Dynamic hidden states underlying working memory guided behaviour

Michael J. Wolff[1,2] , Janina Jochim[1] , Elkan G. Akyürek[2] & Mark G. Stokes[1]

[1] Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom

[2] Department of Experimental Psychology, University of Groningen, Groningen, The Netherlands

Corresponding author: Mark G. Stokes, mark.stokes@psy.ox.ac.uk

Summary of Main Finding: Wolff and colleagues show that 'activity-silent' brain states play an important role in working memory. Using a novel perturbation method to 'ping the brain', they uncover hidden neural states that reflect temporary information held in mind, and predict memory performance. They argue that dynamic hidden states could underpin working memory.

Abstract

Recent theoretical models propose that working memory is mediated by rapid transitions in 'activity-silent' neural states (e.g., short-term synaptic plasticity). According to the dynamic coding framework, such hidden state transitions flexibly configure memory networks for memory-guided behaviour, and dissolve them equally fast to allow forgetting. We developed a novel perturbation approach to measure mnemonic hidden states in electroencephalogram (EEG). By 'pinging the brain' during maintenance, we show that memory item-specific information is decodable from the impulse response, even in the absence of attention and lingering delay activity. Moreover, hidden memories are remarkably flexible: An instruction cue that directs people to forget one item is sufficient to wipe the corresponding trace from the hidden state. In contrast, temporarily unattended items remain robustly coded in the hidden state, decoupling attentional focus from cue-directed forgetting. Finally, the strength of hidden-state coding predicts the accuracy of working memory guided behaviour, including memory precision.

Working memory (WM) is a core cognitive function critical for flexible, intelligent behaviour[1]. Until recently, it was widely assumed that information is maintained in WM by maintaining specific activity states that represent the specific memoranda [2,3]. However, accumulating evidence increasingly shows that successful maintenance in WM is not strictly dependent on an unbroken chain of corresponding delay activity [4], and that item-specific activity states could reflect other cognitive processes. For example, in monkey studies persistent activity ramps up with expectation of the probe [5–8]. Similarly, in the human it has been shown that unattended WM content is not reflected in the neural signal, even when it is still clearly maintained [9–11]. Evidence for WM in the absence of persistent delay activity suggests that WM can be maintained in 'activity silent' neural states [4].

Recent theories acknowledge that brain activity is highly dynamic, even when the contents of working memory remain stable [12]. Multiple neurophysiological mechanisms could underlie such dynamics [13–15]. According to a dynamic coding model of WM [4], behaviourally relevant sensory input drives a memory item-specific neural response, which triggers an item-specific change in the functional state of the system. Depending on the precise neural mechanism, this functional state could be activity-silent (e.g., short-term synaptic plasticity [14,16–19]), and maintained throughout the memory delay to serve as the neural context for subsequent processing. Items in WM would be read-out via the context-dependent response to a probe stimulus during recall [13,20]. Crucially, this model predicts that dynamic hidden states are constructed when new information is encoded, and dissolved as soon as it is forgotten. This model also predicts that dynamic hidden states should determine the quality of a representation maintained in WM.

To probe hidden neural states, we developed a functional perturbation approach to 'ping the brain'. Analogous to the idea of active sonar (or echolocation), the response to a well-characterised impulse stimulus can be used to infer the current state of the system [4,13]. We recently validated this general approach using non-invasive electroencephalography (EEG) in a proof of principle study[21]. The presentation of a high contrast, neutral visual stimulus evoked neural activity that clearly discriminated the previously presented visual stimulus. Here, we exploit this approach to track the functional dynamics of hidden states for WM.

Across two experiments, we show that the content of WM can be decoded from the impulse response during the maintenance interval, while forgotten information leaves effectively no trace. In Experiment 2, we also demonstrate robust hidden-state representation for unattended content in WM, providing a plausible mechanism for maintenance that is independent of the activity associated with the focus of attention. Finally, we also find evidence that the quality of working memory varies with the decodability of these hidden states.

Results

Experiment 1:

In Experiment 1, 30 human participants performed a visual WM task while EEG was recorded. At the beginning of each trial (see Fig. 1a), two memory items were presented, but a retrospective cue (retro-cue) presented during the delay instructed participants which item would actually be probed [22,23]. The other item could be simply forgotten. The retro-cue in this design is essential to differentiate WM from basic stimulation history [24]. During a subsequent memory delay, we then presented a high contrast "impulse" stimulus. Memory performance for the cued item was tested after the impulse by a centrally presented memory probe (Fig. 1b). Time-frequency decomposition of lateralised activity in posterior sensors (Fig. 1c) shows significant lateralization in the alpha range (8-12 Hz) after the presentation of the cue (permutation test, n = 30, $p < 0.001$, corrected, cluster-forming threshold $p < 0.05$). This pattern is consistent with a shift in spatial attention [25] according to the retro-cue, which confirms that the cue manipulation was effective.

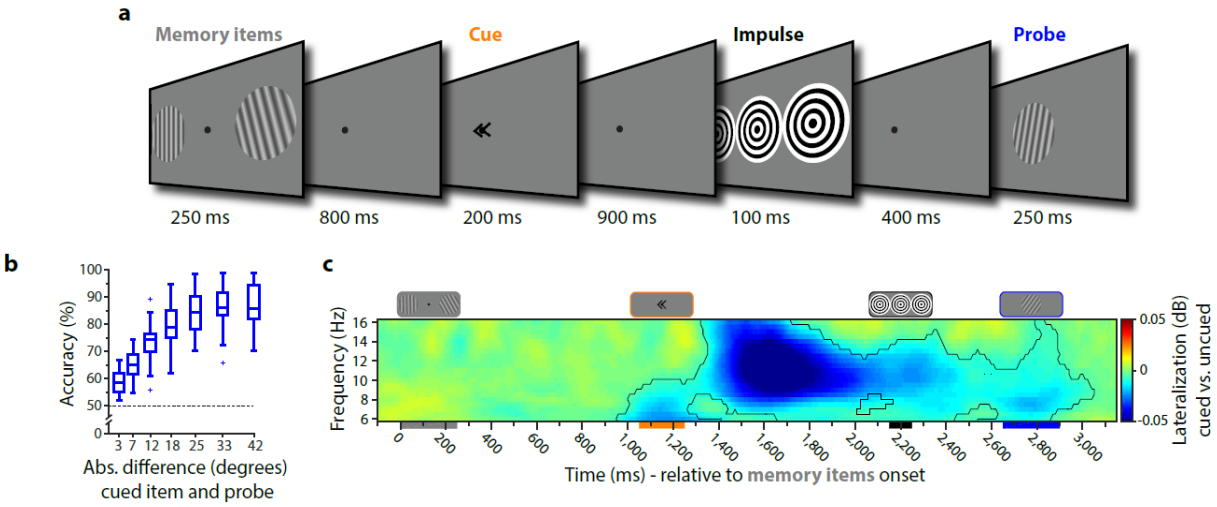


Figure 1. Experiment 1 task structure, behavioural performance and attention-related alpha band activity. a. Trial schematic. Two memory items were presented (randomly oriented grating stimuli), and participants were instructed to memorize both orientations. A retro-cue then indicated which item would actually be tested at the end of the current trial (100% valid). The

impulse stimulus (high contrast, task-irrelevant visual input) was then presented during the subsequent delay while participants should have only the cued item in WM. At the end of the trial, a forced-choice probe was presented at the centre of the screen. Participants indicated whether the probe was rotated clockwise or anti-clockwise relative to the orientation of the cued item. **b.** Boxplots show WM accuracy as a function of the absolute angular difference (in degrees) between the memory item and the probe. Data points outside of the 1.5 * interquartile range are shown separately (small crosses). **c.** Time-frequency representation of the difference between the contra- and ipsilateral posterior electrodes relative to the cued hemifield. The highlighted cluster in the alpha frequency band (8-12 Hz) indicates significant contralateral desynchronization (permutation test, $n = 30$, cluster-forming threshold $p < 0.05$, corrected significance level $p < 0.05$). The coloured bars under the x-axis represent the timings of the corresponding stimuli illustrated on top.

