Distractors induce space-specific neural biases in visual working memory

Running title: Space-specific distractor effects on mnemonic biases

Authors & Affiliations:

Sanchit Gupta[1,2,*] and Devarajan Sridharan[1,3,*]

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

[2]IMI Ph.D. Program, Indian Institute of Science, Bangalore 560012, India

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

***Correspondence to:**

Sanchit Gupta,
IMI Graduate Program,
Centre for Neuroscience,
Indian Institute of Science,
C. V. Raman Avenue,
Bangalore 560012, India
E-mail: sanchitgupta@iisc.ac.in

Devarajan Sridharan,
Associate Professor and
DST Swarna Jayanti 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 in working memory is remarkably resilient to distraction. Yet, recent evidence suggests that distractors containing task-relevant features can disrupt working memory by inducing subtle biases in mnemonic representations. With multivariate decoding of human electroencephalography recordings, we show that temporally unpredictable distractors produce spatially-antagonistic mnemonic biases, across the visual hemifields. Grating distractors produced either an attractive or a repulsive mnemonic bias – a shift in the neural representation of the memorandum toward or away from the distractor's orientation – depending, respectively, on whether the distractor appeared in the same hemifield as the memorandum, or opposite to it. Behavioral biases closely tracked these neural effects. We devised a two-tier ring attractor model with cross-hemifield inhibition, which comprehensively explains how the distractor's timing, encoding strength, and input gating control these mnemonic biases. Our results provide a mechanistic account of distractor-induced biases, across space and time, in visual working memory.

Keywords: distractor interference; attractive and repulsive biases; mnemonic representation; multivariate decoding; inter-hemispheric inhibition; ring attractor model.

Lay Summary

Working memory – the capacity to momentarily store important items or events – is remarkably robust to distraction. Yet, when a salient distractor with features resembling the memorized items appears, it can subtly alter (bias) working memory. Applying state-of-the-art neural decoding to human electrophysiological data, we show the such distractor biases are space-specific. The neural memory representation begins to resemble the distractor (attractive bias) when the distractor appears at the same location as the memory items, but becomes less like the distractor (repulsive bias) when the distractor appears further away. We tie together these experimental findings with a simple, computational model. The findings enhance our understanding of neural mechanisms underlying working memory and may be relevant for cognitive therapies addressing memory 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 drive, 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 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 distractors may interfere, albeit weakly, with information in WM[5,8,10,11] with some distractors being more potent than others[8,9,11–16]. In particular, distractors may induce subtle, yet reliable, biases in working memory. Specifically, 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 distractor biases have been reported even for complex feature combinations. 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 biases, in neural 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 biased 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, it is the stable coding in higher cortical areas that is hypothesized to mediate the remarkable resilience of WM to distracting sensory inputs[8,24].

Yet, previous studies on distractor biases overwhelmingly employed conventional and, often, simplistic WM task designs: for example, most tasks involved a single, central distractor presented at fixed times during the delay epoch[5,8–12,14–16,25]. As a result, how distractors that occur unpredictably in space and time bias mnemonic representations remains unknown: no previous study, to our knowledge, has explored such spatiotemporal bias effects by systematically varying the distractor's location and its timing relative to the memorandum.

Moreover, the vast majority of previous studies investigating mnemonic bias employed functional magnetic resonance imaging (fMRI) to measure brain responses[5,10,11]. The poor temporal resolution of fMRI precludes addressing fundamental questions about the rapid neural mechanisms that mediate distractor-induced biasing of mnemonic representations. In particular, key questions about mechanisms that occur at millisecond timescales, including how the distractor's timing, and its input gating, affect mnemonic biases are challenging to address. Because the WM memorandum and distractor occur within a few hundred milliseconds of each other, the fMRI response evoked by each is conflated with the other and, therefore, not readily disentangled.

Here, we addressed these challenges with multivariate stimulus decoding of high-density human electroencephalography (EEG) recordings. We characterized how distractors induce mnemonic biases across space and time in participants (n=24) performing a continuous response, visual WM task (Fig. 1A), involving an unpredictable, task-irrelevant distractor grating. We report novel evidence for space-specific, antagonistic biases in decoded mnemonic representations. The fine temporal resolution of EEG enabled identifying how the distractor's timing, encoding fidelity and input gating control these biases, with millisecond precision. We consolidate these findings with a two-tier ring attractor model[26,27], that provides a mechanistic, neuro-computational account of these biases.

Results

Distractors induce a space-specific bias 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 side 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 whether space-specific biases were induced by 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) 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.

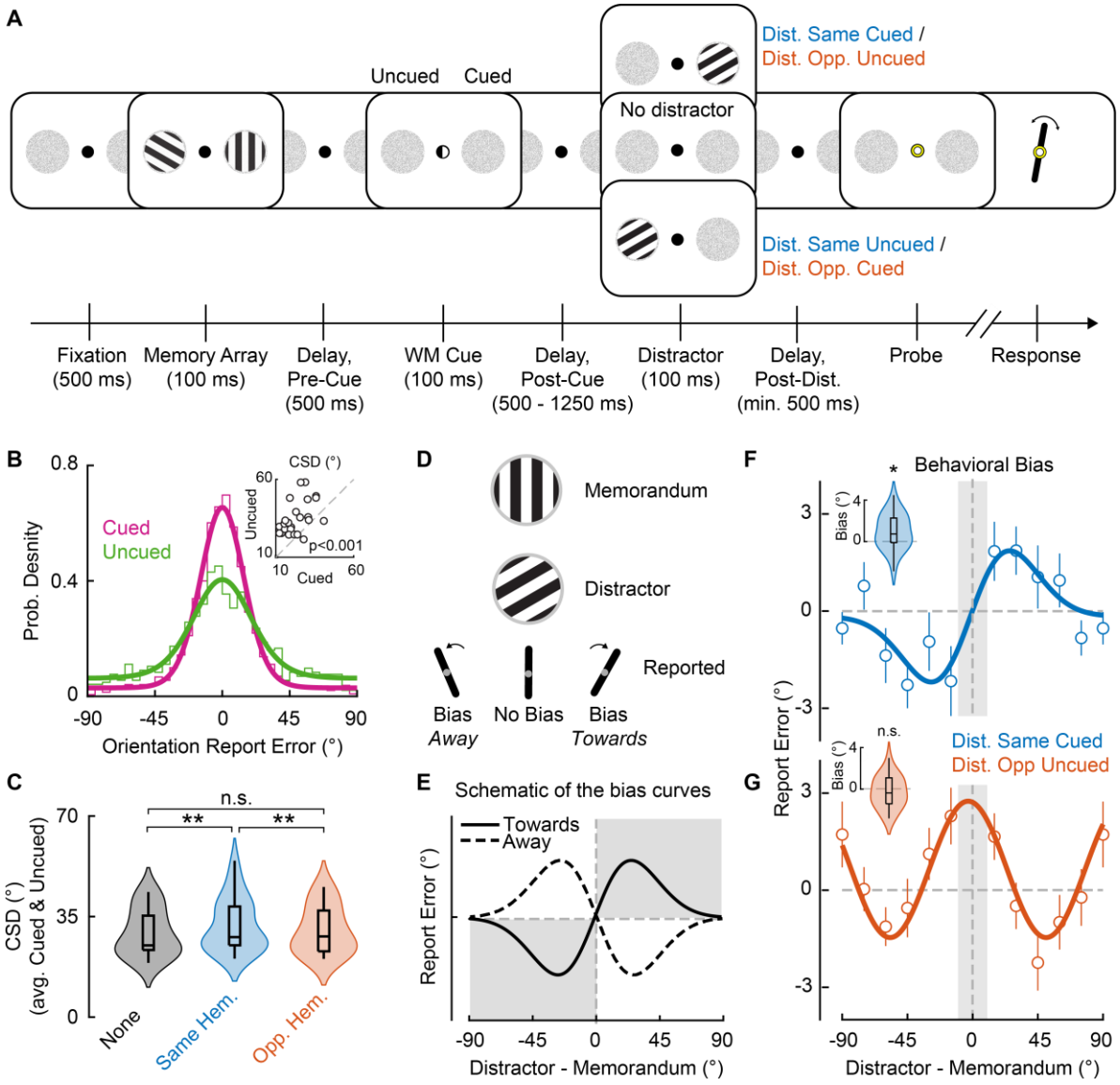


Figure 1. Distractors induce space-specific biases in visual working memory.

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 (post-cue), on 80% of the trials, a high contrast distractor grating – identical in size and spatial frequency

as the memoranda – 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 reproduced. **B.** Distribution (thin histogram) of the orientation report errors for the cued (magenta) and uncued (green) trials (n=24, data pooled). 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.** The distractor (lower grating) may bias the memorandum's (upper grating) report away from itself (left bar), towards itself (right bar), or not at all (middle bar). **E.** Schematic of bias curves. 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: biases toward, or away from, the distractor's orientation, respectively. Gray shading: strength of bias (Methods). **F.** Behavioral bias for the cued memorandum when the distractor appeared in the same hemifield. Solid line: Difference of Gaussians or sine fit (Methods). Error bars: s.e.m. Gray shaded rectangle: untested relative orientation values (±10°). (*Inset*) Violin plot: bias strengths across participants. **G.** Same as in panel F, but for trials in which the distractor appeared in the opposite hemifield as the cued memorandum. (All panels) asterisks: *: $p<0.05$; **: $p<0.01$; n.s.: not significant.

Next, we analyzed the effect of the 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), 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; data pooled across cued and uncued locations). In other words, the distractor affected the precision of orientation reports for the memorandum when it occurred, subsequently, in the same hemifield but not when it occurred in the opposite hemifield.

It is possible that the increase in orientation report error was merely a consequence of non-specific interference in WM maintenance by the distractor – a salient, flashed stimulus. Alternatively, it is possible that the increase in error arose from a specific bias in WM content induced by the distractor's orientation. To disambiguate these hypotheses, we tested whether the behavioral orientation report of the probed memorandum was systematically 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 grating were systematically biased towards the distractor's orientation (Fig. 1E) when the distractor appeared in the same hemifield as the cued grating ($Bias_{Cued,Same}$=0.9°±0.4°, p=0.018, BF=3.78). By contrast, when the distractor appeared in the opposite hemifield orientation reports were marginally biased away, albeit non-significantly ($Bias_{Cued,Opp}$=-0.3°±0.3°, p=0.346, BF=0.32) (Fig. 1F,G). Distractor-induced biases were not

statistically significant for orientation reports of the uncued grating ($Bias_{Uncued,Same}$=1.8°±1.1°, p=0.278, BF=0.60; $Bias_{Uncued,Opp}$=1.0°±1.06°, p=0.067, BF=0.314); the comparatively fewer trials (30%) in which uncued gratings were probed rendered estimates of the bias curve noisy thereby precluding a strong conclusion for this location.

In sum, the distractor induced an attractive bias for cued orientation reports only when it occurred in the same hemifield as the cued memorandum, but not when it occurred in the opposite hemifield. This space-specific bias induced by the distractor indicates that maintenance in working memory is mediated, at least in part, by a spatially lateralized, intra-hemispheric process. To pinpoint mechanisms underlying this bias, and to identify its neural origins, we analyzed neural representations of the memory array and distractor with multivariate stimulus decoding of electrophysiological recordings.

Distractors induce antagonistic biases in mnemonic neural representations

To better understand the neural basis of the bias induced by the distractor on the memoranda, we also computed the bias in the underlying neural representations. For this, first, we performed trial-wise decoding of the memorandum and distractor orientation from time-resolved EEG activity using a multivariate decoder based on the pair-wise Mahalanobis distance[29,30]. The orientation of both memory array gratings could be successfully decoded in an interval from ~50 ms to ~600 ms following their onset (Fig. 2A,B) (p<0.001, cluster-forming threshold p<0.05; n=23 participants for whom EEG data was available, see Methods), as could the distractor's orientation following its onset (Fig. 3A,B) (p<0.001, cluster-forming threshold p<0.05), thereby confirming the success of the decoding approach.

