142 (2015).

68. Cavanagh, S. E., Towers, J. P., Wallis, J. D., Hunt, L. T. & Kennerley, S. W. Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. *Nat Commun* **9**, (2018).

69. Berezuk, C., Scott, S. C., Black, S. E. & Zakzanis, K. K. Cognitive reserve, cognition, and real-world functioning in MCI: A systematic review and meta-analysis. *Journal of Clinical and Experimental Neuropsychology* vol. 43 Preprint at https://doi.org/10.1080/13803395.2022.2047160 (2021).

70. Myers, N. E., Chekroud, S. R., Stokes, M. G. & Nobre, A. C. Benefits of flexible prioritization in working memory can arise without costs. *J Exp Psychol Hum Percept Perform* **44**, 398 (2017).

71. Fischer, J. & Whitney, D. Serial dependence in visual perception. *Nature Neuroscience 2014 17:5* **17**, 738–743 (2014).

72. Bays, P. M. & Husain, M. Dynamic shifts of limited working memory resources in human vision. *Science (1979)* **321**, 851–854 (2008).

73. Ma, W. J., Husain, M. & Bays, P. M. Changing concepts of working memory. *Nature Neuroscience 2014 17:3* **17**, 347–356 (2014).

74. Delorme, A. & Makeig, S. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods* **134**, 9–21 (2004).

75. Lopez-Calderon, J. & Luck, S. J. ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Front Hum Neurosci* **8**, (2014).

76. Rolston, J. D., Gross, R. E. & Potter, S. M. Common median referencing for improved action potential detection with multielectrode arrays. *Annu Int Conf IEEE Eng Med Biol Soc* **2009**, 1604–1607 (2009).

77. Pion-Tonachini, L., Kreutz-Delgado, K. & Makeig, S. ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *Neuroimage* **198**, 181–197 (2019).

78. Junghöfer, M., Elbert, T., Tucker, D. M. & Rockstroh, B. Statistical control of artifacts in dense array EEG/MEG studies. *Psychophysiology* **37**, 523–532 (2000).

79. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods* **164**, 177–190 (2007).

80. Jarosz, A. F. & Wiley, J. What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *J Probl Solving* **7**, 2 (2014).

81. Bart Krekelberg. Bayes factor toolbox. Preprint at https://zenodo.org/records/13744717 (2024).

82. Feredoes, E., Heinen, K., Weiskopf, N., Ruff, C. & Driver, J. Causal evidence for frontal involvement in memory target maintenance by posterior brain areas during distracter interference of visual working memory. *Proc Natl Acad Sci U S A* **108**, (2011).

83. Dotson, N. M., Hoffman, S. J., Goodell, B. & Gray, C. M. Feature-Based Visual Short-Term Memory Is Widely Distributed and Hierarchically Organized. *Neuron* **99**, 215-226.e4 (2018).

Supplementary Information for

Neural bases of space-specific distractor biases in visual working memory

Deepak V Raya, Sanchit Gupta, and Devarajan Sridharan

Supplementary Information: Results

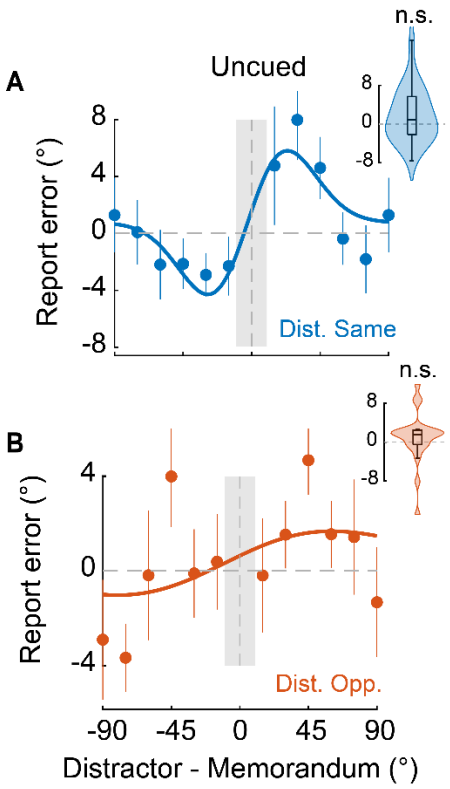
Error-correcting dynamics in a two-tier attractor model

We present here additional simulations clarifying the two-tier architecture for the WM attractor model employed in our simulations. We employed a two-tier model, rather than a single tier model for modelling robust mnemonic maintenance. Recent literature suggests that mnemonic representations in higher order brain regions are less susceptible to interference from distractors than those in early sensory cortices[1–4]. Top-down feedback from the higher order brain region has been hypothesized to help mitigate distractor interference in the early sensory cortices[1,2,4–6]. To test whether a similar mechanism could also be modeled, we performed simulations (n=10^4, across 10 runs) by varying the strength of top-down feedback from the HC to VC attractors in both hemifields ($W_E^{HC\rightarrow VC}$ term in equation 4, main text) by a factor of 2x. Indeed, increasing the top-down feedback to the VC markedly reduced maintenance error for both the cued and the uncued VC attractor (SI Fig. S6A, CSD reduced (%): VC,C=5.94±0.55%, range: 4.19 to 8.11%; VC,U= 8.47±0.68%, range: 4.61 to 12.12%). In our model, such error correcting dynamics from the HC were essential to overcome the disruptive effect of the distractor on the VC; they helped reinstate an accurate mnemonic representation in the VC attractor following distractor disappearance (Fig. 5C-D, main text).