Decoding parametric memory items: To decode the memory items used in this experiment, we developed a parametric variant of distance-based discrimination (see Online Methods, Fig. 2a-d). As shown in Fig 2a, this capitalises on the parametric structure of the stimulus space [26], whilst maintaining the statistical advantages of the Mahalanobis distance metric used in previous EEG/MEG decoding studies [21,27] (see Online Methods). To summarise briefly here: for a given trial, we compare the activity pattern across electrodes to the corresponding activity pattern observed in the remaining trials, averaged by orientation-difference to the test trial (at a bin width of 30 degrees). This procedure is repeated for all trials and all time-points. If the pattern of activity contains information about item orientation, we expect greater pattern dissimilarity (i.e., Mahalanobis distance) at larger angular differences. Fig. 2b shows distance as a function of reference angle and time after the presentation of the left and right item separately (upper/lower respectively). Distance values were then converted into a decoding accuracy score (Fig. 2c) and averaged across both items at each time-point (Fig. 2d). Item orientation could be decoded from 56 ms until 1026 ms after onset (permutation test, $n = 30$, $p < 0.001$ (corrected), cluster-forming threshold $p < 0.05$). This is consistent with previous empirical evidence that EEG is sufficiently sensitive to detect subtle differences in scalp-level activity patterns associated with different stimulus orientation [21]. The current decoding results further validate the utility of multivariate pattern analysis for two simultaneously presented orientation gratings. For completeness, we also decode item-specific orientation during the retro-cue epoch (Supplementary Fig. 1).

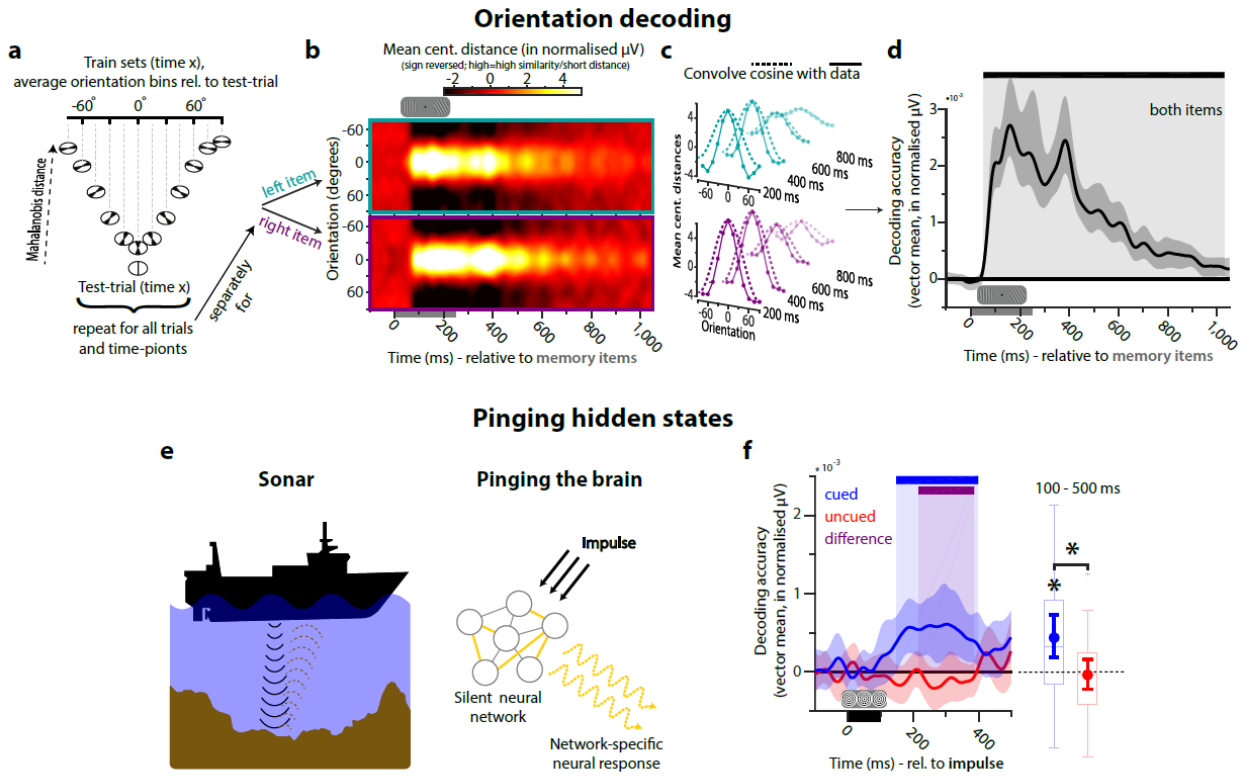


Figure 2. Orientation decoding in EEG and pinging hidden states of WM. a-d. Decoding procedure. **a.** The dissimilarity in the neural pattern between a single trial and all other trials is computed as a function of orientation difference (binned: 30 degrees). **b.** Average distance to template of all trials for each time-point during and after memory item presentation, plotted separately for the left and the right memory item (upper/lower respectively). Distances are mean centred and sign reversed (high = small distance/high similarity) for visualization. **c.** A cosine is convolved with the data. **d.** The vector mean of the convolved tuning curves (i.e., decoding accuracy) over time, averaged over left and right items. The black bar indicates significant decoding (permutation test, $n = 30$, cluster-forming threshold $p < 0.05$, corrected significance level $p < 0.05$). Error shading is the 95 % C.I. of the mean. **e.** Pinging hidden states. Analogy to active sonar: differences in hidden state are inferred from differences in the measured response to a well-characterised impulse. **f.** Decoding results in the impulse epoch. The blue bar indicates significant decoding of the cued item. The purple bar indicates significant difference in decodability between the cued and uncued item (permutation test, $n = 30$, cluster-forming threshold $p < 0.05$, corrected significance level $p < 0.05$). Error shading is the 95 % C.I. of the mean. The boxplots and superimposed circles with error-bars (mean and 95 % C.I. of the mean)

represent average decoding from 100 to 500 ms after impulse onset. Data points outside of the 1.5 * interquartile range are shown separately (small crosses). Significant average decoding and significant difference in average decodability between the cued and uncued item are marked by asterisks (permutation test, $n = 30$, $p < 0.05$).