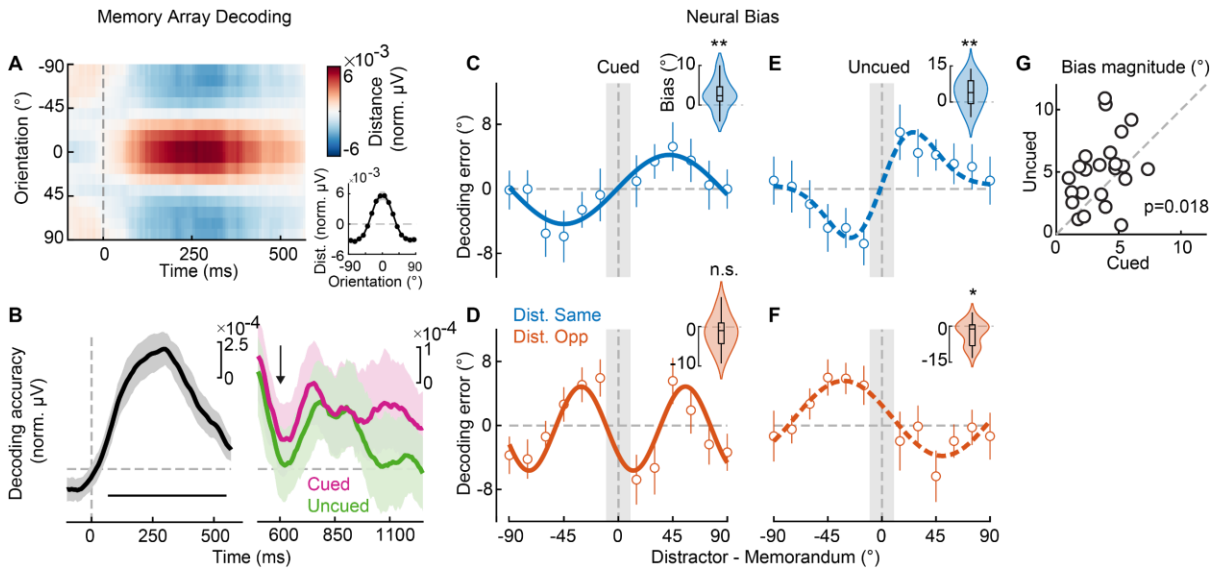


Figure 2. Distractor's orientation biases mnemonic neural representations.

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). Data averaged across participants (n=23) and over the left and the right stimuli. (*Inset*) Orientation tuning curve corresponding to the time of peak decoding accuracy. **B.** (*Left*) 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 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. (*Right*) Same as in the left panel, but showing decoding accuracy separately for cued and uncued gratings following cue onset (arrowhead). **C.** Neural bias curves, following the same format as in Figure 1E, showing the bias in the decoded neural representation of the cued memorandum induced by the distractor, for trials in which the distractor appeared in the same hemifield as the cued memorandum. Error bars: s.e.m. (*Inset*) Violin plot showing the distribution of the neural bias across participants. Other conventions are the same as in Figure 1F. **D.** Same as in panel C, but

for trials in which the distractor appeared in the opposite hemifield as the cued memorandum. **E.** Same as in panel C, but showing the neural bias in the decoded neural representation for the uncued memorandum, for trials in which the distractor appeared in the same hemifield as the uncued memorandum. **F.** Same as in panel D, but for the uncued memorandum for trials in which the distractor appeared in the opposite hemifield as the uncued memorandum. **(D-F).** Other conventions are the same as in panel C. **G.** Magnitude (absolute value) of bias strength for the cued (x-axis) versus the uncued (y-axis) memoranda. Data averaged across distractor locations. Circles: individual participants. Dashed line: line of equality.

Next, we analyzed neural biases in the decoded memory array orientation during the maintenance epoch upon the appearance of the distractor (Fig. 1A, Delay, Post-Dist.). The decoded orientation of the cued grating was systematically biased towards the distractor's orientation (attractive bias) when the latter stimulus was presented in the same hemifield as the former (Fig. 2C) ($Bias_{DistSame}$=2.7°±0.9°, p=0.007, BF=12.52) but not when it was presented in the opposite hemifield as the former (Fig. 2D) ($Bias_{DistOpp}$=-1.2±0.9°, p=0.191, BF=0.51). Similarly, the decoded orientation of the uncued grating was systematically biased towards the distractor's orientation when the latter stimulus was presented in the same hemifield as the former (Fig. 2E) ($Bias_{DistSame}$=3.9°±1.2°, p=0.006, BF=12.18). Surprisingly, the decoded orientation of the uncued grating was biased away from the distractor's orientation (repulsive bias) when the two were presented in opposite hemifields (Fig. 2F) ($Bias_{DistOpp}$=-3.0±1.0°, p=0.018, BF= 5.31). A 2-way ANOVA revealed a significant main effect of cueing on neural bias magnitude ($F_{(1,22)}$=11.06, p=0.018), and a post-hoc test revealed that the absolute magnitude of the distractor-induced neural bias, was greater for the uncued, as compared to the cued, gratings (Fig. 2G) (p=0.018, BF=2.96; Tukey's h.s.d. procedure).

We extended these results by visualizing them with a “neural tuning bias map” (Methods): a plot of the decoded orientation tuning curves for each value of the difference between the distractor and the memorandum’s grating angles. SI Figure S1 shows representative tuning bias maps for a bias toward or away from the distractor. Plotting the tuning bias maps revealed a pattern that closely matched the “attractive bias” map when the distractor appeared in the same hemifield as the remembered grating and matched the “repulsive bias” map when it appeared on the opposite hemifield. Both patterns were more visually pronounced in the uncued than in the cued hemifield, confirming the trends observed with statistical analysis.

We confirmed that this pattern of biases could not be explained merely by distractor-evoked EEG activity “leaking into” activity evoked by the memory array. Such an explanation is unlikely, in the first place, because it cannot readily account for the repulsive bias – away from the distractor’s orientation – for the memory item in the opposite hemifield. Nevertheless, to test this hypothesis, we trained an orientation decoder for each memorandum at each time point following memory array onset and tested if it could predict the distractor’s orientation at the corresponding time points. These decoders performed at chance when decoding the distractor’s orientation both when it appeared in the same or the opposite hemifield as the respective memorandum (SI Fig. S2A) ($p>0.05$, cluster-based permutation test). The results suggest that the bias in memorandum decoding that we observed was not merely due to the distractor’s orientation leaking into maintained activity of the remembered grating orientations.

In sum, the mnemonic neural representation of the cued grating, during WM maintenance, was systematically biased toward the distractor's orientation when these two stimuli were presented on the same hemifield. Neural representations of the uncued grating also showed similar patterns of an attractive bias and, in addition, also showed a strong repulsive bias away from the distractor’s orientation when the latter was presented in the cued (opposite) hemifield. In

other words, the distractor induced space-specific patterns of antagonistic (attractive and repulsive) biases during maintenance in visual WM.

Distractor timing, encoding strength and input gating control bias magnitude

To further understand the mechanistic basis of these distractor-induced biases, we investigated specific factors that might affect the strength of these biases in WM. Based on recent literature[8], we hypothesized that i) the strength of the distractor's encoding following presentation, ii) its timing during the maintenance epoch, and iii) its input gating would modulate distractor-induced bias. While the first two factors (encoding strength, timing) are easily quantified (see next), quantification of input gating is more challenging[8].

To quantify distractor input gating, we adopted the following working hypothesis: we proposed that information in the prioritized location (cued side) would be protected more strongly against distractor interference. We, then, explored neural signatures associated with the distractor that exhibited marked hemifield asymmetries between the cued and uncued sides. Specifically, we quantified and compared three distractor-related neural signatures across the two sides: i) the strength of distractor encoding, as indexed by its multivariate decoding accuracy, ii) the bias in the distractor's neural representation induced by the memoranda on each side and iii) distractor-evoked event-related potential (ERP) components, N1 and P2/P3a, which are known to index the allocation of spatial attention and working memory respectively[31–33].

The strength of distractor encoding was not different based on whether it appeared on the cued or the uncued side (Fig. 3B) (decoding accuracy: Cued=$10.76\pm2.96\times10^{-4}$, Uncued=$11.43\pm2.70\times10^{-4}$, p=0.903, BF=0.22). Nor was there a bias in the distractor's neural representation induced by the cued or the uncued gratings (Fig. 3C) (Bias: Cued=1.07±0.95°, p=0.212, BF=0.39; Uncued=-0.92±1.07°, p=0.196, BF=0.31). Similarly, the distractor-evoked

contralateral N1 ERP (Methods) was not different between the cued and the uncued hemifield (Fig. 3D) (Cued=-0.89±0.20, Uncued=-0.87±0.19, p=0.784, BF=0.23). In contrast, the distractor-evoked P2/P3a ERP was significantly higher when the distractor appeared in the cued, as compared to the uncued hemifield (Fig. 3D) (Cued=0.26±0.10, Uncued=0.18±0.11, p=0.036, BF=1.69). Consequently, we quantified P2/P3a amplitude as a direct indicator of distractor input gating.

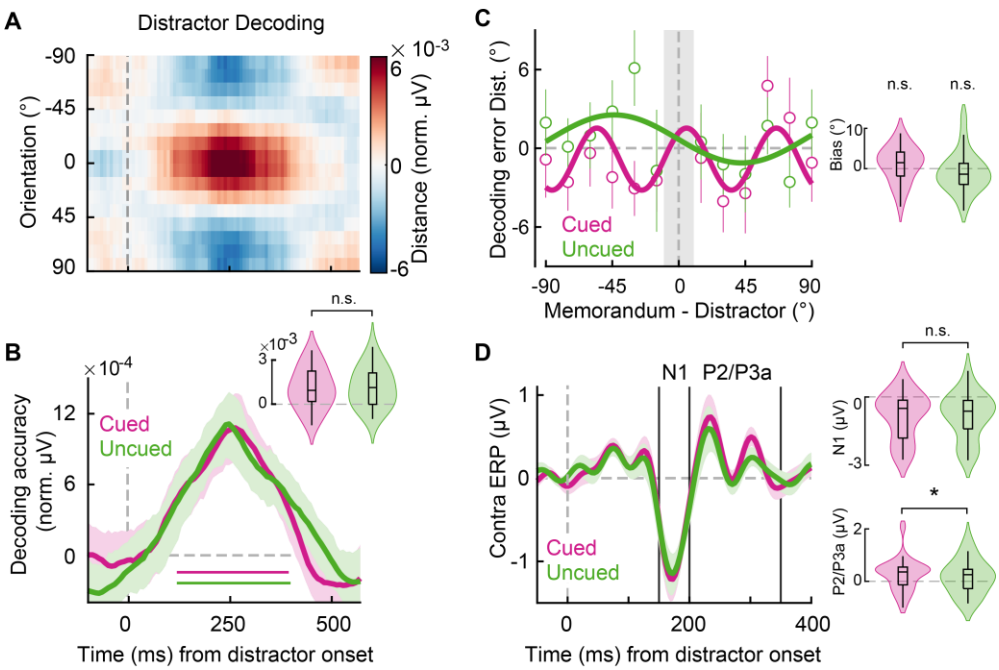


Figure 3. P2/P3a ERP provides a robust marker of distractor gating into working memory.