Moreover, during recall, the behavioral feature estimate of the memorandum was computed from both VC and HC attractor representations. Recent research indicates that visual WM maintenance of involves multiple brain regions, including both sensory and higher-order areas[7–11]. It is possible, then, that a (weighted) average of the memorandum's feature representation across multiple brain regions contributes to a more stable behavioral readout. We hypothesize that such a readout improves precision by cancelling out independent noise and drift in mnemonic representations across different brain regions. We tested this hypothesis by

observing the average precision of “behavioral” readouts over n=10,000 simulations based on the decoded representations of each memorandum’s orientation, at the end of the delay period. In a first set of simulations, these readouts were estimated from the VC attractor alone (equation 7, main text). In a second set of simulations, we averaged the readouts from both the VC and the HC. The averaged readout was numerically more accurate as compared to the estimate from any single brain region alone (SI Fig. S6B, CSD reduction (%): Cued=1.18±0.2%, range: 0.58 to 2.68%; Uncued=2.12±0.15%, range: 1.48 to 2.71%). Consequently, in the simulations in the main text, we averaged the HC and VC attractor readouts to obtain the model’s behavioral readout.

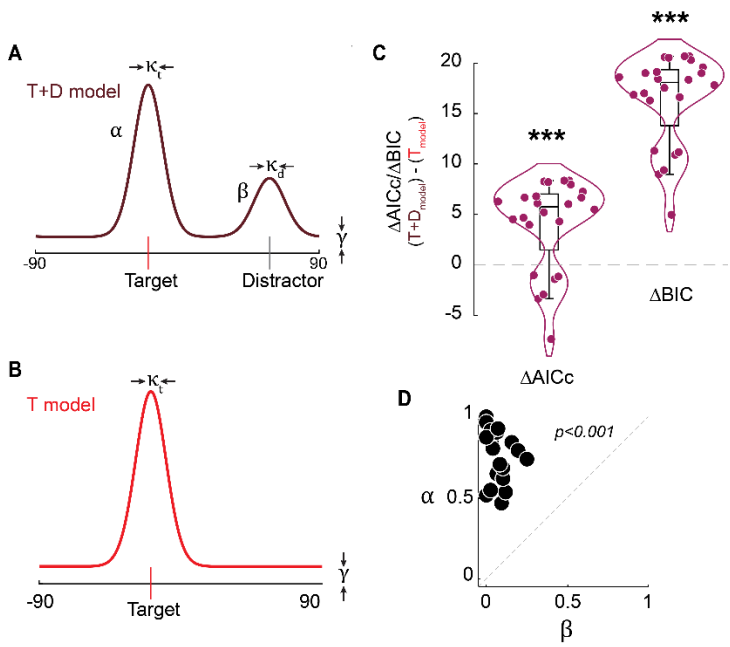
Supplementary Information: Figures and Tables



Supplementary Figure S1. Distractor-induced behavioral biases for the uncued memorandum.

A. Same as in Figure 1G (main text) but showing the average behavioral bias for the uncued memorandum when the distractor appeared in the same hemifield as the former. Other conventions are the same as in Figure 1G.

B. Same as in panel A, but for trials in which the distractor appeared in the hemifield opposite to the uncued memorandum. Other conventions are the same as in panel A and Figure 1H (main text).



Supplementary Figure S2. von Mises mixture model fit and evaluation.

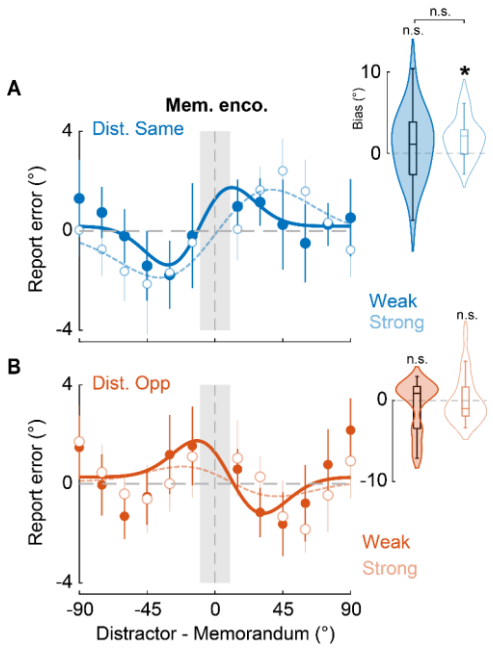
A. Schematic of a mixture model comprising two von Mises and a uniform distribution in which participants responses are modeled as a combination of target (memorandum) and distractor orientation reports, and guesses ("target+distractor"/T+D model). κ_t and κ_d denote the precision associated with the target and distractor responses respectively. α and β denote the contribution of the target and distractor to the orientation responses whereas, γ denotes the contribution of the uniform component ($\alpha + \beta + \gamma = 1$).

B. Same as in A, but for a mixture model comprising one von Mises and a uniform distribution to model participants' responses based on target orientation reports ("target alone"/T model) and guesses alone. Other conventions are the same as in panel A.

C. Model comparisons. (*Left*) Difference between corrected Akaike information criterion (AICc) values for the "target+distractor" model fit and "target alone" model fit. A lower (more negative) value indicates a preferred model. Individual points: Participants. Violins, and box and whisker

plots follow the same conventions as in Figure 1F-G (insets). (*Right*) Same as in the left panel but for the Bayesian information criterion (BIC). Asterisks: significant difference of median values between models, based on permutation tests. *p<0.05, **p<0.01, ***p<0.001, n.s.: not significant.