Pinging hidden states: According to the dynamic coding framework, we hypothesised that the input/output mapping of neural circuits maintaining information in WM should systematically reflect the memory content [4]. We tested this using an impulse stimulus to 'ping' potentially hidden neural states (Fig. 2e). As predicted, the impulse-specific response clearly differentiated the content of WM (Fig. 2f), even though the driving input ('ping') was held constant on each trial. The decodability of the cued item showed a significant cluster from 148 to 398 ms after impulse stimulus onset (permutation test, $n = 30$, $p = 0.002$, corrected, cluster-forming threshold $p < 0.05$). Average decodability from 100 to 500 ms was also significant ($p = 0.004$). Cued item decoding was also higher than task-irrelevant (uncued) item decoding (cluster: 216 to 386 ms, $p = 0.009$, corrected; average: $p = 0.028$). Indeed, the uncued item showed no evidence for decoding (no corrected clusters; average: $p = 0.687$), suggesting that content can be rapidly purged from WM when instructed, leaving effectively no trace in the neural state.

To test whether the impulse response reflects a literal 'reactivation' of item-specific activity observed during encoding (e.g., Fig. 2b), we also examined whether a classifier trained on the activity elicited by the memory stimuli during encoding could be used to decode the memory item during the impulse epoch (and vice versa). However, we found no evidence for significant cross-generalization between discriminative activity patterns during encoding and discriminative activity driven by the impulse (corrected clusters, $p > 0.347$). We propose that the impulse stimulus simply acts as a functional ping to recover hidden states, rather than a literal 'reactivation' of a latent representation[21].

Trial-wise variability in decoding the impulse response also predicted variability in WM performance. Higher decoding trials of the cued item were accompanied by higher performance than low decoding trials (permutation test, $n = 30$, $p = 0.043$; Fig 3a, left). There was also a complementary cost for decoding the uncued item (i.e., a high decoding score for the uncued item led to a decrease in accuracy on the cued item; $p = 0.002$; Fig 3b, left), suggesting that

participants might have failed to discard the uncued item (or simply did not use the cue properly) on some trials, contributing to error in performance. Finally, the difference between the accuracy effect of the cued and the uncued item was also significant (permutation test, $n = 30$, $p < 0.001$).

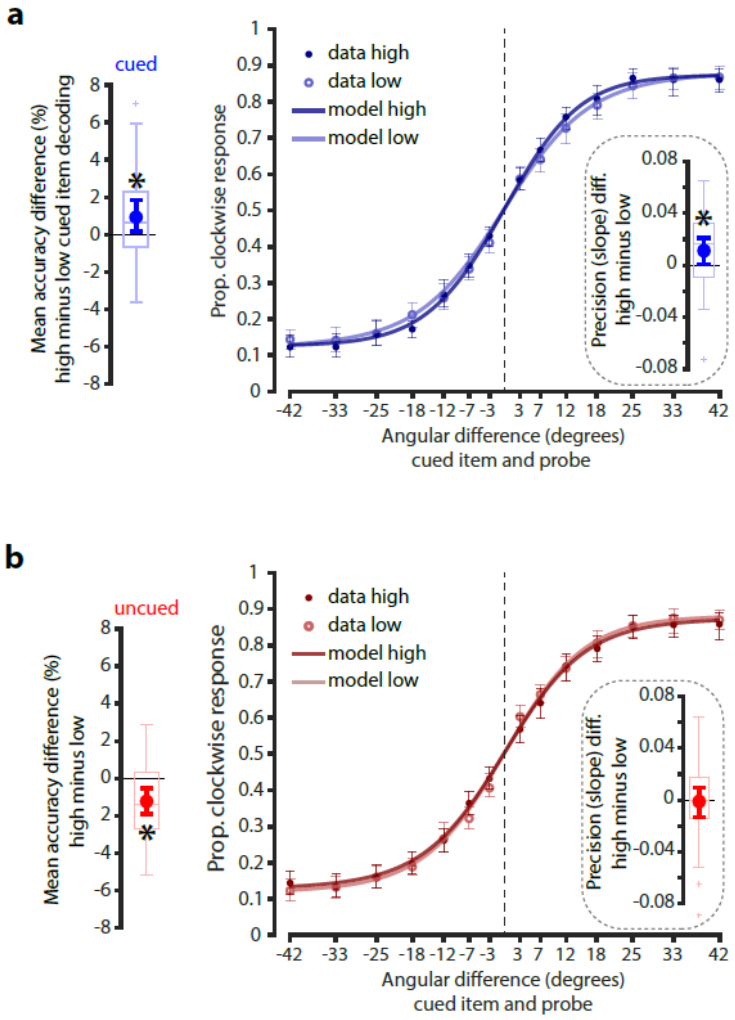


Figure 3. Relationship between item-specific impulse decoding and WM accuracy. a. Difference in overall WM task performance between high and low cued item decoding trials (left). Proportion clockwise response for high and low decoding trials as a function of the angular difference between the memory item and the probe (right). Inset shows the difference in the slope parameter (a measure of memory precision) between high and low decoding trials. Data points outside of the 1.5 * interquartile range are shown separately in the boxplots (small crosses). Superimposed circles and error-bars are the mean and 95% C. I. of the mean. **b.** The same convention as in a. but for the decoding of the uncued item. Significant differences in

accuracy/precision between high and low decoding trials are highlighted by asterisks (permutation test, $n = 30$, $p < 0.05$).

In principle, the relationship between trial-wise decoding and WM performance may rest on an increase in guess-rate (i.e., due to forgetting or failure to encode), or a reduction in precision, or both [28,29]. To separate these possible contributions, we modelled the behavioural profile over degrees of angular rotation between the memory item and the probe stimulus (see Online Methods) [30,31]. We found that the link to behaviour is most likely driven by a decrease in precision (the slope parameter of the model) for weakly encoded hidden states of WM (permutation test, $n = 30$, $p = 0.023$, one-tailed; Fig. 3a, right), while no evidence for an effect in guess rate (the asymptote parameter) was found ($p = 0.867$, one-tailed). Modelling the observed uncued item accuracy effect was inconclusive (Fig. 3b, right), with no evidence for either a precision or guess rate effect ($p = 0.443$ and $p = 0.184$ respectively, one-tailed). Finally, we found no evidence that trial-wise item decoding during the initial presentation of the memory stimuli relates to memory performance (Supplementary Fig. 2a), further suggesting that the relationship between accuracy and decoding triggered by the impulse is not due to a failure to encode the memory item.

Experiment 2

Recently, it has been proposed that information in WM can be represented in qualitatively different states [32–34], with attended items encoded in activity states measurable with standard recordings of delay activity, whereas activity-silent states could underlie the representation of currently unattended information in WM. In Experiment 2 ($n = 19$) we test whether unattended but nevertheless remembered information in WM can still be decoded from the impulse response. Again, two memory items were presented at the start of the trial, however both were ultimately relevant as they would both be probed. Priority was manipulated by blocking the order in which items would be probed (Fig. 4a), and instructing participants accordingly. Because there was no other clue as to which item was being probed first or second, non-random responses already indicate that participants used this blocked information (Fig. 4b). This was further supported by lateralised changes in alpha power (Fig. 4c). During and shortly after the initial presentation of the memory stimuli, there was a relative decrease in power at sensors contralateral to the initially prioritised item, consistent with selective allocation of attention

(permutation test, $n = 19$, $p = 0.023$, corrected, cluster-forming threshold $p < 0.05$). Moreover, this pattern reversed after the response to the first item ($p = 0.009$, corrected), consistent with the assumption that participants then shift the originally de-prioritised item into the focus of attention in WM in preparation for the second probe [35].