A. Orientation tuning map for distractor decoding for each time point relative to distractor onset. Other conventions are the same as in Figure 2A. **B.** Distractor decoding accuracy for each time point computed separately for trials in which the distractor appeared in the cued (magenta) or the uncued (green) hemifields. (*Inset*) Violin plots showing the distribution of decoding accuracy for distractor in the cued (magenta) or the uncued (green) hemifields. Shading: s.e.m. Other conventions are the same as in Figure 2B. **C.** Bias in the decoded neural representation of the distractor induced by either the cued memorandum (magenta) or the uncued memorandum

(green) for trials in which the distractor appeared in the same hemifield as the respective memorandum. (*Insets*) Violin plots showing the distribution of the neural bias in the distractor representation across participants induced by the cued (magenta) and the uncued (green) memoranda. Error bars: s.e.m. Other conventions are the same as in Figure 2C. **D.** Event-related potentials (ERP) locked to distractor onset (dashed vertical line) 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 150 ms, 200 ms and 350 ms demarcate the boundaries for computing N1 and P2/P3a amplitudes, respectively. Shading: s.e.m. (Insets) Violin plots showing the distribution of the N1 and P2/P3a amplitudes evoked by the distractor in the cued (magenta) and uncued (green) hemifields, respectively. Other conventions are the same as in panel C.

Event-related potentials (ERP) locked to distractor onset (dashed vertical line) 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 150 ms, 200 ms and 350 ms demarcate the boundaries for computing N1 and P2/P3a amplitudes, respectively. Shading: s.e.m. (*Insets*) Violin plots showing the distribution of the N1 and P2/P3a amplitudes evoked by the distractor in the cued (magenta) and uncued (green) hemifields, respectively. Other conventions are the same as in panel C.

With this definition, we probed whether the distractor's timing, encoding strength and input gating would affect the magnitude of distractor-induced bias in behavioral orientation reports. As before, due to the relative paucity of trials in which the uncued grating was probed for behavior, we performed these analyses with the cued grating probed trials only. Moreover, as a first step, we confirmed that these factors did not mutually correlate with each other (SI Results Section on *Distractor timing, strength of encoding and gating are independent factors*).

For the first analysis, we divided (median-split) trials based on the time at which the distractor appeared during the working memory epoch; we term these “early” and “late” distractor trials. 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 orientation 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>10^2) (Fig. 4A). By contrast, this bias was not significantly different from zero for early distractor trials (“early”: behavioral $Bias_{DistSame}$=-0.3°±0.8°, p=0.668, BF=0.23) (Fig. 4A).

For the second analysis, we divided trials based on the strength of the distractor’s decoded orientation quantified with the cosine similarity metric (Methods); we term these the “weak” and “strong” distractor trials. Distractor-induced orientation bias, when the distractor appeared on the same hemifield as the cued grating, was significantly positive for strong distractor decoding trials (“strong”: behavioral $Bias_{DistSame}$=1.7°±0.7°, p=0.031, BF=1.98) (Fig. 4B). By contrast, this bias was not significantly different from zero for weak distractor decoding trials (“weak”: behavioral $Bias_{DistSame}$=0.5°±0.6°, p=0.447, BF=0.31) (Fig. 4B).

For the final analysis, we divided trials based on the amplitude of the distractor-evoked contralateral P2/P3a component, for trials in which distractor appeared on the same hemifield as the cued grating. Distractor-induced orientation bias was not significantly different from zero for high P2/P3a-amplitude trials (“strong”: behavioral $Bias_{DistSame}$=0.4°±0.5°, p=0.224, BF=0.29) (Fig. 4C). By contrast, this bias was significant, and attractive, for weak P2/P3a amplitude trials (“weak”: behavioral $Bias_{DistSame}$=1.2°±0.5°, p=0.013, BF=2.53) (Fig. 4C).

In other words, distractors that appeared later during working memory maintenance, as well as those that were encoded more robustly, produced a stronger bias in orientation reports of

the cued memorandum. Moreover, the effectiveness of distractor input gating – as quantified operationally by the distractor-evoked P2/P3a amplitude – reliably predicted the magnitude of bias. We synthesized these findings, as well as the other findings described so far, into a dynamical attractor model of visual WM.

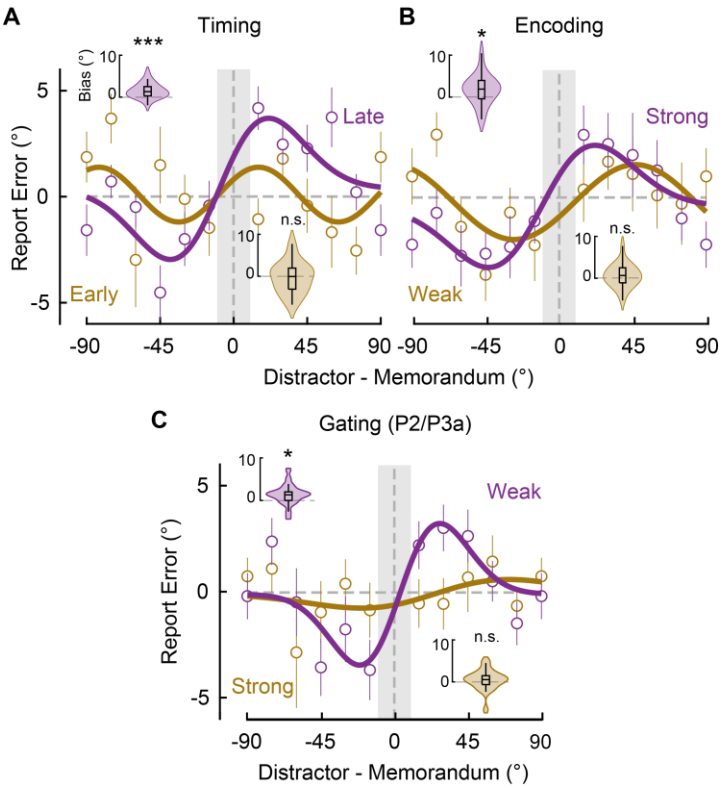


Figure 4. Distractor's timing, encoding strength and gating predict mnemonic biases.

A. Behavioral bias curves as in Figure 1F, but separately for trials in which the distractor appeared early (brown) or late (purple) during the delay period. Inset: Violin plots showing the distribution of bias strengths across participants for early (brown) and late (purple) distractor trials. Other conventions are the same as in Figure 1F. **B.** Same as in panel A, but for trials with weak (brown) or strong (purple) distractor encoding. Other conventions are the same as in panel A. **C.** Same as in panel A, but for trials with high (brown) or low (purple) distractor-evoked P2/P3a amplitude, an index of distractor gating (see text for details). Other conventions are the same as in panel A.

An attractor model explains distractor-induced biases in visual WM

To provide a parsimonious, mechanistic explanation of these distractor-induced biases, we designed a two-tier bump attractor model. Based on extensive recent literature[3–5,9,13,18,34–37], working memory maintenance in our model involves both the sensory (visual) cortex (VC), and higher cortex (HC), such as the prefrontal cortex (PFC) or the parietal cortex (PPC). In fact, the latter regions both are reported to be involved in robust maintenance in the presence of distractors[9,23,24,38]. Moreover, we include a key element of novelty in our model: while distinct attractors independently encode stimuli in the two hemifields, they interact with each other via topographic, cross-hemifield inhibitory connections. We provide a brief description of the model below; a detailed description is provided in the Methods.

The VC and the HC were modeled with separate ring attractors; we also employed distinct attractors for encoding the memorandum in each (cued, uncued) hemifield, yielding a total of 4 attractor networks (Fig. 5A, top). Each attractor network comprised both excitatory (E) and inhibitory (I) neurons (512 each) with dynamics described by rate-based models (Methods, equations 1-2). Recurrent connections among E neurons within an attractor were local and topographic (Fig. 5A, bottom), whereas all other connection types (E-I, I-E, I-I) were global (see Methods). VC neurons provided input to HC neurons via topographic, excitatory feedforward connections and received input from HC via reciprocal, topographic feedback connections (Methods, equations 3-4). Based on our experimental observation of a lower distractor-induced neural bias for the cued, as compared to the uncued representation (Fig. 2G), we expected cued information in the VC to be more robust to distractors. Therefore, HC→VC feedback connections were made 20% stronger for the cued, than the uncued, hemifield (Fig. 5A, thick HC-VC connection). In addition, previous studies suggest that mnemonic information in higher order areas (e.g., PFC, PPC) is relatively robust to distractors[8,9,13,23,38]. Consequently, we transiently

disconnected the feedforward connections from the VC to the HC during the maintenance epoch, thereby rendering the HC mnemonic information immune to disruption by the distractor upon its presentation. Finally, based on previous reports of competitive visual inhibition across the hemifield[39–43], we also modeled topographic, cross-hemispheric inhibitory connections between the VC attractors in each hemisphere (Fig. 5A, black circles, Fig. 5B, top; equation 3); these connections played a critical role in explaining several, critical experimental observations (see next).

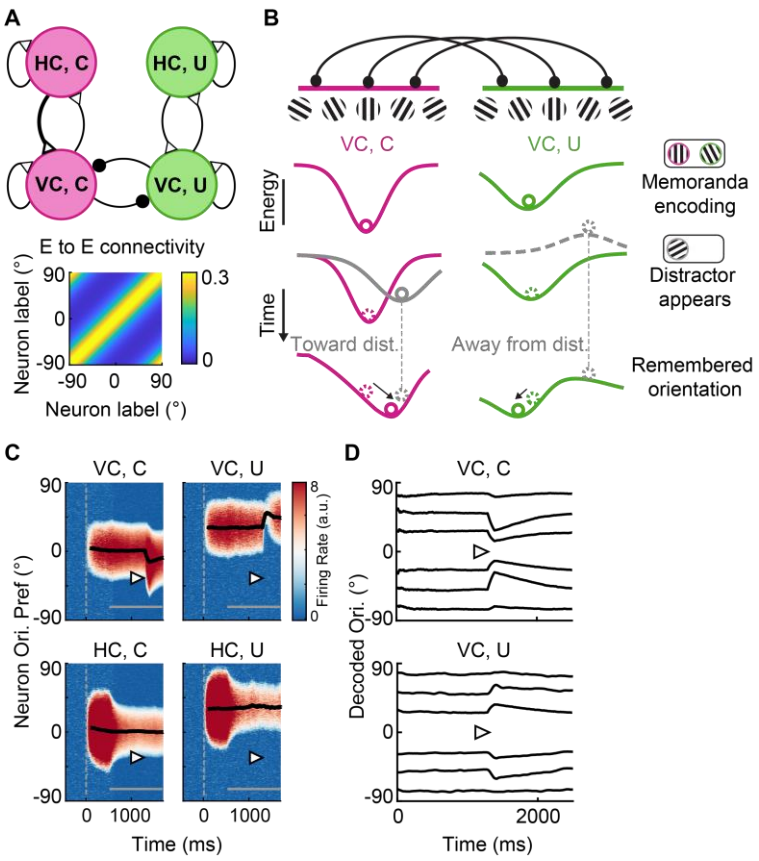


Figure 5. Two-tier attractor model explains spatially antagonistic, mnemonic biases.

A. (*Top*) Two-tier ring attractor model. The visual cortex (VC, bottom) and higher cortex (HC, top) were modeled with separate ring attractors. Distinct attractors also encoded stimuli in the Cued (left, Magenta) and the Uncued (right, Green) hemifields. Top-down input from HC to VC was stronger for the cued (thick black line) compared to the uncued side. Mutual inhibition between

the VC,C and VC,U attractors was mediated by topographic, cross-hemispheric connections (black circles). (*Bottom*) Recurrent E-E connection weights, showing local and topographic connectivity (see text for details). **B.** Energy landscape diagram explains antagonistic distractor-induced biases. *(Topmost)* Mutual inhibition between VC,C and VC,U. (*Top*) The cued memorandum (magenta circle) is maintained more robustly, at a lower energy state, than the uncued memorandum (green circle). (*Middle*) In this example, the distractor (gray, solid) appears in the cued hemifield and causes a valley (energy minimum) in the cued attractor landscape. Topographic, cross-hemifield inhibition induces a peak (maximum) at the spatially corresponding location in the uncued landscape (gray, dashed). (*Bottom*) The final attractor landscape carries a bias in the cued representation (magenta circle) *towards* the distractor's orientation and a bias in the uncued representation (green circle) *away* from the distractor's orientation. **C.** Firing rate heat maps of the VC,C (top, left), VC,U (top, right), HC,C (bottom, left) and HC,U (bottom, right) attractors; hotter colors indicate higher activity. Y-axis: E neurons arranged in order of orientation preference on the ring; x-axis: Time. In this example, the onset (dashed vertical line) of the memorandum gratings (cued: 0°; uncued 30°) causes a localized bump in firing rate at the respective location on the ring, which persists throughout the delay period (gray horizontal bar) for all attractors. Thick black curve: orientation decoded from the population activity vector for each attractor. Black unfilled triangle: Distractor grating (-30° orientation) presented on the cued hemifield during the delay period. **D.** Decoded orientation from the VC,C (top panel) and VC,U (bottom panel) attractors across time for six different memorandum orientations; each curve represents a distinct simulation. Black unfilled triangle: Distractor grating (0° orientation) presented in the cued hemifield.