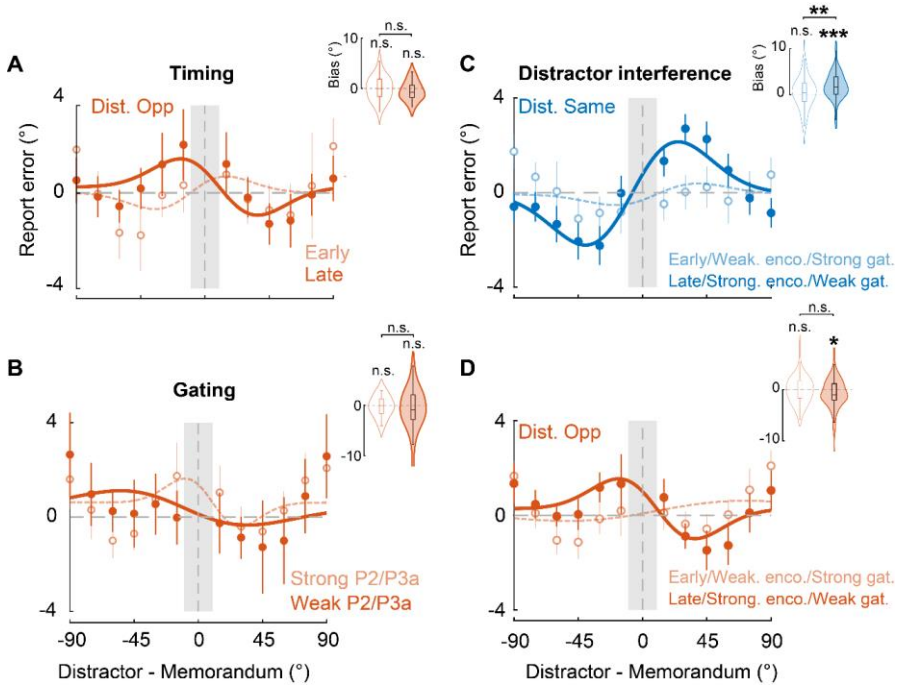
D. Scatter showing the distribution of α (target weight) and β (distractor weight) obtained by fitting participants' responses with the "target+distractor" combined mixture model. Points: Individual participants. Dashed diagonal line: line of equality.



Supplementary Figure S3. Effect of memorandum encoding on distractor-induced bias.

A. Same as in Figure 2C (main text), but showing behavioral bias curves for the cued memorandum based on a median split of memorandum encoding strength – quantified with neural decoding accuracy following memorandum presentation (gray shaded bar, Fig. 2B, left) – when the distractor appeared in the same hemifield as the memorandum (n=23). Solid curve and filled violin plot (*inset*): Weaker memorandum encoding trials. Dashed curve and open violin plot (*inset*). Stronger memorandum encoding trials.

B. Same as in panel A, but when the distractor appeared in the hemifield opposite to the cued memorandum. Other conventions are the same as in Figures 2C-D (main text).



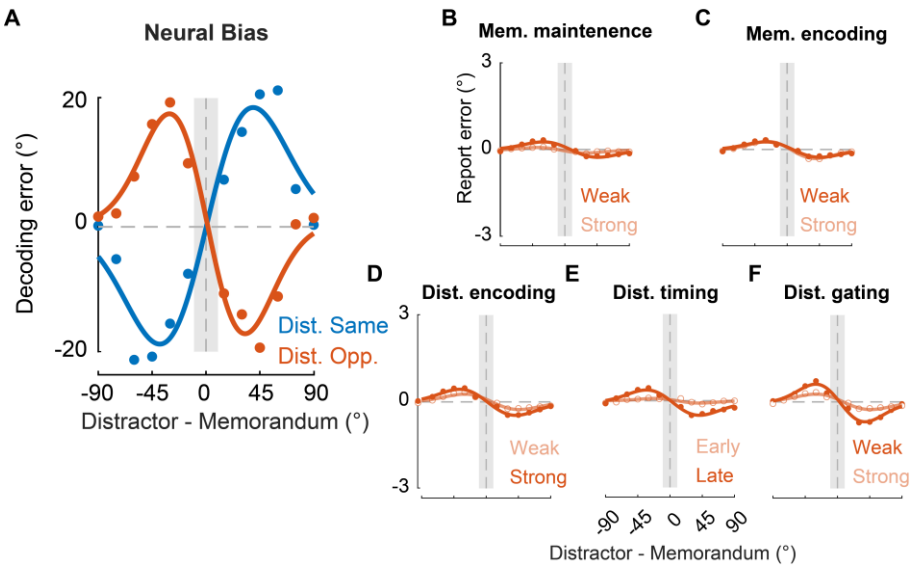
Supplementary Figure S4. Distractor-opposite and omnibus analysis results.

A. Same as in Figure 4B-C (main text) but showing the behavioral bias curves for the cued memorandum when the distractor appeared on the hemifield opposite to the cued memorandum – for trials in which the distractor appeared early (dashed, open symbols, unfilled violins) or late (solid, filled symbols, filled violins) during the delay period.

B. Same as panel A, but for trials with strong (dashed) or weak (solid) distractor-evoked P2/P3a amplitude. (A-B) Other conventions are the same as in Figures 4B-C.

C. Same as panel Figure 4B (main text), but showing the results of an omnibus analysis combining across conditions with high distractor interference – late timing, strong encoding, weak gating (solid, filled symbols, filled violins) – or low distractor interference – early timing, weak encoding, strong gating (dashed, open symbols, unfilled violins) – for trials in which the distractor appeared on the same hemifield as the cued memorandum.

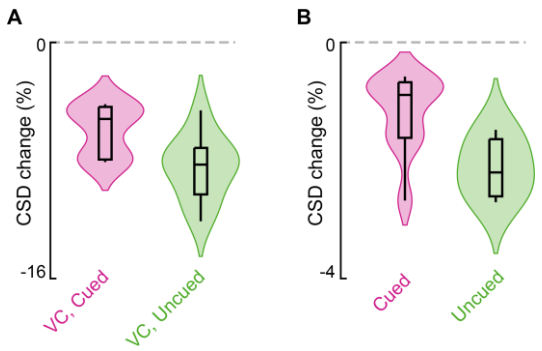
1487 **D.** Same as panel C, but when the distractor appeared on hemifield opposite to the cued
1488 memorandum. (C-D), other conventions are the same as in Figure 4B (main text) and SI Figure
1489 S4A. (All panels) *: $p<0.05$; **: $p<0.01$; ***: $p<0.001$; n.s.: not significant.



Supplementary Figure S5. Simulations of neural bias and distractor-opposite effects.