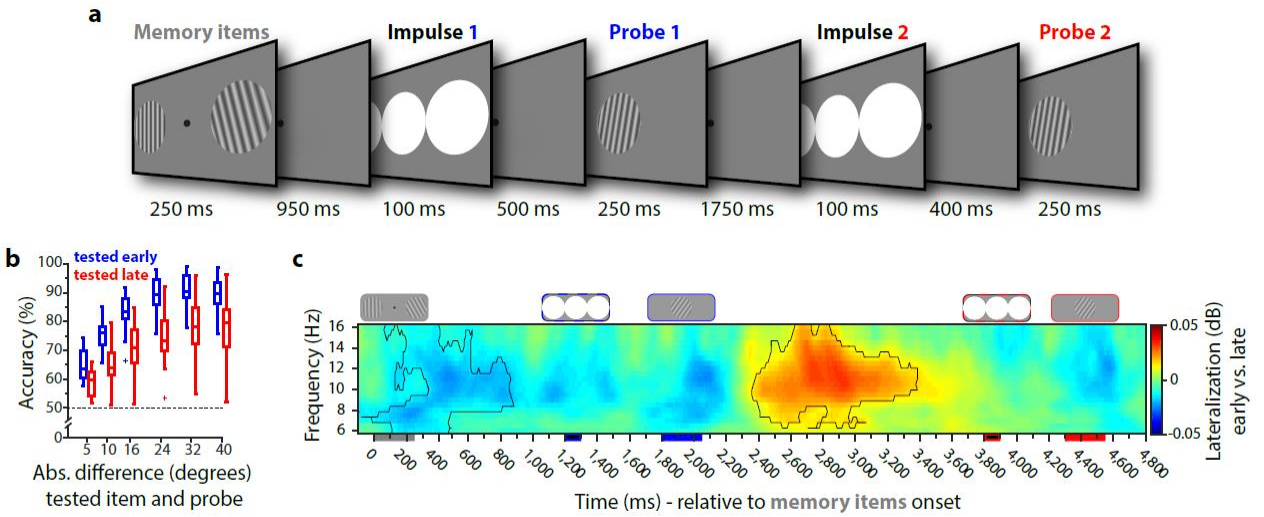


Figure 4. Experiment 2 task structure, behavioural performance and attention-related alpha band activity. a. Trial schematic. Two memory items were presented. Participants were instructed to maintain both items and were told at the start of each block which order the items would be tested. The first impulse was presented within the first memory delay (maintain both items, but attend the prioritised item), after which the prioritised item was probed. The second impulse was presented during the subsequent memory delay (maintain and attend only the now-prioritised item), after which the remaining item was probed. **b.** Boxplots show the accuracy of the early and late tested item as a function of the absolute angular difference (in degrees) between the memory item and the probe. Data points outside of the 1.5 * interquartile range are shown separately in the boxplots (small crosses). **c.** Time-frequency representation of the difference between the contra- and ipsilateral posterior electrodes relative to the presentation side of the early tested memory items. Highlighted areas indicate significant difference (permutation test, $n = 19$, cluster-forming threshold $p < 0.05$, corrected significance level $p < 0.05$).

Decoding during stimulus presentation: We first analysed decoding during the initial processing of the memory stimuli. The results are plotted separately as a function of test-time (early or late

in the trial) as this could be meaningfully classified from the beginning of the trial (Fig. 5a). As expected, decoding the prioritised item (cluster: 74 to 1,200 ms, $p < 0.001$, corrected, cluster-forming threshold $p < 0.05$; average: $p < 0.001$), relative to the de-prioritised item (cluster: 82 to 542 ms, corrected, $p < 0.001$, corrected; average: $p < 0.001$) was more robust (average: $p = 0.013$). While decoding of the unattended item drops to chance relatively quickly after item presentation, the attended item shows significant decoding until the end of the epoch, replicating previous evidence showing that maintenance of only attended WM items is represented in the recorded brain activity patterns [9–11].

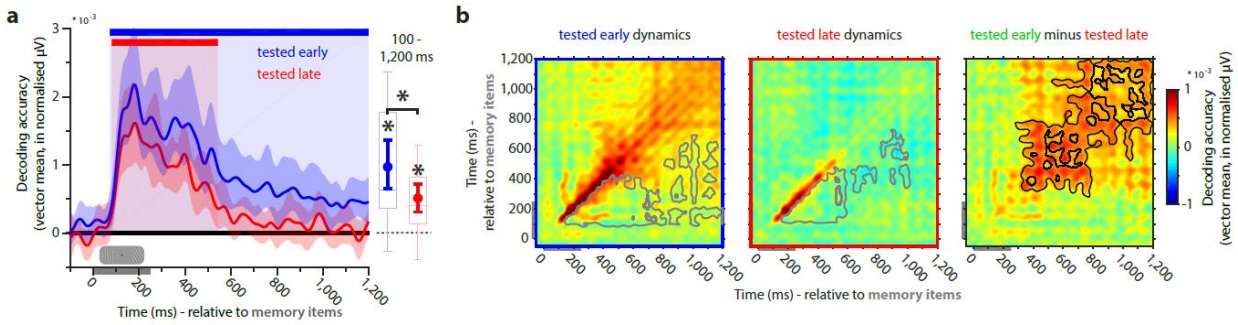


Figure 5. Priority-dependent encoding and maintenance in WM. a. Decodability of the item that is tested early (blue) and the item that is tested late (red) during memory item presentation. Blue and red bars indicate significant decoding clusters for the early- and late-tested item, respectively (permutation test, $n = 19$, cluster-defining threshold $p < 0.05$, corrected significance level $p < 0.05$). Error shading is 95% C.I. of the mean. Boxplots and superimposed circles with error bars (mean and 95 % C.I. of the mean) represent average decodability from 100 ms after stimulus onset until the end of the epoch. Significant average decoding and average difference between the decodability of the early and late item are marked by an asterisk (permutation test, $n = 19$, $p < 0.05$). **b.** Cross-temporal decoding matrices of the early (left) and late-tested (middle) item derived from training and testing on all time-point combinations, and the difference between the decoding of the early and late tested item (right). The grey outline indicates time-points of significantly lower decoding relative to both equivalent time-points along the diagonal, which is taken as evidence for dynamic coding (permutation test, $n = 19$, cluster-defining threshold $p < 0.05$, corrected significance level $p < 0.05$). The black outline (right) indicates

significantly higher decodability of the early compared to the late tested item (permutation test, $n = 19$, cluster-defining threshold $p < 0.05$, corrected significance level $p < 0.05$).

The difference between attended and unattended item-maintenance in WM was even more apparent when comparing their cross-temporal decoding matrices. Minimal cross-temporal generalization during and shortly after memory item presentation suggested highly dynamic item encoding: orientation discriminative patterns change over time. This was supported by significant dynamic coding clusters during item encoding for both the early and late tested item, where off-diagonal time-points show significantly lower decodability than both corresponding on-diagonal time-points (permutation test, $n = 19$, cluster-defining threshold $p < 0.05$, corrected significance level $p < 0.05$; see Online Methods; Fig. 5b, left and middle). However, the attended item clearly showed a more time-invariant decoding pattern at the end of the epoch than the unattended item, apparent by both significantly higher decodability on same time-point as well as cross time-point decoding ($n = 19$, $p = 0.023$, corrected, cluster-forming threshold $p < 0.05$; Fig. 5b right). This further suggests that while the attended item also has a corresponding WM maintenance signature in stable activity patterns, the unattended item does not.