With this model, we sought to replicate the attractive and repulsive biases observed experimentally. We explain the expected sequence of events, first, with a schematic (Fig. 5B);

detailed simulations are presented subsequently. The presentation of either memorandum produces a persistent dip in the VC attractor's energy landscape at the orientation of the respective stimulus, either cued (Fig. 5B, left column, magenta) or uncued (Fig. 5B, right column, green). Next, distractor presentation – in this case, in the same hemifield as the cued memorandum – causes another dip in the VC,C attractor's energy landscape at the distractor's orientation during maintenance (Fig. 5B, left column, middle row, gray trace). This dip interacts with the maintained representation of the cued memorandum and causes it to drift towards the distractor's orientation (Fig. 5B, left column, bottom row, magenta trace), yielding an attractive bias towards the distractor (Fig. 5B, left column, bottom row, magenta circle). By contrast, the distractor also creates a peak at its own orientation in the VC,U attractor's energy landscape in the opposite hemifield; this energy peak arises as a result of the topographic, cross-hemispheric inhibitory connections (Fig. 5B, right column, middle row, gray trace). This peak causes the maintained representation of the VC,U stimulus to drift away from the distractor's orientation (Fig. 5B, right column, bottom row, green trace), yielding a repulsive bias (Fig. 5B, right column, bottom row, green circle).

Next, we describe the results of simulating the model. We modeled the initial memory array presentation with an additive input current to the VC-E neurons of each attractor (cued, uncued) using a von Mises spatial profile centered on the orientation of the respective grating (κ_M=2.7°, equation 5). This initiated localized bumps of activity in VC attractors, at locations corresponding to the orientations of the encoded gratings (Fig. 5C, top rows; activity at cued: 0°; uncued: 30°); similar activity was also initiated in the corresponding HC attractors (Fig. 5C, bottom rows). Upon disappearance of the memory array, during the delay period (Fig. 5C, gray underbar), the bump of activity was maintained through recurrent excitation in each attractor, as in conventional bump attractor models[26,27,44] (Fig. 5C, red patches). A distractor was presented at t=1300 ms during the delay period (Fig. 5C, unfilled triangle); distractor presentation was

modeled, also with a von Mises input profile (κ_D=2.3°, equation 6) centered on the distractor's orientation (Fig. 5C; distractor: -30°).

Neural estimates of each memorandum's orientation were decoded directly from the location of peak activity in the orientation map in the VC, in direct analogy to decoding orientation from the occipital EEG electrodes (e.g., Figs. 5C-D, black traces). By contrast, behavioral estimates were obtained by averaging the memorandum's orientation decoded from both VC and HC at the end of the trial; this choice rendered behavioral estimates more robust to noise (see SI Results Section for a detailed analysis of error-correcting dynamics). While these simulations were performed with a fixed value for each parameter (SI Table S1), the reported effects were fairly robust to minor changes in these parameter values.

The simulations replicated five key experimental observations in the data:

i) <u>The distractor induces an attractive mnemonic bias in its hemifield:</u> We observed that the distractor produced an attractive bias for the cued memorandum in the VC attractor when the former was presented in the latter's hemifield. Figure 5C (top left) shows the firing rate heat map for an exemplar simulation in which the distractor (Fig. 5C, top left, black unfilled triangle) biased the decoded VC,C representation (Fig. 5C, top left, black trace overlaid on heat map) towards itself. We repeated these simulations with six different cued memoranda of different orientations (Fig. 5D, top), with a distractor always appearing at 0° orientation (Fig. 5D, top, unfilled triangle). In each case, the distractor biased the decoded neural representations towards itself, to varying degrees (see next).

ii) <u>The distractor induces a repulsive mnemonic bias in the opposite hemifield:</u> In addition, the distractor also produced a repulsive mnemonic bias in the VC attractor in the opposite hemifield. Figure 5C (top right) shows the firing rate heat map for an exemplar simulation in which the distractor (Fig. 5C, top right, black unfilled triangle) appeared in the cued memorandum's hemifield and biased the uncued (VC,U) representation (Fig. 5C, top right, black trace) away from itself. As before, repeating the simulations with six different uncued memoranda, of different

orientations (Fig. 5D, bottom), revealed systematic patterns of bias away from the distractor's orientation for the uncued memorandum.

We quantified the bias for each of these distractor conditions. In line with the experimental findings, the distractor produced an attractive mnemonic bias for the cued memorandum in its hemifield (Fig. 6A) and a repulsive mnemonic bias in the opposite hemifield (Fig. 6B). Importantly, the magnitude of this bias varied parametrically with the angular difference between the distractor and the memoranda: the strongest bias appearing for intermediate angle differences, and the weakest appeared for extreme differences (0° or ±90°). Similar trends were observed for the uncued memorandum also (Figs. 6C-D). Behaviorally, the distractor produced a strong attractive behavioral bias in its hemifield but only a weak repulsive behavioral bias in its opposite hemifield (Fig. 6E-F; both for the cued memorandum). This disparity occurred because the inhibitory effect decayed relatively quickly, influenced by error-correcting dynamics arising from the HC attractor (see next).

iii) <u>Late distractors produce stronger biases:</u> Following the appearance of the distractor, VC neural representations were either biased toward (same hemifield) or away from (opposite hemifield) the distractor's orientation. Yet, the top-down feedback from HC to the VC mitigated distractor interference in the VC through error-correcting dynamics[8,9,24] (SI Results Section on *Error correcting dynamics*). Specifically, late distractors appeared later during the WM maintenance epoch and were temporally more proximal to the response probe than early distractors; consequently, the former received ameliorative top-down feedback for a shorter duration than the latter. This produced a higher "residual" bias for late (Fig. 6G, purple) compared to early (Fig. 6G, brown) distractor trials at the end of the delay epoch, matching our experimental findings.

iv) <u>Strongly encoded distractors produce stronger biases:</u> As mentioned above, the distractor induced a dip in the VC attractor's energy landscape. We varied the extent and spread of this dip to mimic differences in the strength of the distractor encoding – strongly encoded

distractors were afforded a 60% higher input current amplitude as compared to weakly encoded distractors. The stronger perturbation of the energy landscape induced by the strongly encoded distractors yielded faster drifts in the memorandum's representation thereby yielding a higher bias for weak (Fig. 6H, purple) compared to strong (Fig. 6H, brown) distractor encoding trials; again, these results matched our experimental observations.

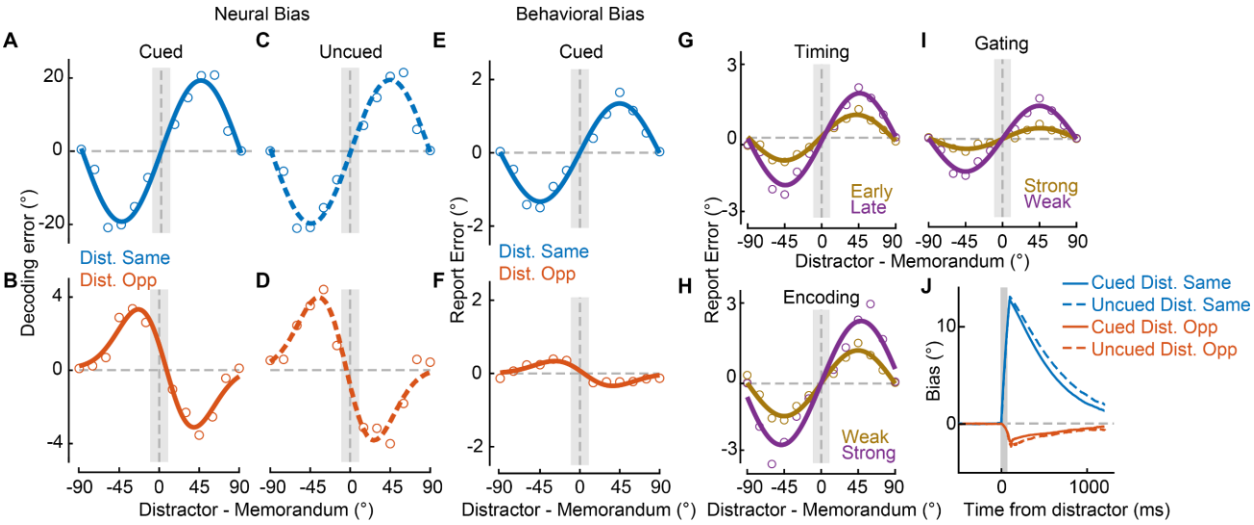


Figure 6. Model simulations replicate distractor biases in working memory.

A-B. Same as in Figures 2C-D, but showing the model's neural bias for the cued representation (VC,C attractor) measured 100 ms following the distractor onset when the distractor appeared in the same (C) or opposite (D) hemifield as the cued memorandum. Other conventions are the same as in Figures 2C-D. **C-D.** Same as in Figures 2E-F, but showing the model's neural bias for the uncued representation (VC,U attractor) measured 100 ms following the distractor onset when the distractor appeared in the same (E) or opposite (F) hemifield as the uncued memorandum. Other conventions are the same as in Figures 2E-F. **E-F.** Same as in Figures 1F-G, but showing the model's behavioral bias for cued grating reports when the distractor appeared in the same (A) or opposite (B) hemifield as the cued grating. Other conventions are the same as in Figures 1F-G. **G-I.** Same as in Figures 4A-C, but showing the effect of distractor's timing (G), encoding strength (H) and input gating (I) on the model's behavioral bias. Other conventions are the same

as in Figures 4A-C. **J.** Model's neural bias as a function of time from distractor onset separately for the cued (VC, C; solid line) and uncued (VC, U; dashed line) attractors when the distractor was presented in the same hemifield (blue) or opposite hemifield (red) as the corresponding memorandum. The bias decay is more rapid for the cued representation (solid lines) due to stronger top-down, error correcting dynamics, as compared to the uncued representation (dashed lines). Gray shaded region: period of distractor presentation. All data averaged across n=10,000 simulations of the model, each initialized with a different random seed.

v) Weakly gated distractors produce stronger biases: Our experimental observations suggest a protective effect of the distractor-evoked P2/P3a potential: both in terms of a stronger P2/P3a in the cued hemifield (Fig. 3D), and a stronger P2/P3a predicting a weaker distractor-induced bias (Fig. 4C) (see Discussion). We propose that such a protective effect is mediated by feedback input from the HC to VC, which enabled distractor gating and rendered mnemonic representations in the VC more robust to interference. To test this hypothesis, we varied the strength of this top-down feedback by a factor of 2x – the magnitude of distractor-induced bias was systematically higher for the weakly-gated distractors (Fig. 6I, purple), compared to the strongly gated distractors (Fig. 6I, brown), consistent with experimental results (see also SI Fig. S3A). Interestingly, the effect of this gating mechanism was also apparent in our model, in another way. Due to stronger top-down feedback from HC to VC attractor on the cued hemifield, error-correcting dynamics were stronger in the cued hemifield. This led to a weaker neural bias for the cued (Fig. 6J, solid curves) compared to the uncued (Fig. 6J, dashed curves) memorandum in the model's neural representation.