A. Simulated neural bias for the cued memorandum when the distractor appeared in the same hemifield (blue) or the opposite hemifield (orange) as the former. Other conventions are the same as in Figure 6A (main text).

B-F. Same corresponding panels as in Figure 6B-F (main text), except showing simulated behavioral bias when the distractor appeared in hemifield opposite to the respective memorandum. Other conventions are the same as in corresponding panels in Figures 6B-F.



Supplementary Figure S6. Error-correcting dynamics in WM.

A. Change in error in the VC attractor readout upon incorporating top-down feedback from the HC attractor. Data shown separately for the cued (magenta) and uncued (green) attractors. Violin plots showing the distribution of values across 10 runs of the model (n=1000 simulations for each run); each simulation initialized with different random seeds, and values were averaged across simulations within a run.

B. Change in error obtained by averaging readouts from the VC and the HC attractor relative to the mean readout error for either attractor (average of VC and HC readout errors). Other conventions are the same as in panel A.

Supplementary Table S1. Network architecture and simulation parameters employed in the model.

A. Network architecture parameters

Parameter	Description	Value
$W_I^{VC,U \to VC,C}$	Inter-hemispheric inhibitory connection strength from uncued visual cortex to cued visual cortex	0.02
$W_I^{VC,C \to VC,U}$	Inter-hemispheric inhibitory connection strength from cued visual cortex to uncued visual cortex	0.02
$W_E^{HC,C \to VC,C}$	Top-down connection strength from higher cortex to visual cortex on the cued hemifield before/after cue onset	0.1/0.12
$W_E^{HC,U \to VC,U}$	Top-down connection strength from higher cortex to visual cortex on the uncued hemifield before/after cue onset	0.1/0.1
$W_E^{VC,C \to HC,C}$	Bottom-up connection strength from visual cortex to higher cortex on the cued hemifield during the stimulus/delay periods	3.0/0.0
$W_E^{VC,U \to HC,U}$	Bottom-up connection strength from visual cortex to higher cortex on the uncued hemifield during the stimulus/delay periods	3.0/0.0
W_{EE}	Strength of local recurrent connections among excitatory neurons	6*
W_{EI}	Strength of global recurrent connections from excitatory to inhibitory neurons	4*
W_{IE}	Strength of global recurrent connections from inhibitory to excitatory neurons	3.4*
W_{IE}	Strength of global recurrent connections among inhibitory neurons	0.85*

*these parameter values were identical for all four attractors (VC,C; VC,U; HC,C; HC,U)

B. Simulation parameters

Parameter	Description	Value
τ_E	Time constant of E neurons, in ms	20
τ_I	Time constant of I neurons, in ms	10
A_M	Peak amplitude of the von Mises input current profile for the memory array	200
A_D	Peak amplitude of the von Mises input current profile for the distractor	10^3
κ_M	Concentration parameter of the von Mises input current profile for the stimulus	2.7
κ_D	Concentration parameter of the von Mises input current profile for the distractor	2.25
C_E	Constant bias current input to excitatory neurons	0.2*
C_I	Constant bias current input to inhibitory neurons	0.5*
σ_E	Standard deviation of input noise for excitatory neurons	3*
σ_I	Standard deviation of input noise for inhibitory neurons	5*

*these parameter values were identical for all four attractors (VC,C; VC,U; HC,C; HC,U)

Supplementary References

1. Lorenc, E. S., Mallett, R. & Lewis-Peacock, J. A. Distraction in Visual Working Memory: Resistance is Not Futile. *Trends Cogn Sci* **25**, 228–239 (2021).

2. Xu, Y. Reevaluating the Sensory Account of Visual Working Memory Storage. *Trends Cogn Sci* **21**, 794–815 (2017).

3. Xu, Y. Revisit once more the sensory storage account of visual working memory. *Vis cogn* **28**, 433–446 (2020).

4. Myers, N. E., Stokes, M. G. & Nobre, A. C. Prioritizing Information during Working Memory: Beyond Sustained Internal Attention. *Trends Cogn Sci* **21**, 449–461 (2017).

5. Suzuki, M. & Gottlieb, J. Distinct neural mechanisms of distractor suppression in the frontal and parietal lobe. *Nature Neuroscience 2012 16:1* **16**, 98–104 (2012).

6. Feredoes, E., Heinen, K., Weiskopf, N., Ruff, C. & Driver, J. Causal evidence for frontal involvement in memory target maintenance by posterior brain areas during distracter interference of visual working memory. *Proc Natl Acad Sci U S A* **108**, (2011).