Decoding of the impulse responses: Critically, we found that both the attended (clusters: 80 to 308 ms, $p = 0.004$, and 332 ms to 434 ms, $p = 0.031$, corrected; average: $p < 0.001$) and unattended items (cluster: 172 to 306 ms, $p = 0.011$, corrected; average: $p = 0.045$) were decodable in the first impulse response (Fig. 6a). This contrasts with the clear cueing differences observed in Experiment 1, and suggests multiple items can be encoded in hidden states and revealed by the impulse, even if only one item is in the focus of attention. It is worth noting, however, that the decodability of the attended item was significantly higher than that of the unattended item (average: $p = 0.031$), consistent with the behavioural evidence for relatively better memory for the initially prioritised item.

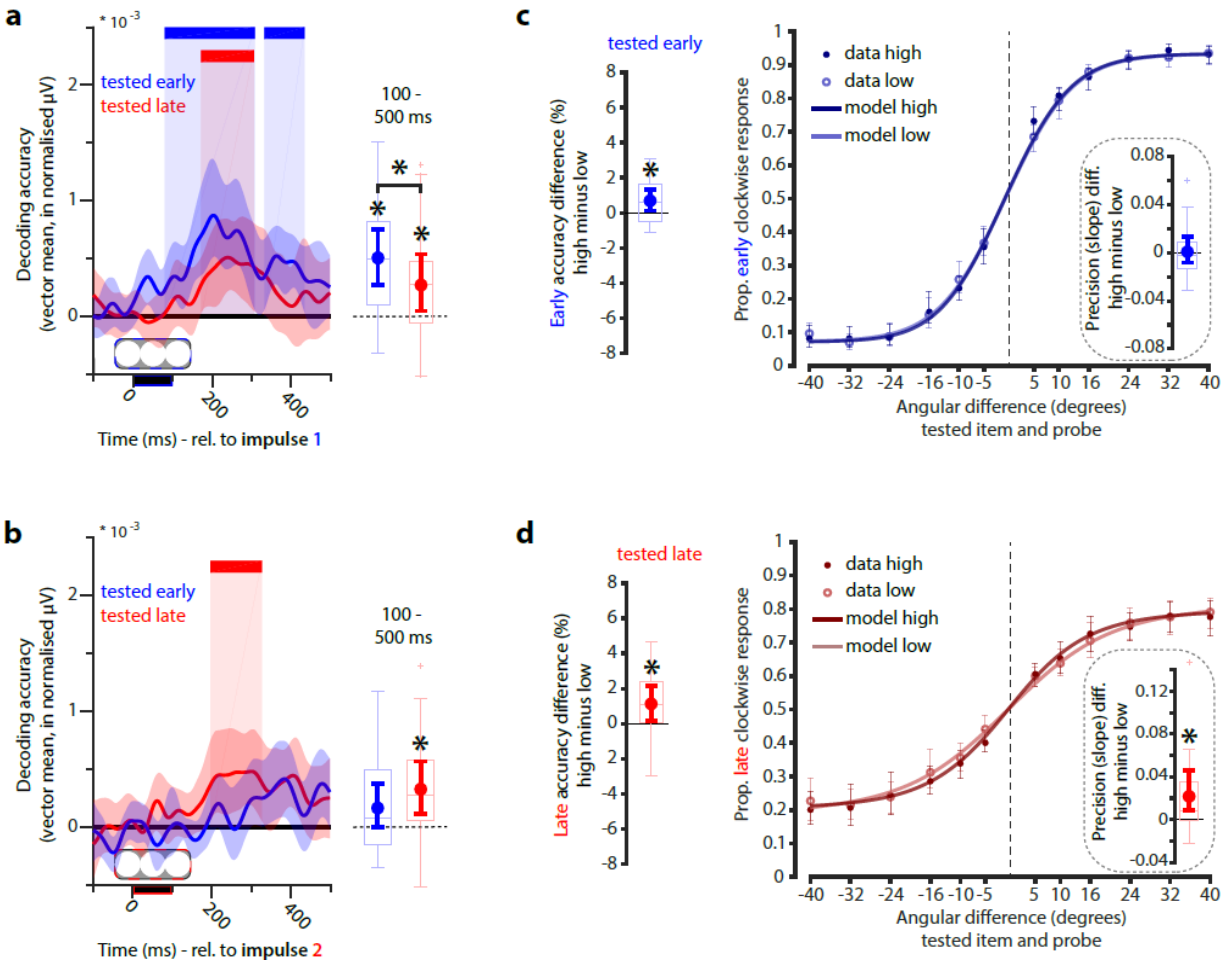


Figure 6. Attended and unattended WM items in early and late epochs and relationship to behavioural performance. a. Item decoding of the early (blue) and late tested item (red) during the first impulse epoch. Coloured bars on top indicate significant decoding clusters of the corresponding items (permutation test, $n = 19$, cluster-defining threshold $p < 0.05$, corrected significance level $p < 0.05$). Error shading is 95% C.I. of the mean. Boxplots and superimposed circles with error bars (mean and 95% C.I. of the mean) represent average decodability from 100 ms after stimulus onset until the end of the epoch. Significant average decoding and average difference between the decodability of the early and late item are marked by an asterisk (permutation test, $n = 19$, $p < 0.05$). **b.** Item decoding during the second impulse epoch, same conventions as **a. c**. Boxplot and superimposed circles and error-bars represent the difference in overall WM task performance between high and low early-tested item decoding trials during the first impulse (left). Proportion of clockwise responses for high and low decoding trials as a function of the angular difference between the memory item and the probe (right). Inset shows

the boxplot and error-bar of the difference in the slope parameter (a measure of memory precision) between high and low decoding trials. **d.** The same convention as in a. but for the decoding of the late-tested item during the late impulse. Significant differences in accuracy/precision between high and low decoding trials are highlighted by asterisks (permutation test, $n = 19$, $p < 0.05$, two-sided and one-sided for accuracy and precision tests, respectively). Data points outside of 1.5 * interquartile range are shown separately in the boxplots (small crosses).

We found no evidence for a relationship between trial-wise differences in alpha lateralization and WM item decodability of the impulse response for either the attended or unattended item (Supplementary Fig. 3). This further suggests that the item-specific impulse response does not even vary with trialwise differences in the focus of attention.

We also found that the remaining relevant and initially unattended item could also be decoded in the second impulse response (cluster: 196 to 326 ms, $p = 0.016$, corrected; average: $p = 0.012$), while decoding the initially prioritised item failed to reach significance in this epoch (clusters: $p > 0.109$, corrected; average: $p = 0.112$; Fig 6b). The now-deprioritised item was presumably cleared from the hidden state because it was no longer relevant, similar to forgetting observed after the retro-cue from Experiment 1.

Again, we also tested for cross-generalization between the decodable patterns of the memory items epoch (Fig. 5a) and the impulse-epochs (Fig. 6a, b). However, like in Experiment 1, we found no evidence that the impulse literally ‘reactivates’ activity patterns associated with initial encoding for either item (all corrected clusters: $p > 0.32$).