In summary, our results uncover space-specific, antagonistic patterns of distractor-induced biases in visual working memory. Behavioral reports, and the mnemonic neural representation of the memorandum, were biased towards the distractor's orientation when the

latter appeared in the same hemifield as the former. By contrast, the mnemonic neural representation was biased away from the distractor's orientation when the latter appeared in the opposite hemifield. Our two-tier ring attractor model parsimoniously replicated each of these experimental observations. Moreover, the model comprehensively explained diverse factors that mediate distractor-induced bias, including distractor timing, strength of encoding and input gating. Overall, the results identify essential neural bases of space-specific distractor biases in visual working memory.

Discussion

We discovered that a temporally unpredictable, task-irrelevant distractor produces space-specific, antagonistic biases in visual working memory. The memorandum's neural representation was biased towards the distractor's orientation when the latter occurred in the same hemifield as the former, but was biased away when these stimuli appeared in opposite hemifields. Moreover, the internally selected cued neural representation was less susceptible to distractor-induced biases, as compared to the uncued representation. These results are consistent with the hypothesis that information about the memorandum is maintained, at least partially, in a lateralized buffer[22,45–48]. Over time, these representations become more labile and prone to interference by the sensory distractor. Yet, information associated with the selected (retro-cued) hemifield is held more robustly and shielded against distractor interference, possibly by "error-correcting" top-down control mechanisms[8,9,13,19,24].

We have ample evidence to believe that the distractor biased the information maintained in WM (mnemonic bias), rather than merely inducing a bias during the behavioral report (response bias). First, the distractor was flashed briefly (100 ms) and was followed immediately by a noise mask (pedestal) to diminish persistent iconic representations. Moreover, the distractor appeared 960 ms, on average, and at least 500 ms, before the probe rendering biases arising from iconic memory unlikely. Second, had the distractor merely produced a response bias, we would have expected such a bias to occur regardless of the distractor's location. In our task, even though the response probe and bar were always presented centrally, we observed distinct patterns of biases – either toward the distractor's orientation, or a null bias (Fig. 1F-G, 2C-F) – depending on the distractor's location relative to the probed memorandum. Third, and most importantly, decoding the neural representation of the memorandum during the maintenance epoch revealed distinct, space-specific, antagonistic patterns of distractor-induced biases, providing direct evidence for distractor induced biases in the representations maintained in WM.

Critically, prioritization in working memory promoted distractor resilience: the selected neural representation, associated with the retro-cued memorandum, was significantly less biased by the distractor as compared to the uncued memorandum's representation. Consistent with previous studies[24,49], this result indicates that prioritized information is more robust to interference, perhaps due to greater top-down feedback stabilizing the cued memorandum's representation[24,49]; indeed, this is the mechanism simulated in our computational model (Fig. 5A, top). On the other hand, it is also possible that such stabilization of the cued representation arises from a mixture of persistent activity and activity-silent states[7,29,44,50]; the latter are arguably less prone to interference from sensory distractors[7]. Future experiments can disambiguate these hypotheses by using a "pinging" protocol that activates latent representations and studying the effect of distractor interference on these activated representations.

Our results provide concurrent validity to the sensory recruitment hypothesis of visual WM, which posits that brain regions involved in processing the memorandum during encoding also contribute to its maintenance in WM[1–5]. We were able to successfully decode both the orientation of the memoranda during the delay epoch, as well as distractor-induced biases in their neural representations from the same posterior (occipital) EEG electrodes (Fig. 2A-B). Moreover, distractor-induced biases decoded from these posterior electrodes were closely tracked by behavioral reports (Fig. 4B). Yet, due to the low spatial resolution of EEG we cannot readily disambiguate whether the neural signatures that we observed originated entirely from the early visual cortex or also included contributions from the parietal cortex[51]. Lastly, distractor-induced bias effects were hemifield-specific; this suggests that mnemonic information may be stored in a lateralized sensory buffer, which is selectively affected by the hemifield in which the distractor appears.

Among the many electrophysiological markers tested, we found the P2/P3a ERP component to be a reliable marker of distractor input gating into working memory. Previous

literature has often linked the P2 component to the gating (or suppression) of distractor information during working memory tasks[52–54]. For instance, higher distractor-evoked P2 amplitude, indicative of greater distractor gating, correlated positively with performance in an auditory WM task[31]. Similarly, the short peak latency P3 component, referred to as the P3a, possibly originates in the frontal lobe[53] and indexes top-down inhibition of distractors[52,53]. In fact, distractor-evoked P3a amplitudes were correlated with frontal lobe grey matter volume across individuals[55], suggesting its involvement in the distractor-related processing. Our study extends these findings and shows that the P2/P3a amplitude contralateral to the distractor was higher at the cued compared to the uncued hemifield; in other words, the P2/P3a may index a stronger input gating of cued mnemonic information stored in a lateralized mnemonic buffer (Fig. 3D). Moreover, the P2/P3a amplitude directly predicted the magnitude of the distractor-induced bias, providing further confirmation for this “gating hypothesis”.

We synthesized our experimental findings with a two-tier ring attractor WM model. Attractor models have been frequently used for explaining diverse phenomena in WM, including variable precision[26], serial biases[44], categorical biases[57] and the like[27]. For instance,–a recent study employed an attractor model to explain serial biases in spatial WM[44]. Other studies employing attractor models have been able to explain the mechanisms behind trial to trial variability and the presence of categorical biases in visual WM[26,27,44]. In our study, such an attractor model was a natural candidate to explain our results: even without elaborate simulations, attractors provide a conceptually simple and intuitive framework for explaining the antagonistic pattern of distractor-induced biases across the hemifields (Fig. 2C-F). Moreover, recent research has shown that maintenance in visual working memory (WM) involves multiple regions across the brain, including both sensory and higher-order cortex[18,35,37]. We, therefore, employed a two-tier attractor model that includes both sensory and higher order brain regions to simulate our experimental findings.

In our model, a higher-order cortical region (HC) stored information relevant to the memorandum and provided top-down influence to the visual cortex (VC) during the delay period to mitigate distractor interference in WM. Candidate regions for the HC include parietal cortex or prefrontal cortex both of which are known to contain robust WM representations, especially in the presence of distractors[9,13,23,24,38]. In our model, the representation in HC was rendered immune to distractor interference by transiently disconnecting its bottom-up input from VC during the delay period. Neurophysiological mechanisms that could mediate such dynamic disconnections include oscillatory coherence[58] or pharmacological gating (e.g., acetylcholine[59]). Some recent studies have shown that disruption in visual WM maintenance occurs in both lower (V1, V4) and higher level cortex[11]. In this case, the coordinated activity across these regions could help improve the robustness of maintenance in WM, even in the presence of distractors[18,35–37,60]. Consistent with these results, our simulations showed that combining VC and HC activity yielded a more accurate behavioral readout of the memorandum's orientation, as compared to either VC or HC activity alone (SI Fig. S3B). Future work, employing fMRI, perhaps in combination with EEG, could help identify the precise brain regions, and their dynamics, that mediate distractor biases in WM.

The model accurately reproduced the effect of distractor timing on mnemonic bias: distractors presented later during the delay epoch induced a stronger bias, and vice versa. This experimental result is in line with literature suggesting that neural WM representations are not static but evolve over time, as dynamic processes[8,30,56]. We leveraged these findings in our model to explain distractor timing effects. In our model, the bias due to error-correcting dynamics in visual WM decays steadily; these dynamics are, in turn, mediated by top-down feedback from higher cortex (e.g., PFC[9,13,23,24,38]). Due to this effect, early distractor trials provided more time for the error-correcting mechanism to operate thereby yielding a weaker behavioral bias, as compared to late distractor trials. While model simulations accurately recapitulated experimental data, future experiments can confirm this hypothesis by causally manipulating the activity of

higher order brain regions (through brain stimulation) during a visual WM task and studying its effects on the extent of error-correcting dynamics observed in lower order brain regions and in behavior.

In conclusion, we identified key neural processes by which distractors induce space-specific biases in neural representations in WM. Complementing these results, our simulations suggest key computational mechanisms that may mediate distractor effects in WM. The causal role of these neural processes in distractor interference can be established with non-invasive brain stimulation methods (e.g., transcranial magnetic stimulation) by focally perturbing activity in specific brain regions (e.g., visual cortex, prefrontal cortex), and at specific times during the delay epoch. Understanding the neural basis of distractor interference in WM could have critical implications for improving WM robustness in individuals with attention and memory deficits, such as patients with attention deficit disorder (ADHD) or mild cognitive impairment (MCI[61]).

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, 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. No power analysis was performed for this study. The sample size in this study (n=24) is similar to that of multiple 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. The 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, 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 side 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.

Training and testing. 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

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

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

Eye-tracking and gaze-based trial exclusion. For all participants, gaze fixation position was monitored binocularly at 500 Hz, and stored offline for analysis. In *post hoc* analyses, any trial in which the subject’s gaze deviated by more than 1 dva from the fixation dot along the azimuthal direction was flagged, and excluded from further analysis. Given that we were primarily interested in the effect of distractors on visual working memory, this criterion for gaze-based exclusion was applied only in a window around (100 ms before to 400 ms after) distractor onset during the delay epoch. We did not reject trials based on gaze deviation during other times in the delay epoch, because, barring the fixation dot and placeholders, there were no other stimuli on screen during this epoch. We also did not reject trials based on gaze deviation during the memory array presentation, because there was no prior information (or pre-cue) that could have systematically biased gaze toward either grating. In addition, no gaze-based rejection was performed for trials in which the distractor did not appear. Finally, for 6/24 participants gaze fixation could not be reliably recorded, largely because these participants wore eyeglasses with strongly reflective 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[62–64], without the need for fitting complex psychophysical models[65].

We computed the effect of the distractor on the precision of each participant's orientation report using the circular standard deviation (CSD)[64], 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 side, versus the opposite side, of the probed stimulus; this analysis was performed separately for the cued and uncued grating reports.

We also computed the effect of the distractor in biasing orientation reports for the probed grating, following a standard protocol[28,63]. 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). For visualization purposes, the bias curves were fit with both a Gaussian derivative function[63] and a sinusoid; for each bias curve, the fit with the lower mean squared error is shown in the corresponding plot.

EEG data acquisition and preprocessing

EEG acquisition. High density EEG data was acquired with the Active Two recording system (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 technical error. EEG analyses were performed with the remaining 23/24 participants.

EEG Preprocessing. EEG data was preprocessed in MATLAB using the EEGLAB[66], ERPLAB[67] 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 >4x 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 (Rolston et al., 2009). 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[68], 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[69]. Briefly, SCADS identifies outlier signal values based on statistical measures like the standard deviation, maximum amplitude and the like. Any trial in which any of the posterior electrodes used for decoding was flagged with SCADS (outlier criterion, l=5.0) was removed entirely from the analysis; this yielded a median trial removal rate of 8.5% per participant.

EEG data analyses

<u>Decoding stimulus orientation.</u> We decoded the orientation of memory array and distractor gratings based on multivariate pattern analysis of EEG responses. Whereas the memory array grating orientations were decoded both during memory array presentation (Fig. 1A) and the delay epoch (Fig. 1A), distractor grating's orientation was decoded only during its presentation epoch

(Fig. 1A). Decoding analysis was performed using the 39 posterior electrodes over occipital cortex; either all of these 39 electrodes or a subset of the 14 electrodes in the hemisphere contralateral to the decoded stimulus, were used for the decoding analysis (SI Fig. S2B; see next). We employed a multivariate decoder based on the Mahalanobis-distance, following a procedure established by previous studies[12,29,30]. Briefly, EEG data were downsampled to 128 Hz by taking an average of each set of two successive samples (original sampling rate: 256 Hz). To improve the decoding accuracy, we used a sliding window (size: 300 ms, shift: 8 ms) and concatenated all the data points in that time window from all electrodes into a single 532 or 1482 dimensional time x channel vector (38 time points in each 300 ms window x 14, or 39, electrodes). We utilized the pattern of temporal variability of the EEG response within each time window for decoding[30]. Therefore, we subtracted the mean activity across the time window for each channel, and applied principal component analysis (PCA) to reduce the number of feature dimensions; the top-k PCA dimensions that explained 95% of the variance across trials were retained for decoding ("encoding vector").