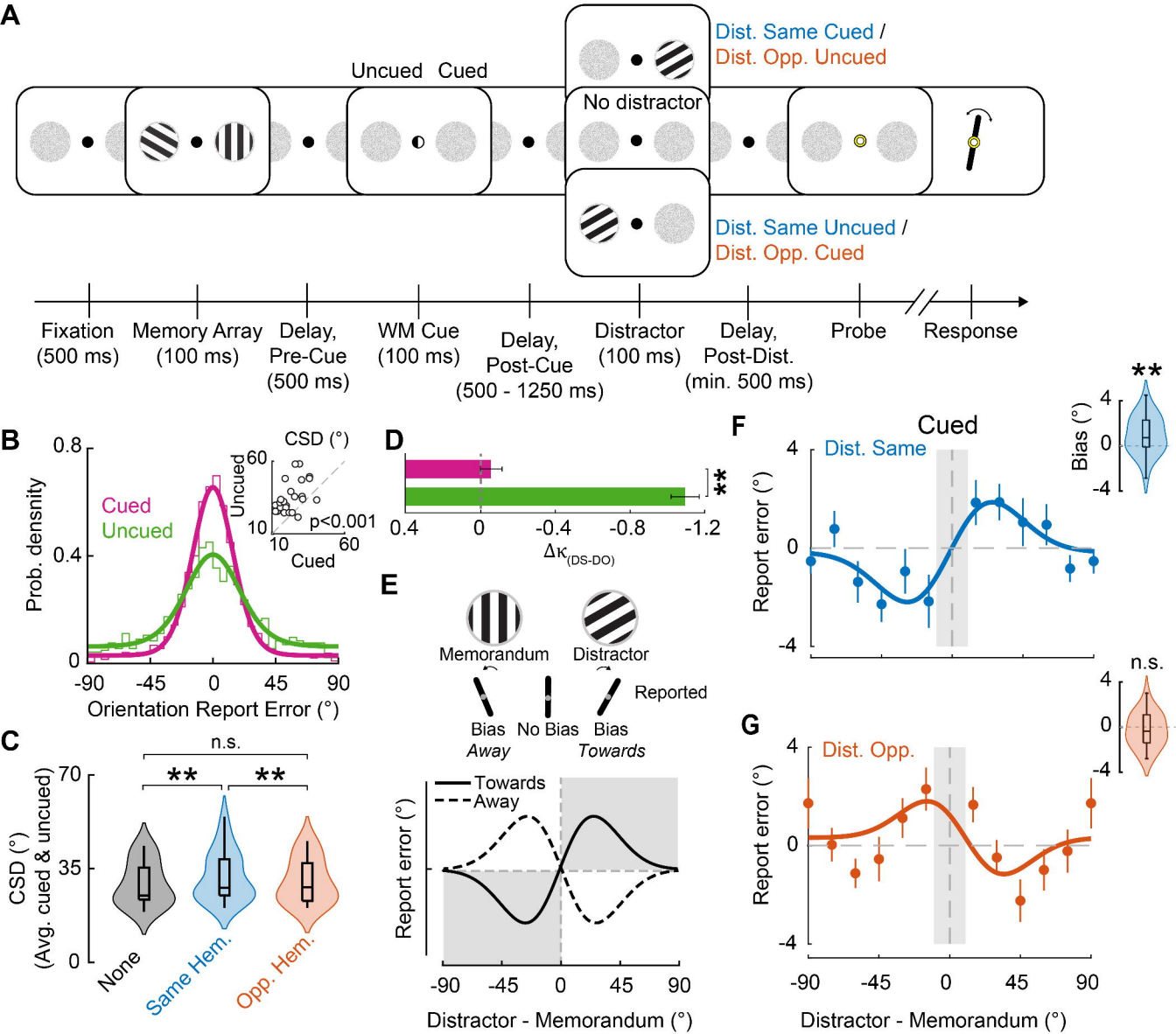
7. Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R. & Haynes, J. D. The Distributed Nature of Working Memory. *Trends Cogn Sci* **21**, 111–124 (2017).

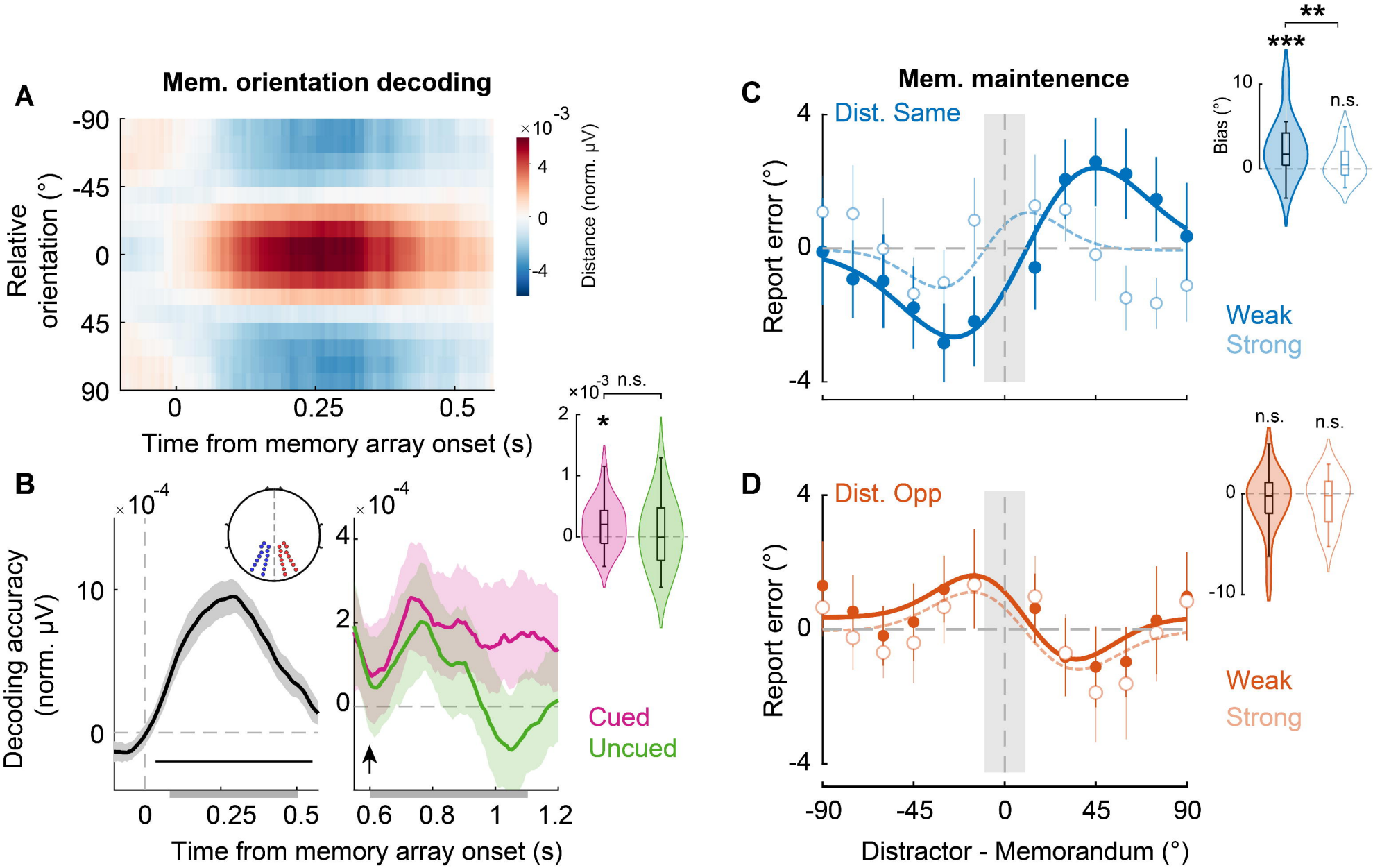
8. Sreenivasan, K. K., Vytlacil, J. & D'Esposito, M. Distributed and dynamic storage of working memory stimulus information in extrastriate cortex. *J Cogn Neurosci* **26**, 1141–1153 (2014).