There was also a positive relationship between trial-wise decoding of the attended items at the first and at the second impulse with WM performance (early: $p = 0.038$, Fig. 6c; late: $p = 0.04$; Fig. 6d), replicating and extending the findings of Experiment 1. As in Experiment 1, we modelled the behavioural profile to test if the positive relationship between decoding and task performance is due to an increase in precision and/or a decrease in guess-rate. While the modelling results were inconclusive for the early-tested item (precision: $p = 0.399$, one-tailed;

guess-rate: $p = 0.329$, one-tailed; Fig. 6c), there was evidence for an effect in precision of working memory for the late item (precision: $p = 0.006$, one-tailed; guess-rate: $p = 0.942$, one-tailed; Fig. 6d), replicating the precision effect of Experiment 1. Note that there was again no relationship between accuracy and item decoding during the encoding phase (Supplementary Fig. 2b).

Experiment 3

We developed the impulse perturbation approach to reveal otherwise hidden neural states, without necessarily transforming the mnemonic representation [4,21]. This contrasts with other studies using retro-cues [10,11,32] or TMS [36] to 'reactivate' a latent item in working memory. However, to test whether our impulse stimulus actually did result in a behaviourally relevant transformation of the memory item (i.e., from a functionally latent to active state), we conducted an additional behavioural experiment (n = 20). Adapting the design of Experiment 1, we now varied the presentation of the stimulus-onset asynchrony (SOA) between impulse and probe onset in Experiment 3 (SOA from 0 to 500ms; see Supplementary Fig. 4a). If the increase in impulse-specific decodability observed in both EEG experiments reflects a functional "reactivation" of an otherwise latent memory item, there should be a corresponding benefit to behaviour.

A repeated measures ANOVA provided no evidence for an effect of SOA ($F(4, 76) = 1.184$, $p = 0.325$). Uncorrected paired comparisons between the no-impulse condition (SOA 0 ms) and all other SOAs provided no evidence for an impulse-specific effect on accuracy for any SOA either (permutation test, $n = 20$, all $p > 0.12$; Supplementary Fig. 4b). This suggests that our impulse stimulus is effective for 'pinging' activity silent neural states, without resulting in any behaviourally relevant transformation of the mnemonic representation.

Discussion

Recent theoretical models of WM predict a key role for activity-silent neural states in maintaining item-specific information [4,17,18]. This raises a particular challenge for contemporary neuroscience that is dominated by measurement and analysis of neural activation states. Here, we address this challenge using a perturbation approach to reveal hidden neural states that code the contents of WM. We show that the response to an impulse stimulus faithfully reflects item-specific information in WM. We further demonstrate that the impulse response reflects both attended and unattended items in WM, yet recently forgotten information leaves no detectable trace in the hidden state. Behavioural modelling further suggests that the hidden-state coding determines the quality of information in WM.

Previous evidence from non-human primates showed that a neutral visual stimulus presented during the WM delay period can elicit distinct patterns of neural activity that depend on recent visual input [37]. Although the previous work could not deconfound previous sensory stimulation and WM proper, the observed effect helped motivate a dynamic coding model for WM [4]. According to this framework, distinct memoranda are associated with distinct changes in neural response profile, which would be readable to downstream systems from the state-dependent response to a retrieval probe [4,18]. Crucially, WM depends on the maintenance of the item-specific neural response profile, rather than an explicit representation of an item in a persistent activity state. We now provide direct evidence for a WM-dependent impulse response decoupled from previous stimulation history, and further demonstrate that this WM state is highly flexible and coupled to behavioural performance. The hidden state for a specific item can be rapidly cleared if it is no longer relevant to the task, providing a striking neural correlate of directed forgetting in WM.

Recent retro-cuing evidence suggests that prioritising one WM item relative to other task-relevant items improves neural decoding of the cued item, whereas decoding of unattended items drops to chance levels even though the unattended information is still ultimately task relevant and retrievable at the end of the trial [10]. Item-specific delay activity therefore seems to reflect the focus of attention, rather than WM per se[32]. The impulse response reported here clearly differs from the typical profile observed for decoding delay activity patterns. In Experiment 2, both attended and unattended items could be decoded from the impulse response of the hidden state as

long as they are both still ultimately required for task performance. This suggests that if the information is successfully maintained in WM, there is a corresponding trace in the hidden state, irrespective of attentional priority. These results highlight the flexibility of WM, independently of switching attention between specific items in WM. Activity states appear to track the focus of attention [10,11,32], whereas hidden states, as revealed by the impulse response, more closely track the actual contents of WM.

Exactly how the proposed hidden state can be used for WM-guided behaviour remains an important open question. Computationally, supervised learning could determine the mapping between the memory-dependent probe response and the correct behavioural response [38], however such a learning strategy seems implausible for real-world behaviour. Trial and error learning of arbitrary patterns does not seem a realistic model for WM, at least for humans. Instead, the inherent dynamics could establish a history-dependent match filter [20], which would be capable of transforming probe input to a common decision signal (i.e., match/no-match, or in our case clockwise/counter-clockwise). In Myers et al. [27], such a mechanism was shown to generate two distinct decision-related signals in an orientation detection task: a signed (i.e., directional) and unsigned difference signal, even though the signed difference was actually irrelevant to behaviour in that task. A similar process could underpin WM encoding in hidden states. The hidden state could establish a flexible, task dependent circuit for WM-dependent decision-making [39]. When the probe stimulus is presented, the hidden state transforms the input to decision-relevant output: e.g., direction of angular rotation. However, because the impulse stimulus used in these experiments does not contain decision-relevant features, the impulse response reflects an input-output transformation of the arbitrary input.

It may be noted that although the response to an arbitrary input is sufficient to ‘read-out’ the hidden state, it is unlikely to constitute an explicit ‘reactivation’ of the memory representation. In contrast, retro-cueing can convert an unattended item to a prioritised state in preparation for the recall [22]. Similarly, a recent transcranial magnetic stimulation study suggests that stimulation of the visual cortex can also render an item active from its latent state [36]. We find no evidence that our impulse stimulus reactivates the same pattern associated with stimulus processing. Moreover, a further behavioural experiment designed to test the possible behavioural consequences of our impulse stimulus provides no evidence that it interacts with the mnemonic representation. Rather, we argue that the impulse response simply ‘echoes’ the representational structure of the

hidden state, but does not drive an explicit transformation of latent memories to a prioritized state.

It has long been assumed that WM maintenance depends on persistent neural activity [2]. Instead, we propose that activity-silent neural states are sufficient to bridge memory delays. Activity-dependent transformations in hidden states determine the temporary coding properties of memory networks: i.e., dynamic coding [4,37]. WM decisions are made by the state-dependent response to subsequent input. However, WM is also classically associated with active manipulation of content in short-term memory [1]. We argue that such transformations are activity dependent, but the results of the transformation can be maintained in short term memory via latent network states. This alternative account does not ignore previous evidence for decodable activity during mnemonic delays, but rather attributes such evidence to focused attention [36], periodic [18] or stochastic [17] updating, and/or response preparation [8]. Interestingly, our current results also show that cue-directed forgetting can rapidly wipe the mnemonic representation from the hidden state. Rapid construction and dissolution of hidden states places important constraints on the basic mechanisms of hidden-state coding.

Although the present study addressed a specific model of WM, it is worth noting that the general impulse response approach for inferring otherwise silent neural states could also be particularly fruitful for exploring other tonic cognitive states, such as task set, attention and expectation. It is becoming increasingly apparent that we need to look beyond simple measures of neural activity, and consider a richer diversity of neural states that underpin context-dependent behaviour. Here we focus on perturbation to illuminate hidden states, but future work will also profit from more direct measures of functionally relevant hidden states (e.g., synaptic efficacy, membrane potentials, extra-cellular transmitter concentrations). This will require more sophisticated measurements in awake behaving animals, coupled with non-invasive approaches like described here for human studies.