For training and evaluation, we employed a leave-one-trial-out cross-validation approach. For each participant, the model was trained on all but one trial, with orientations binned relative to the orientation in the left-out trial into one of 16 different bins (bin centers: 11.25° apart, bin width: ±30°). Then the Mahalanobis distance was computed between the test trial's encoding vector and the mean encoding vector of each training bin. The computed distances were de-meaned and the sign of the distances was reversed for ease of visualization. To obtain a smooth decoding estimate, this procedure was repeated 8 times by sliding the bin centers by 1.4° each time, and the resultant decoding profiles were averaged across repetitions (e.g., Fig. 2A, 3A). To estimate decoding accuracy, the sign-reversed distances were multiplied elementwise by a cosine kernel, symmetric around 0° (period: 180°), and averaged. This procedure was performed for each participant, memory array location (left/right), distractor location (left/right), and each time

window, separately. The memory array gratings' orientation was decoded using the 14 electrodes contralateral to the respective grating whereas the distractor's orientation was decoded using all 39 electrodes (SI Fig. S2B).

Estimating distractor-induced bias in neural representations. To compute the distractor-induced bias in the neural representation of the memory array gratings, we first decoded the orientation of the latter gratings. For this, first, we smoothed the decoded tuning curves using a moving average window (size: 80 ms) from 150 ms before to 450 ms after distractor onset. The mean orientation of the bin corresponding to the peak of the tuning curve was taken to be the decoded orientation for that timepoint and trial. Following this, the neural bias was computed from the decoded orientation in a way similar to the behavioral bias (see Methods section on *Estimating precision and bias in behavior*). Briefly, we plotted the median error in memorandum's orientation decoding (decoded – actual orientation) as a function of the angle difference between the distractor and the grating's actual orientation (distractor – actual orientation). Similar, to the behavioral bias, 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 grating's actual orientation. We repeated this procedure separately for the distractor-same (decoded grating and the distractor appeared in same hemifield) and distractor-opposite (decoded grating and the distractor appeared in opposite hemifields) conditions, and separately for the cued and uncued gratings. The maximum value of the bias in the entire time window (-0.15 to 0.45s relative to distractor onset) was taken as the neural bias (Fig. 2C-F).

For further clarity, rather than visualizing only the decoded orientation (maximum of the decoded orientation tuning curve) we also visualized the entire tuning curve as a function of the difference between the distractor and the probed orientation; this representation was computed at the time when the bias reached its maximum value. In this 2-D plot, the color indicates the

decoding similarity (inverse of the Mahalanobis distance to the orientation along the y-axis) as a function of the difference between the distractor and probed orientation (SI Fig. S1). Visualizing the entire tuning curve demonstrated robust biases, both toward and away from the distractor orientation in the neural representations of the cued and uncued memoranda.

Distractor-evoked ERP estimation. The event related potential (ERP) was measured by signal-averaging across the same set of 14 contralateral (or ipsilateral) electrodes as employed in the decoding analyses. First, the data were high-pass filtered with a lower cut-off at 0.5 Hz using an FIR filter (N=500). Subsequently, the data were epoched from 800 ms before to 700 ms after distractor onset and then averaged across channels and trials to obtain the ERP waveform. This was done separately for each of the four distractor conditions (cued/uncued x left/right hemifield). Consistent with previous studies[31,33], the N1 amplitude was defined as the mean signal value from 150 to 200 ms post distractor onset and P2/P3a amplitude as the mean signal value from 200 to 350 ms post distractor onset; virtually identical results were obtained when the P2 and the P3a amplitudes were analyzed in separate time windows.

Median-split analyses based on distractor encoding strength, timing and gating. We computed a measure of association between the strength of neural distractor encoding and behavioral bias. For this analysis, the strength of distractor encoding was quantified as the value of decoded cosine similarity at t=250 ms after distractor onset (time of peak distractor decoding, Fig. 3B). Then, we divided trials into “weak” and “strong” distractor encoding trials based on a median split of the cosine similarity metric; the median split was performed for each participant and for each distractor condition (cued/uncued x left/right hemifield) separately. Weak and strong distractor encoding trials were then aggregated across the conditions, and the behavioral bias in the cued

orientation report was computed separately for the weak and strong trials with a procedure identical to that described previously (see Methods section on *Estimating precision and bias in behavior*). The association between other metrics like the distractor timing or the strength of gating and behavioral bias was computed in an identical manner, except that the median split was performed based on the respective metric – time of distractor onset relative to the cue (early versus late) or amplitude of the distractor-evoked P2/P3a component over the electrodes contralateral to the distractor (weak versus strong gating).

Computational model of distractor biases in WM

<u>i) Network architecture.</u> We synthesized our behavioral and neural findings by modifying and extending a bump attractor model of working memory[12]. The visual cortex (VC) and higher cortex (HC), for each of the cued and uncued visual hemifields, were modeled with distinct ring attractors (total of 4 ring attractor models). Each attractor model comprised 1024 neurons in total out of which 512 were excitatory (E) and 512 were inhibitory (I). Each neuron was described by a standard rate model and represented stimulus orientation topographically, depending on its location along the ring (Fig. 5A; 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[26,27].

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

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

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[26]

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

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 side. 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).

Delay: 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 to subsequent sensory input. 600 ms after the memory array onset, the cue appeared. To mimic prioritization of cued information in WM[24], we reduced the noise terms for both the cued attractors ($\sigma_E^{VC,C}, \sigma_I^{VC,C}, \sigma_E^{HC,C}$ and $\sigma_I^{HC,C}$ in equations 1 and 2) by 60% and increased top-down feedback to cued VC attractor ($W_E^{HC,C\rightarrow VC,C}$ in equation 4) by 20%, following cue onset.

Distractor presentation: The distractor was presented for a duration of 0.1 seconds, at a random time (distributed exponentially) between 1200 ms and 1950 ms after the memory array onset, mimicking the behavioral paradigm. The distractor was presented either in the cued or the uncued hemifield with equal probability and was encoded by the visual cortex (VC) attractor representing

the corresponding visual hemifield. Distractor presentation was modeled by adjusting the spatial profile of the input current to the excitatory neurons of the VC attractor as follows:

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

Where $U_i^{VC,C}$ is the input current to the excitatory neuron i of VC,C attractor, A_D is the amplitude of the input current, κ_D is the concentration parameter determining the spread, θ_i is the orientation preference of excitatory neuron i and Θ_D is the distractor's orientation. In addition, we enhanced the cross-hemifield inhibitory weights ($W_I^{VC,U\rightarrow VC,C}$ or $W_I^{VC,C\rightarrow VC,U}$ in equation 3) multiplicatively (by ~30x) during distractor presentation; this enabled modeling the bias away from the distractor's orientation observed in the distractor-opposite condition (Fig. 2F).

<u>iii) Quantifying simulated behavioral and neural metrics.</u> We ran a total of 10,000 simulations with 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:

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

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[3–5,9,13,18,34], 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.

The model's neural bias was computed with the decoded orientation from the VC attractor's activity around the distractor onset, using a procedure identical to that described in EEG data analysis (see section on *Estimating distractor-induced bias in neural representations*). Orientation was decoded from the VC attractor 100 ms after distractor onset (Fig. 6J), matching the peak time of neural bias observed in the experimental data.

Statistical Analyses

We utilized the non-parametric Wilcoxon signed rank test for comparing median errors or behavioral and neural bias across conditions (e.g., cued versus uncued or distractor-same versus opposite) (Fig. 1B, F-G; 1E, 2C-G; 3B-D; 4A-C). All tests were two-sided, unless stated otherwise. In addition, we also report the Bayes factor (BF)[70], a statistical measure that compares the likelihood of the alternate hypothesis to that of the null hypothesis, under the observed data. The Bayes factor was computed with the Matlab Bayes Factor toolbox. To study the effect of distractor condition and cueing on precision in visual WM (Fig. 1C), we performed a 2-way ANOVA with the circular standard deviation as the dependent variable and distractor condition (distractor-same, distractor-opposite and no-distractor) and cueing (cued and uncued) as predictors; participants was treated as random effects. *Post-hoc* analyses were performed by averaging the values across the cue conditions and comparing CSD values for distractor-same versus distractor-

opposite conditions with Tukey's h.s.d. procedure, using the *multcompare* function in Matlab. To find significant timepoints in decoding accuracies plotted as a function of time (Fig. 2B and 3B), we used the two-sided cluster-based permutation test[71] with a cluster-forming threshold of $p<0.05$ and with 10,000 permutations of the data.

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

Competing interests

Devarajan Sridharan is a research consultant at Google.

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

During the preparation of this work the first author used ChatGPT, Open AI in order to rephrase sentences. After using this tool/service, the first author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

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

S.G. and D.S. designed the study; S.G. conducted the experiment; S.G. analyzed the data; S.G. and D.S. designed the model; 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. Feldmann-Wüstefeld, T. & Vogel, E. K. Neural Evidence for the Contribution of Active Suppression During Working Memory Filtering. *Cerebral Cortex* **29**, (2019).

26. 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).

27. 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).

28. 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).

29. 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).

30. 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).

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

32. 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).

33. 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).

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

35. 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).

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

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

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

39. 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).

40. 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).

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

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

43. 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).

44. 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).

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

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

47. 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).

48. 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).

49. 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).

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

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

52. 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).

53. 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).

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

55. 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).

56. 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).

57. 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.

58. 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).

59. 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).

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

61. 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).

62. 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).

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

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

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

66. 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).