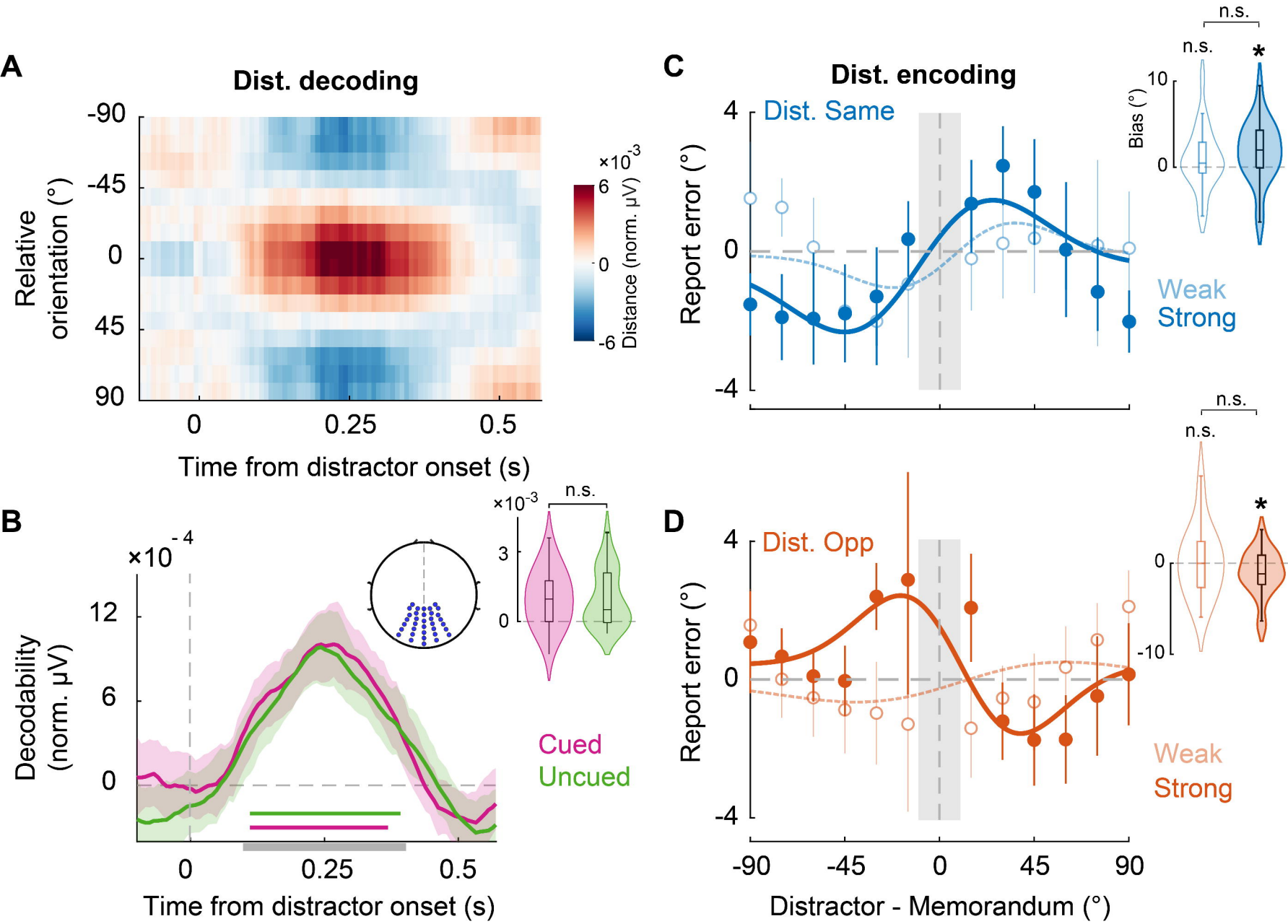
9. Mejías, J. F. & Wang, X. J. Mechanisms of distributed working memory in a large-scale network of macaque neocortex. *Elife* **11**, (2022).