Acknowledgements

We would like to thank the Biotechnology & Biological Sciences Research Council (BB/M010732/1 to M.G.S.), and the National Institute for Health Research Oxford Biomedical Research Centre Programme based at the Oxford University Hospitals Trust, Oxford University. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health. We would also like to thank E. Spaak, A. Cravo, and N. Myers for helpful comments, and all our volunteers for their participation.

Author Contributions

M.J.W., M.G.S., and E.G.A. designed the study. M.J.W. and J.J. collected the data. M.J.W. analysed the data. M.J.W., M.G.S., E.G.A., and J.J. wrote the manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

1. Baddeley, A. Working memory: looking back and looking forward. *Nat. Rev. Neurosci.* **4,** 829–839 (2003).
2. Curtis, C. E. & D'Esposito, M. Persistent activity in the prefrontal cortex during working memory. *Trends Cogn. Sci.* **7,** 415–423 (2003).
3. Goldman-Rakic, P. Cellular basis of working memory. *Neuron* **14,** 477–485 (1995).
4. Stokes, M. G. 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn. Sci.* **19,** 394–405 (2015).
5. Watanabe, K. & Funahashi, S. Neural mechanisms of dual-task interference and cognitive capacity limitation in the prefrontal cortex. *Nat. Neurosci.* **17,** 601–611 (2014).
6. Watanabe, K. & Funahashi, S. Prefrontal Delay-Period Activity Reflects the Decision Process of a Saccade Direction during a Free-Choice ODR Task. *Cereb. Cortex* **17,** i88–i100 (2007).
7. Miller, E. K., Erickson, C. A. & Desimone, R. Neural Mechanisms of Visual Working Memory in Prefrontal Cortex of the Macaque. *J. Neurosci.* **16,** 5154–5167 (1996).
8. Barak, O., Tsodyks, M. & Romo, R. Neuronal Population Coding of Parametric Working Memory. *J. Neurosci.* **30,** 9424–9430 (2010).
9. LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K. & Postle, B. R. Decoding Attended Information in Short-term Memory: An EEG Study. *J. Cogn. Neurosci.* **25,** 127–142 (2012).
10. Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K. & Postle, B. R. Neural Evidence for a Distinction between Short-term Memory and the Focus of Attention. *J. Cogn. Neurosci.* **24,** 61–79 (2011).

11. Sprague, T. C., Ester, E. F. & Serences, J. T. Restoring Latent Visual Working Memory Representations in Human Cortex. *Neuron* **91,** 694–707 (2016).
12. Sreenivasan, K. K., Curtis, C. E. & D'Esposito, M. Revisiting the role of persistent neural activity during working memory. *Trends Cogn. Sci.* **18,** 82–89 (2014).
13. Buonomano, D. V. & Maass, W. State-dependent computations: spatiotemporal processing in cortical networks. *Nat. Rev. Neurosci.* **10,** 113–125 (2009).
14. Barak, O. & Tsodyks, M. Working models of working memory. *Curr. Opin. Neurobiol.* **25,** 20–24 (2014).
15. Murray, J. D. *et al.* Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc. Natl. Acad. Sci.* **114,** 394–399 (2017).
16. Fujisawa, S., Amarasingham, A., Harrison, M. T. & Buzsáki, G. Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. *Nat. Neurosci.* **11,** 823–833 (2008).
17. Lundqvist, M. *et al.* Gamma and Beta Bursts Underlie Working Memory. *Neuron* **90,** 152–164 (2016).
18. Mongillo, G., Barak, O. & Tsodyks, M. Synaptic Theory of Working Memory. *Science* **319,** 1543–1546 (2008).
19. Hempel, C. M., Hartman, K. H., Wang, X.-J., Turrigiano, G. G. & Nelson, S. B. Multiple Forms of Short-Term Plasticity at Excitatory Synapses in Rat Medial Prefrontal Cortex. *J. Neurophysiol.* **83,** 3031–3041 (2000).
20. Sugase-Miyamoto, Y., Liu, Z., Wiener, M. C., Optican, L. M. & Richmond, B. J. Short-Term Memory Trace in Rapidly Adapting Synapses of Inferior Temporal Cortex. *PLOS Comput Biol* **4,** e1000073 (2008).

21. Wolff, M. J., Ding, J., Myers, N. E. & Stokes, M. G. Revealing hidden states in visual working memory using electroencephalography. *Front. Syst. Neurosci.* **9,** (2015).
22. Griffin, I. C. & Nobre, A. C. Orienting Attention to Locations in Internal Representations. *J. Cogn. Neurosci.* **15,** 1176–1194 (2003).
23. Landman, R., Spekreijse, H. & Lamme, V. A. F. Large capacity storage of integrated objects before change blindness. *Vision Res.* **43,** 149–164 (2003).
24. Harrison, S. A. & Tong, F. Decoding reveals the contents of visual working memory in early visual areas. *Nature* **458,** 632–635 (2009).
25. Worden, M. S., Foxe, J. J., Wang, N. & Simpson, G. V. Anticipatory biasing of visuospatial attention indexed by retinotopically specific alpha-band electroencephalography increases over occipital cortex. *J. Neurosci. Off. J. Soc. Neurosci.* **20,** (2000).
26. Saproo, S. & Serences, J. T. Spatial Attention Improves the Quality of Population Codes in Human Visual Cortex. *J. Neurophysiol.* **104,** 885–895 (2010).
27. Myers, N. E. *et al.* Testing sensory evidence against mnemonic templates. *eLife* **4,** e09000 (2015).
28. Zhang, W. & Luck, S. J. Discrete fixed-resolution representations in visual working memory. *Nature* **453,** 233–235 (2008).
29. Bays, P. M. & Husain, M. Dynamic Shifts of Limited Working Memory Resources in Human Vision. *Science* **321,** 851–854 (2008).
30. Murray, A. M., Nobre, A. C. & Stokes, M. G. Markers of preparatory attention predict visual short-term memory performance. *Neuropsychologia* **49,** 1458–1465 (2011).
31. Prins, N. & Kingdom, F. A. A. Palamedes: Matlab routines for analyzing psychophysical data. *http://www.palamedestoolbox.org* (2009).

32. Larocque, J. J., Lewis-Peacock, J. A. & Postle, B. R. Multiple neural states of representation in short-term memory? It's a matter of attention. *Front. Hum. Neurosci.* **8,** 5 (2014).
33. Olivers, C. N. L., Peters, J., Houtkamp, R. & Roelfsema, P. R. Different states in visual working memory: when it guides attention and when it does not. *Trends Cogn. Sci.* (2011). doi:10.1016/j.tics.2011.05.004
34. 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.* 1–22 (2016). doi:10.3758/s13414-016-1108-5
35. Ede, F. van, Niklaus, M. & Nobre, A. C. Temporal expectations guide dynamic prioritization in visual working memory through attenuated alpha oscillations. *J. Neurosci.* 2272–16 (2016). doi:10.1523/JNEUROSCI.2272-16.2016
36. Rose, N. S. *et al.* Reactivation of latent working memories with transcranial magnetic stimulation. *Science* **354,** 1136–1139 (2016).
37. Stokes, M. G. *et al.* Dynamic Coding for Cognitive Control in Prefrontal Cortex. *Neuron* **78,** 364–375 (2013).
38. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503,** 78–84 (2013).
39. Martínez-García, M., Rolls, E. T., Deco, G. & Romo, R. Neural and computational mechanisms of postponed decisions. *Proc. Natl. Acad. Sci.* **108,** 11626–11631 (2011).