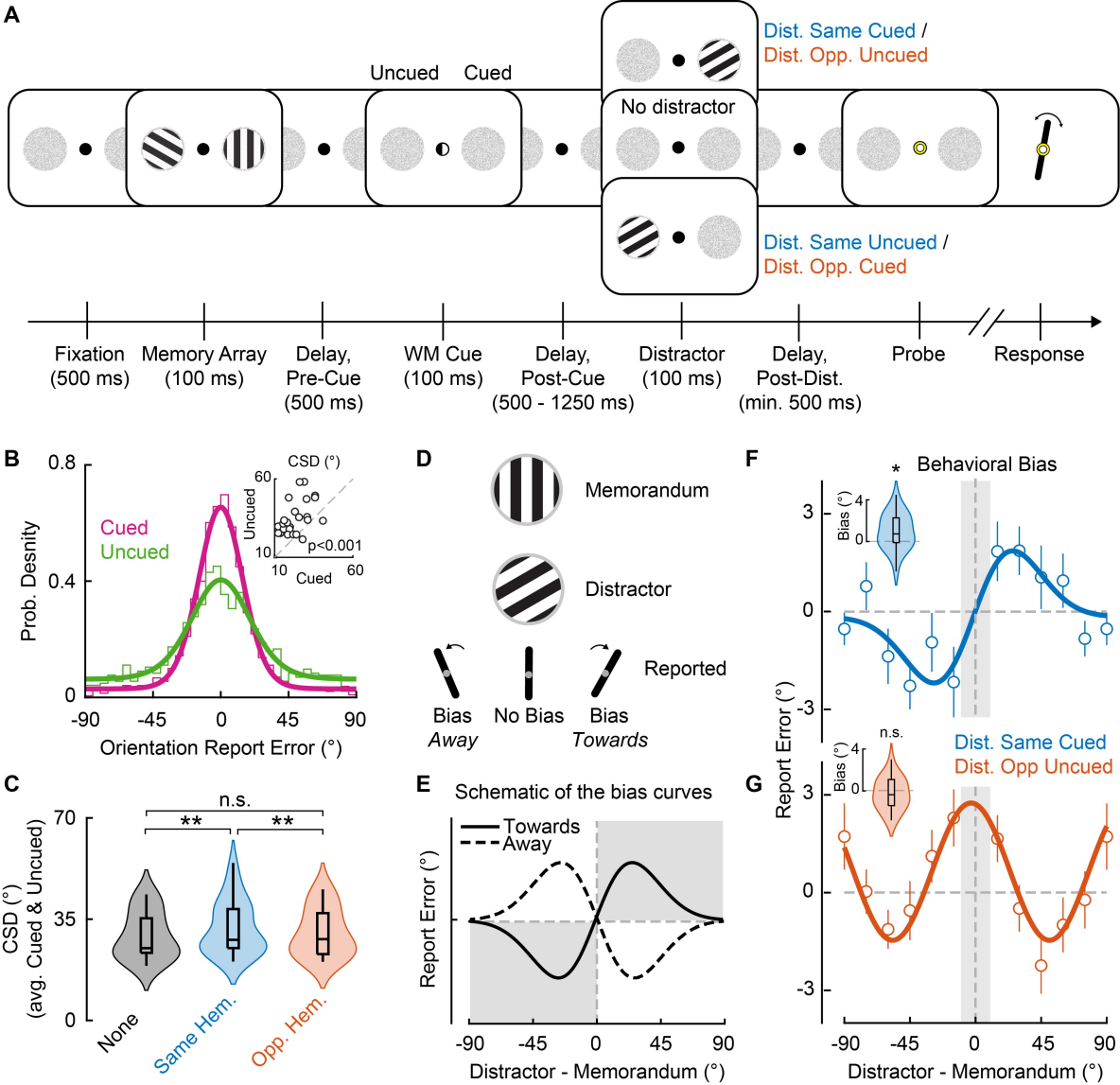
67. Lopez-Calderon, J. & Luck, S. J. ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Front Hum Neurosci* **8**, (2014).

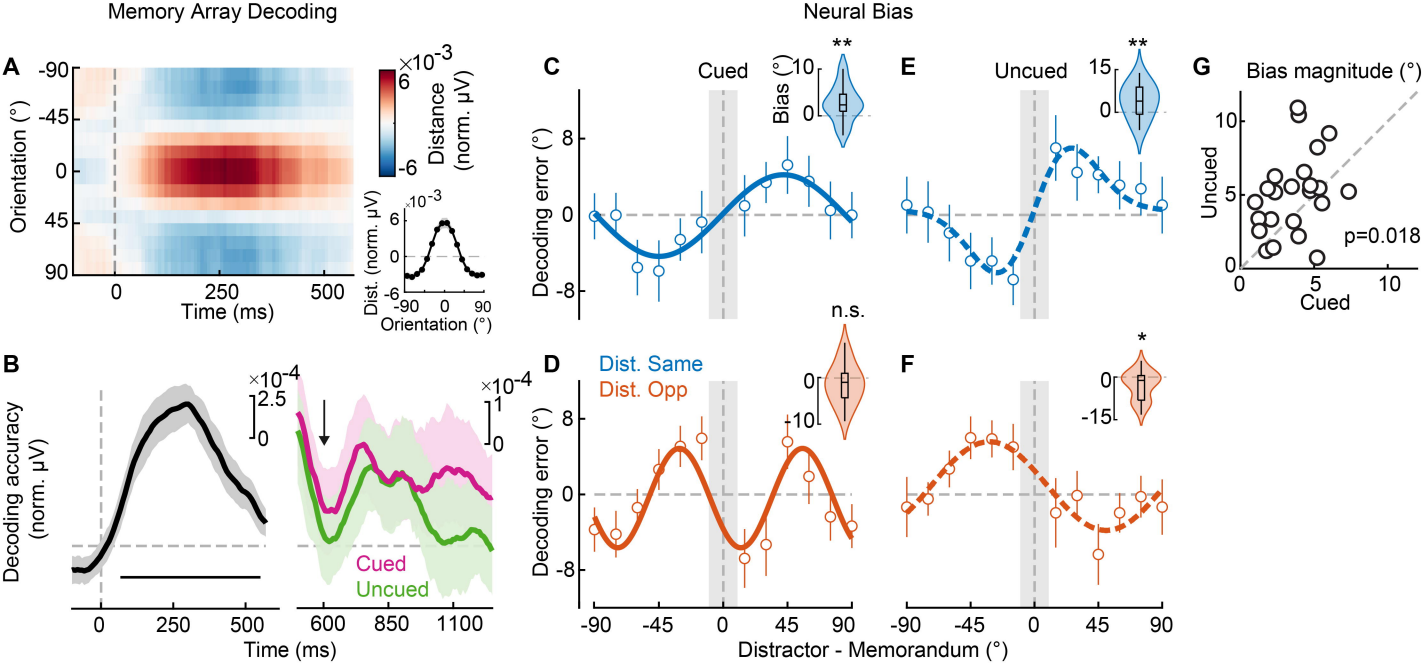
68. Pion-Tonachini, L., Kreutz-Delgado, K. & Makeig, S. ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *Neuroimage* **198**, 181–197 (2019).

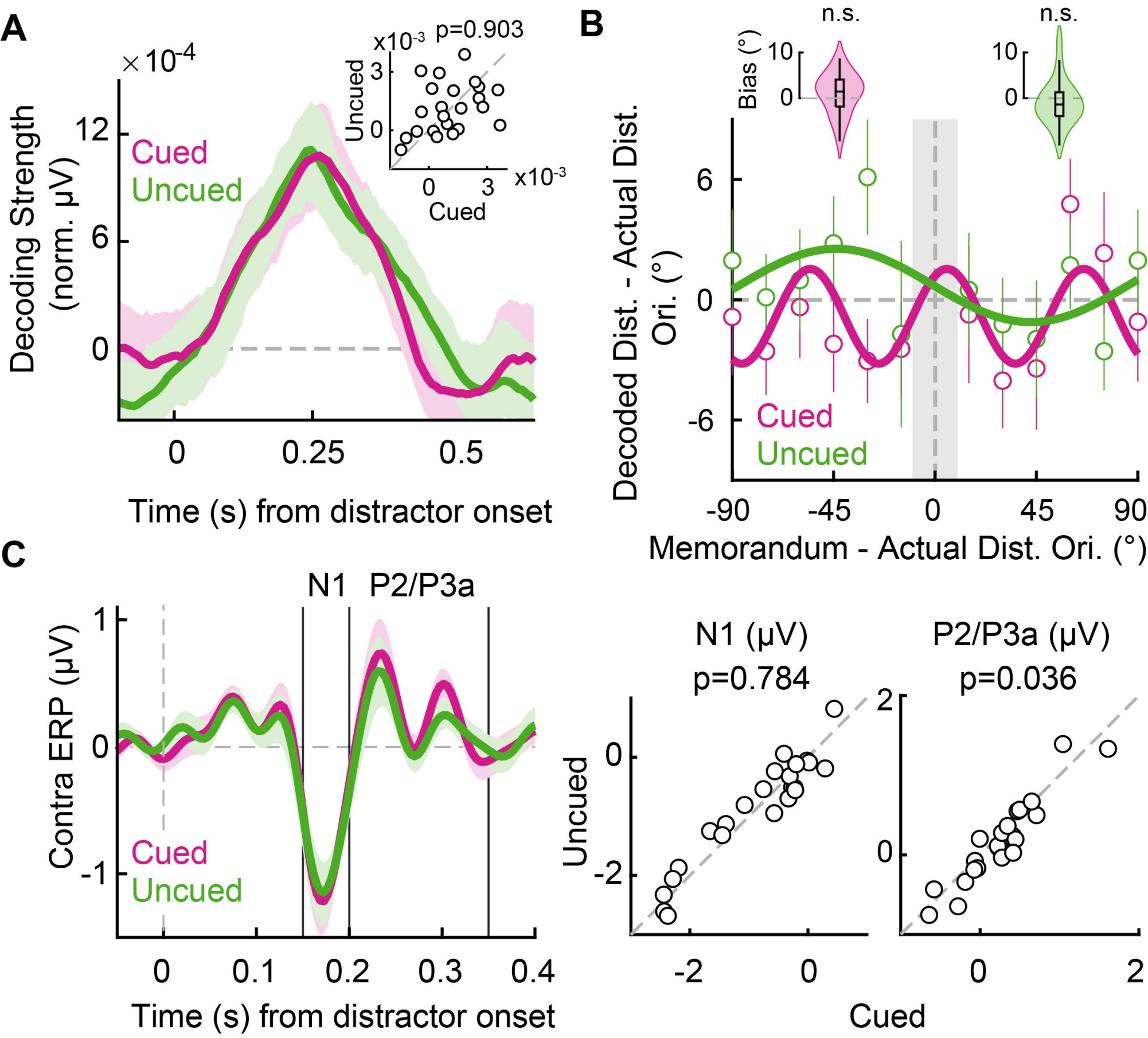
69. Junghöfer, M., Elbert, T., Tucker, D. M. & Rockstroh, B. Statistical control of artifacts in dense array EEG/MEG studies. *Psychophysiology* **37**, 523–532 (2000).

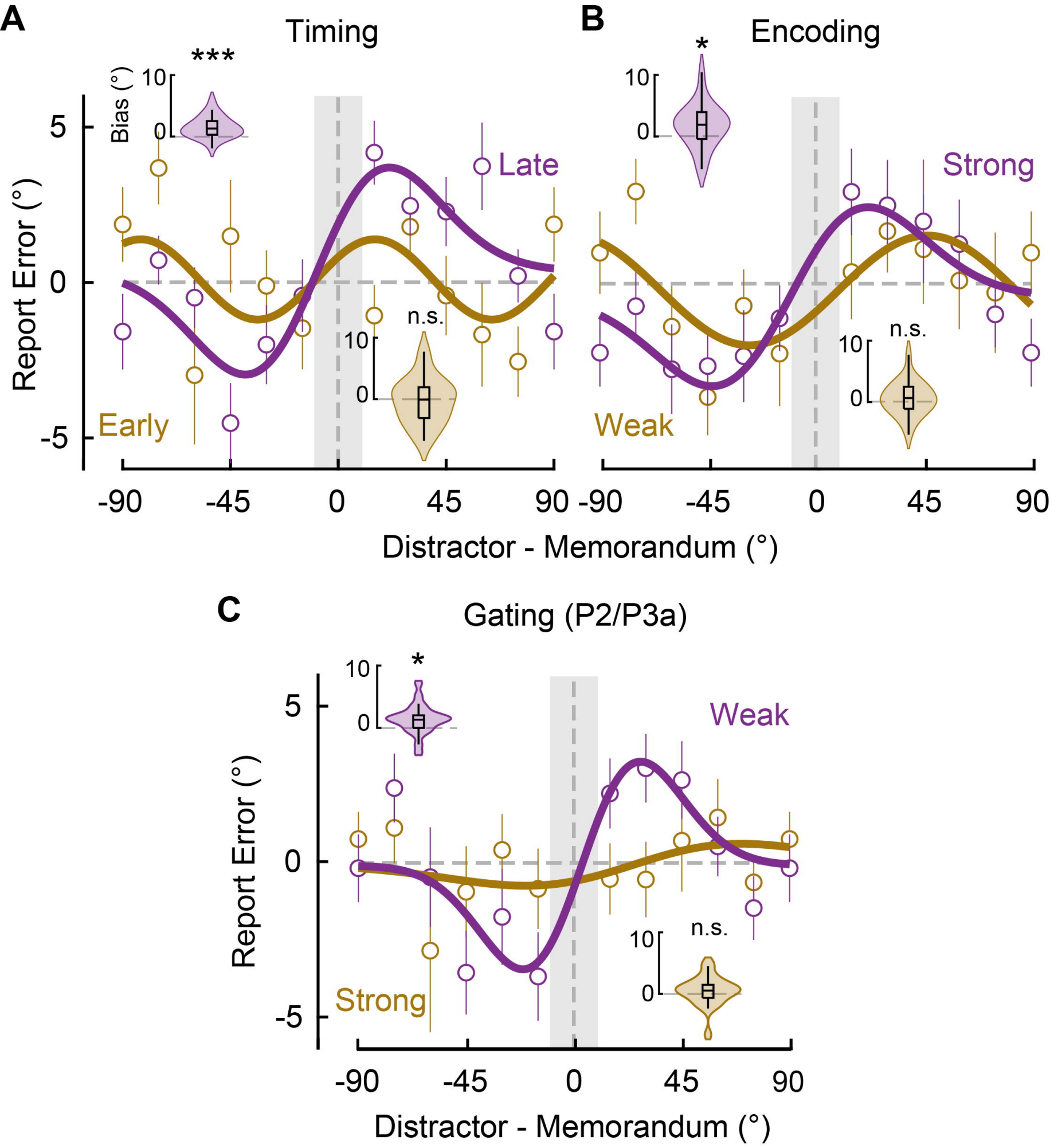
70. Jarosz, A. F. & Wiley, J. What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *J Probl Solving* **7**, 2 (2014).