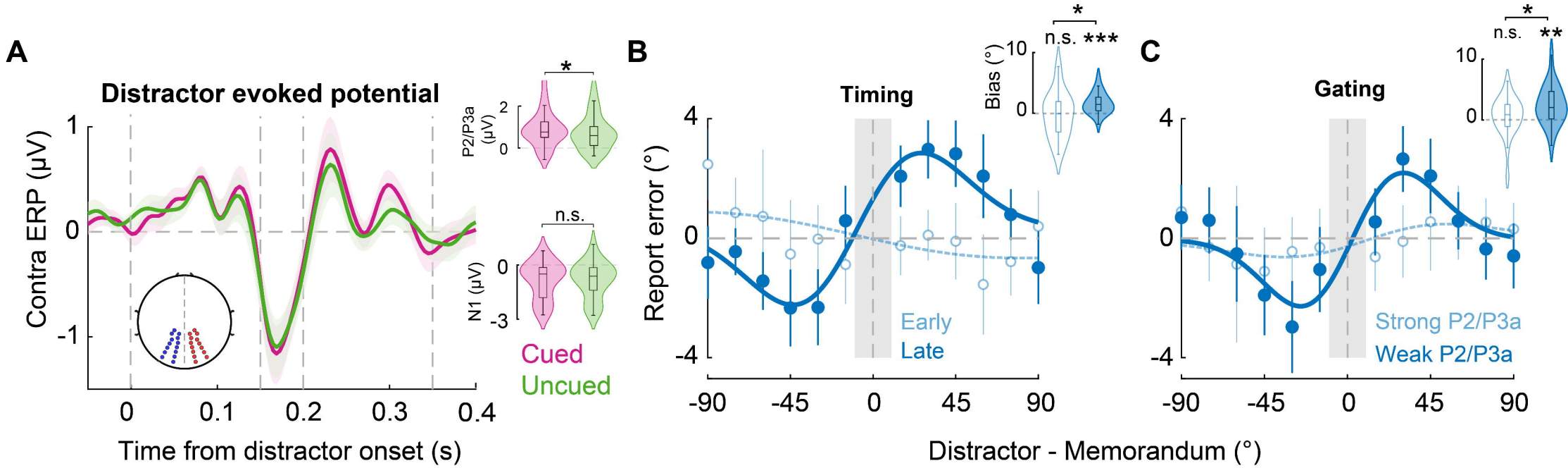
10. Dotson, N. M., Hoffman, S. J., Goodell, B. & Gray, C. M. Feature-Based Visual Short-Term Memory Is Widely Distributed and Hierarchically Organized. *Neuron* **99**, 215-226.e4 (2018).

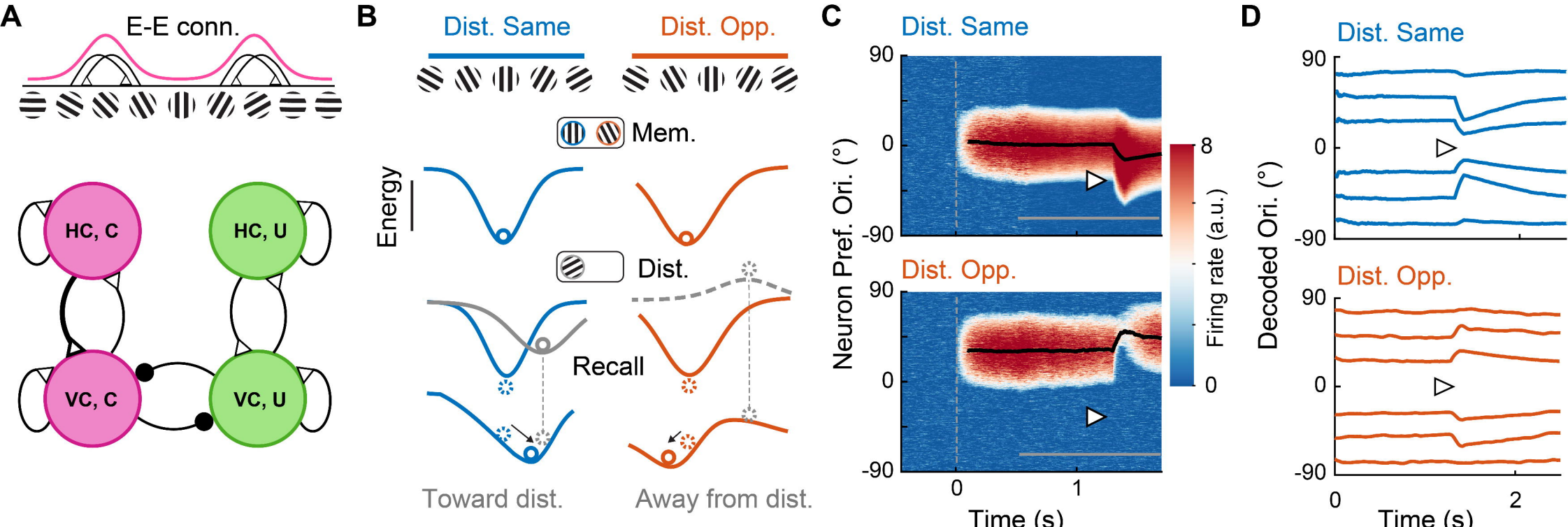
11. Voitov, I. & Mrsic-Flogel, T. D. Cortical feedback loops bind distributed representations of working memory. *Nature 2022 608:7922* **608**, 381–389 (2022).