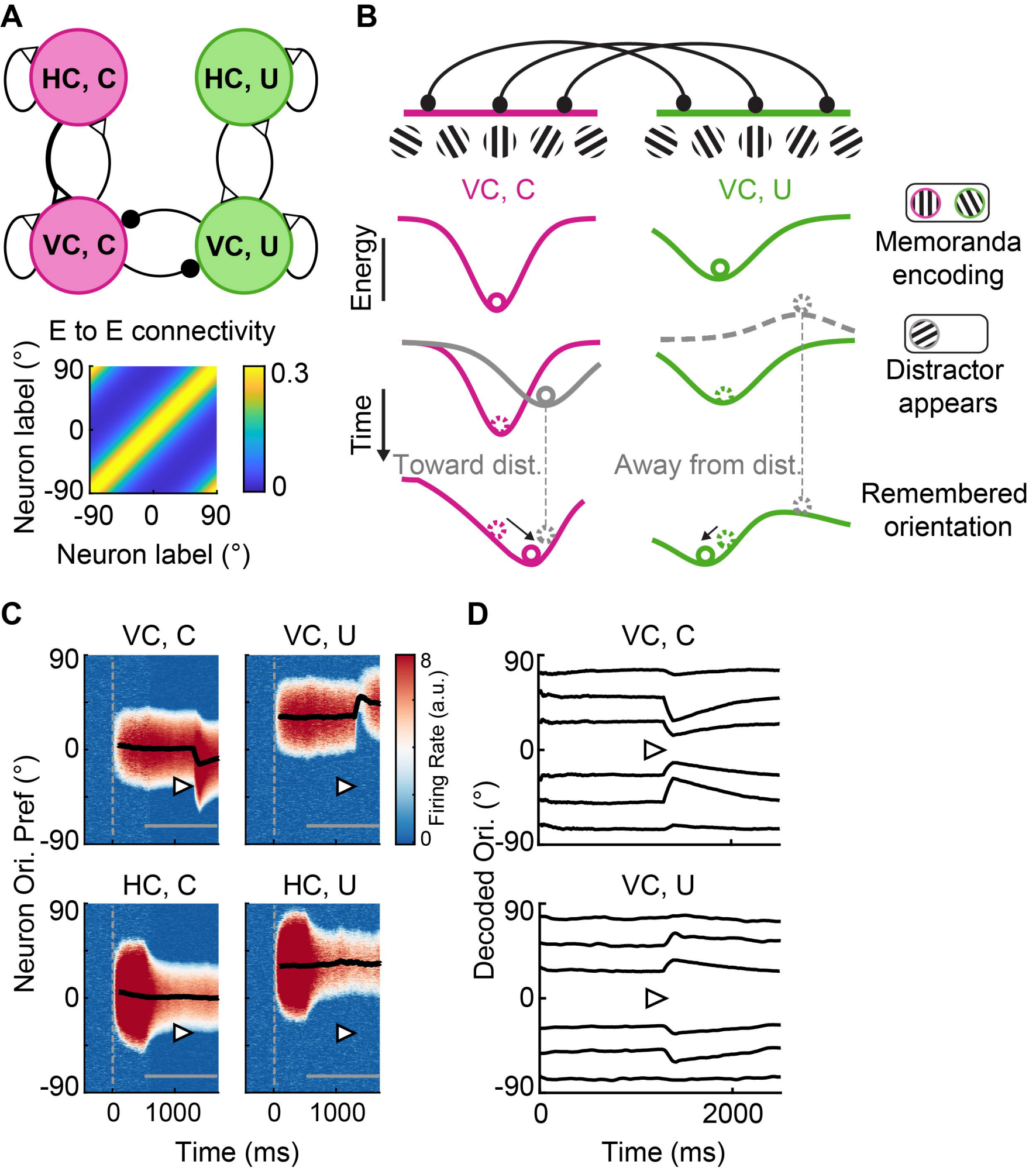
71. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J Neurosci Methods* **164**, 177–190 (2007).

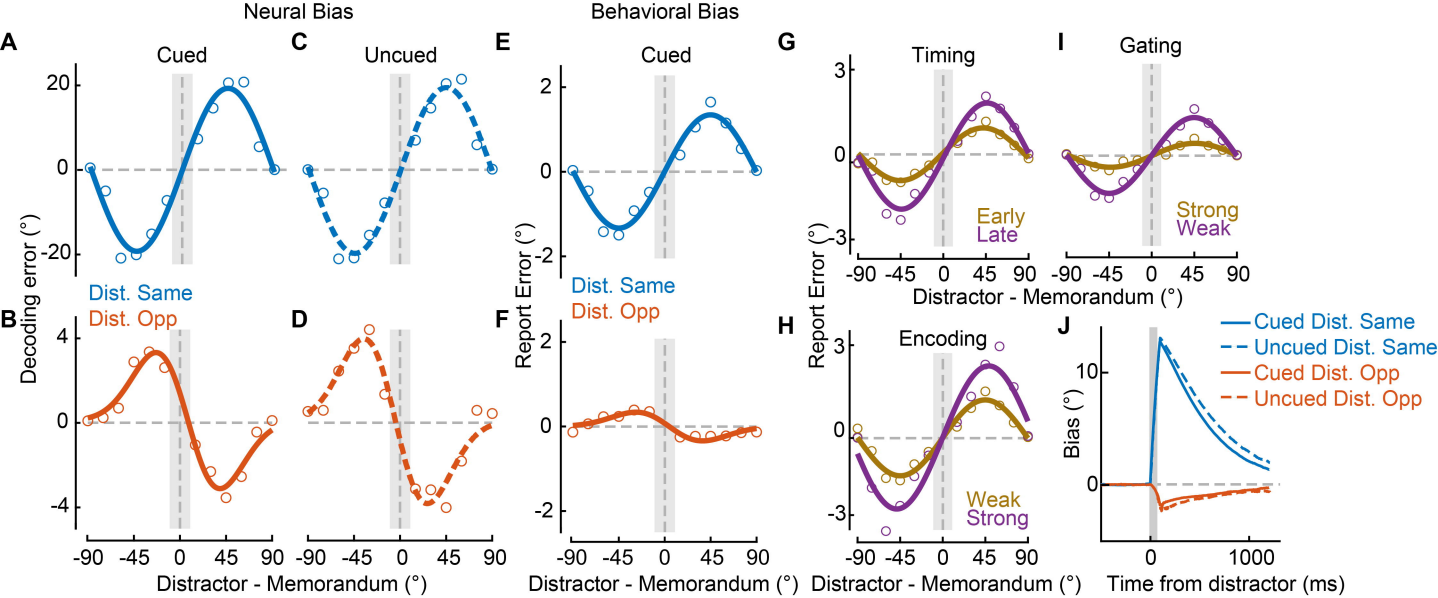

A
Uncued
Cued
No distractor
Dist. Same Cued /
Dist. Opp. Uncued
Dist. Same Uncued /
Dist. Opp. Cued
Fixation (500 ms)
Memory Array (100 ms)
Delay, Pre-Cue (500 ms)
WM Cue (100 ms)
Delay, Post-Cue (500 - 1250 ms)
Distractor (100 ms)
Delay, Post-Dist. (min. 500 ms)
Probe
Response
B
Prob. Desnity
Cued
Uncued
CSD (°)
Uncued
Cued
p<0.001
Orientation Report Error (°)
C
CSD (°) (avg. Cued & Uncued)
n.s.
**
**
None
Same Hem.
Opp. Hem.
D
Memorandum
Distractor
Reported
Bias Away
No Bias
Bias Towards
E
Schematic of the bias curves
Towards
Away
Report Error (°)
Distractor - Memorandum (°)
F
Behavioral Bias
Bias (°)
Report Error (°)
G
n.s.
Dist. Same Cued
Dist. Opp Uncued
Distractor - Memorandum (°)


Memory Array Decoding
Neural Bias
A
Orientation (°)
Time (ms)
Distance (norm. μV)
Dist. (norm. μV)
Orientation (°)
B
Decoding accuracy (norm. μV)
Time (ms)
Cued
Uncued
C
Cued
Bias (°)
Decoding error (°)
D
Dist. Same
Dist. Opp
n.s.
E
Uncued
F
Distractor - Memorandum (°)
G
Bias magnitude (°)
Uncued
Cued
p=0.018


A
×10-4
Decoding Strength (norm. µV)
12
6
0
Cued
Uncued
0
0.25
0.5
Time (s) from distractor onset
x10-3
p=0.903
Uncued
3
0
0
3
x10-3
Cued
B
n.s.
n.s.
Bias (°)
10
0
10
0
Decoded Dist. - Actual Dist. Ori. (°)
6
0
-6
Cued
Uncued
-90
-45
0
45
90
Memorandum - Actual Dist. Ori. (°)
C
N1
P2/P3a
Contra ERP (µV)
1
0
-1
Cued
Uncued
0
0.1
0.2
0.3
0.4
Time (s) from distractor onset
N1 (µV)
p=0.784
P2/P3a (µV)
p=0.036
Uncued
0
-2
2
0
-2
0
0
2
Cued


A
Timing
***
Bias (°)
10
0
5
0
-5
Report Error (°)
Late
Early
n.s.
10
0
-90
-45
0
45
90
B
Encoding
*
10
0
Strong
Weak
n.s.
10
0
-90
-45
0
45
90
Distractor - Memorandum (°)
C
Gating (P2/P3a)
*
10
0
5
0
-5
Report Error (°)
Weak
Strong
n.s.
10
0
-90
-45
0
45
90
Distractor - Memorandum (°)


A
HC, C
HC, U
VC, C
VC, U
E to E connectivity
Neuron label (°)
90
0
-90
-90
0
90
0.3
0
Neuron label (°)
B
VC, C
VC, U
Energy
Memoranda encoding
Distractor appears
Time
Toward dist.
Away from dist.
Remembered orientation
C
VC, C
VC, U
HC, C
HC, U
Neuron Ori. Pref (°)
90
0
-90
8
0
Firing Rate (a.u.)
0
1000
Time (ms)
D
VC, C
VC, U
Decoded Ori. (°)
90
0
-90
0
2000
Time (ms)


Neural Bias
Behavioral Bias
A
Cued
C
Uncued
E
Cued
G
Timing
I
Gating
Decoding error (°)
Report Error (°)
Dist. Same
Dist. Opp
Early
Late
Strong
Weak
B
D
F
H
Encoding
J
Weak
Strong
Cued Dist. Same
Uncued Dist. Same
Cued Dist. Opp
Uncued Dist. Opp
Bias (°)
Distractor - Memorandum (°)
Time from distractor (ms)