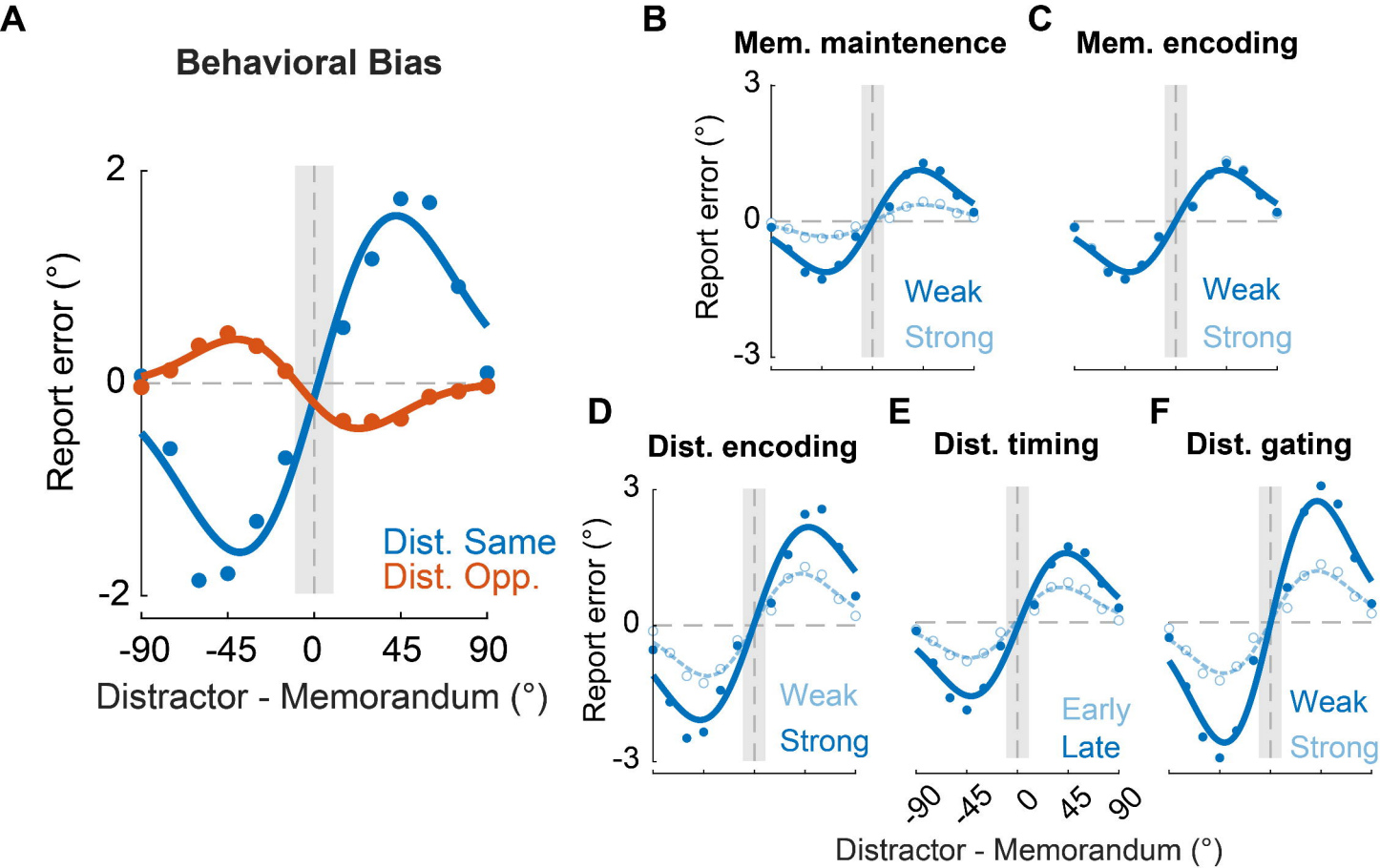

A
Uncued
Cued
No distractor
Dist. Same Cued /
Dist. Opp. Uncued
Dist. Same Uncued /
Dist. Opp. Cued
Fixation (500 ms)
Memory Array (100 ms)
Delay, Pre-Cue (500 ms)
WM Cue (100 ms)
Delay, Post-Cue (500 - 1250 ms)
Distractor (100 ms)
Delay, Post-Dist. (min. 500 ms)
Probe
Response
B
Prob. density
Cued
Uncued
CSD (°)
p<0.001
Orientation Report Error (°)
C
CSD (°) (Avg. cued & uncued)
n.s.
None
Same Hem.
Opp. Hem.
D
Δκ(DS-DO)
E
Memorandum
Distractor
Reported
Bias Away
No Bias
Bias Towards
Towards
Away
Report error (°)
Distractor - Memorandum (°)
F
Cued
Dist. Same
Bias (°)
G
Dist. Opp.


A
Mem. orientation decoding
Relative orientation (°)
-90
-45
0
45
90
0
0.25
0.5
Time from memory array onset (s)
×10^-3
4
0
-4
Distance (norm. µV)
B
Decoding accuracy (norm. µV)
×10^-4
10
0
0
0.25
0.5
×10^-4
4
2
0
0.6
0.8
1
1.2
Time from memory array onset (s)
Cued
Uncued
×10^-3
n.s.
2
1
0
*
C
Mem. maintenence
Dist. Same
Report error (°)
4
0
-4
Weak
Strong
Bias (°)
10
0
***
**
n.s.
D
Dist. Opp
Report error (°)
4
0
-4
-90
-45
0
45
90
Distractor - Memorandum (°)
Weak
Strong
0
-10
n.s.
n.s.


A
Dist. decoding
Relative orientation (°)
-90
-45
0
45
90
0
0.25
0.5
Time from distractor onset (s)
×10-3
6
0
-6
Distance (norm. µV)
B
×10-4
12
6
0
Decodability (norm. µV)
0
0.25
0.5
Time from distractor onset (s)
Cued
Uncued
×10-3
3
0
n.s.
C
Dist. encoding
Dist. Same
Report error (°)
4
0
-4
Weak
Strong
Bias (°)
10
0
n.s.
n.s.
*
D
Dist. Opp
Report error (°)
4
0
-4
Weak
Strong
0
-10
n.s.
n.s.
*
-90
-45
0
45
90
Distractor - Memorandum (°)


A
Distractor evoked potential
Contra ERP (µV)
1
0
-1
0
0.1
0.2
0.3
0.4
Time from distractor onset (s)
P2/P3a (µV)
2
0
*
N1 (µV)
0
-3
n.s.
Cued
Uncued
B
Timing
Report error (°)
4
0
-4
-90
-45
0
45
90
Early
Late
Bias (°)
10
0
n.s.
***
*
C
Gating
Strong P2/P3a
Weak P2/P3a
n.s.
**
*
Distractor - Memorandum (°)


A
E-E conn.
HC, C
HC, U
VC, C
VC, U
B
Dist. Same
Dist. Opp.
Mem.
Energy
Dist.
Recall
Toward dist.
Away from dist.
C
Dist. Same
Dist. Opp.
Neuron Pref. Ori. (°)
90
0
-90
Firing rate (a.u.)
8
0
Time (s)
0
1
D
Dist. Same
Dist. Opp.
Decoded Ori. (°)
90
0
-90
Time (s)
0
2


A
Behavioral Bias
Report error (°)
2
0
-2
Dist. Same
Dist. Opp.
-90
-45
0
45
90
Distractor - Memorandum (°)
B
Mem. maintenence
C
Mem. encoding
3
0
-3
Weak
Strong
D
Dist. encoding
E
Dist. timing
F
Dist. gating
Early
Late
Distractor - Memorandum (°)