Online Methods

Participants

Thirty healthy adults (13 female, mean age 24.9 years, range 18-38 years) were included in the analyses of Experiment 1, 19 (10 female, mean age 24.7 years, range 18-39 years) in Experiment 2, and 20 in Experiment 3 (13 female, mean age 21, range 18-29 years). During data collection and preprocessing, 4 additional participants of Experiment 1, 1 additional participant of Experiment 2, and 6 additional participants of Experiment 3 were excluded from all analyses due to either low average performance on the memory task (below 60% accuracy) or excessive eye-movements (more than 30% of trials contaminated). No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those reported in previous publications [21,28]. All participants of Experiment 1 and 2 received monetary compensation of £10/h, and participation in Experiment 3 contributed to course credits. All participants gave written informed consent. Experiments 1 and 2 were approved by the Central University Research Ethics Committee of the University of Oxford and Experiment 3 was approved by the Departmental Ethical Committee of the University of Groningen.

Apparatus and Stimuli

The experimental stimuli were generated and controlled by Psychtoolbox [40], a freely available MATLAB extension. The stimuli were presented on a 23" screen running at 100 Hz and a resolution of 1920 by 1080 in Experiment 1, on a 22" screen at a resolution of 1680 by 1050 in Experiment 2, and on a 19 inch CRT screen running at 100 Hz and a resolution of 1280 by 1024 in Experiment 3. Viewing distance was set at 64 cm in Experiment 1, 67.5 cm in Experiment 2 and approximately 60 cm (not controlled) in Experiment 3, to ensure that the visual angles of stimuli were the same across experiments even though the screen parameters were different. A standard keyboard was used for response input by the participants.

All reported stimuli were the same in all experiments, unless explicitly mentioned otherwise. A grey background (RGB = 128, 128, 128; 20.5 cd/m^2; 28.6 cd/m^2 in Experiment 3) was maintained throughout the experiments. A black fixation dot with a white outline (0.242°) was presented in the centre of the screen throughout all trials. Memory items and memory were sine-wave gratings presented at 20% contrast, with a diameter of 6.69° and spatial frequency of 0.65 cycles per degree. The phase was randomized within and across trials. The memory items were

presented at 6.69° eccentricity and for each trial the orientations were randomly selected without replacement from a uniform distribution of orientations. The impulse stimulus was 3 adjacent 'bullseyes' in Experiment 1. Each 'bullseye' was of the same size and spatial frequency as the memory items. To reduce strain on the eyes, and to minimise forward masking in Experiment 3, the impulse stimulus in Experiments 2 and 3 consisted of 3 adjacent white circles. In Experiment 1 and 2 the probes had the same contrast and spatial frequency as the memory items, and was presented in the centre of the screen. In Experiment 3 the probe screen included a high contrast black and white square-wave grating in the centre and two white lateralized circles on the outside (the same location and size as the preceding lateral impulse circles). The angle differences between a memory item and the corresponding memory probe were uniformly distributed across 7 angle differences in Experiment 1 (±3°, ±7°, ±12°, ±18°, ±25°, ±33°, ±42°), 6 angle differences in Experiment 2 (±5°, ±10°, ±16° ±24°, ±26°, ±32°, ±40°) and a single angle difference (±16°) in Experiment 3.

Procedure

Experiment 1: Participants completed a retro-cue visual working memory task. Each trial began with the onset of a fixation dot at the centre of the screen. After 1000 ms, the memory item array was shown for 250 ms, consisting of two randomly oriented low-contrast gratings left and right of fixation. After a delay of 800 ms an arrow was shown for 200 ms in the centre of the screen, pointing either to the left or to the right, and thus cueing which of the two previously presented items would be tested. The number of left and right cued trials was equal and the order was randomized for each participant. The impulse stimulus was presented for 100 ms, 900 ms after the offset of the retro-cue. After another delay of 400 ms, the memory probe was shown for 250 ms. Participants were instructed to indicate if the orientation of the probe relative to the orientation of the memory item was rotated clockwise by pressing the "m" key with the right index finger, or counter-clockwise by pressing the "c" key with the left index finger. A high or low frequency feedback tone was played after response, indicating if the answer was correct or incorrect, respectively. The next trial started within 400 to 700 ms (determined randomly). Participants completed 1344 trials in total, which lasted approximately 3 hours (including breaks). Trial conditions were randomized across the whole session. See Figure 1a for a trial schematic.

Experiment 2: Participants completed a visual working memory task where two items were serially tested. The experiment began by instructing the participant which of the two memory items would be tested early, and which one would be tested late. This rule never changed within a session. Each trial began with the onset of a fixation dot at the centre of the screen. After 1000 ms, the memory item array was shown for 250 ms, consisting of two randomly oriented low-contrast gratings left and right from fixation. After a delay of 950 ms, the first impulse was presented for 100 ms. After a delay of 500 ms, the first memory probe was presented for 250 ms, probing the first item. The response input was the same as in Experiment 1. After a fixed delay of 1750 ms after the offset of the first probe, the second impulse was shown for 100 ms. Following a delay of 400 ms, the second memory probe was presented for 250 ms, probing the late-tested item. After the second response, two feedback tones were played, one for each response, separately indicating whether the first and second answers were correct. Participants completed two sessions of the task on two separate days, separated by approximately 1-2 weeks. The testing order of the memory items was fixed within each sessions, and switched between sessions (i.e. left item tested first in one session, right item tested first in the other session). The order of the testing rule between sessions, (whether the left item would be tested first in the first or in the second session) was counterbalanced across participants (odd numbered left first, even numbered right first). Each session consisted of 864 trials, and lasted approximately 3 hours including breaks. See Figure 3a for a trial schematic.

Experiment 3: The task was almost the same as Experiment 1, including the same timings of the memory items, cue, probe and overall trial duration. The one key difference was the timing of the impulse stimulus. While the delay between cue offset and probe onset was held constant at 1,400 ms across all trials (the same as in Experiment 1), the SOA between impulse and probe onset was 0, 50, 100, 250 or 500 ms (determined pseudo randomly across the session). No impulse was shown in the 0 ms SOA condition. The impulse remained on the screen until the probe stimulus was presented. This was to ensure the least possible interference of the impulse on probe processing (i.e., rapid onset and offset of the white circles immediately before probe presentation could deteriorate probe visibility), as well as keeping the different SOA conditions as similar as possible (longer SOA would include an additional offset). Participants completed 280 trials (approximately 30 minutes). See Supplementary Fig. 4a for a trial schematic.

Data collection and analyses were not performed blind to the conditions of the experiments.

Due to the within-subject design in all three experiments, randomization of conditions between subjects was not applicable.

EEG Acquisition

The EEG signal was acquired from 61 Ag/AgCl sintered electrodes (EasyCap, Herrsching, Germany) laid out according to the extended international 10-20 system. Data was recorded at 1000 Hz using NeuroScan SynAmps RT amplifier and Scan 4.5 software in Experiment 1 and Curry 7 software in Experiment 2 (Compumedics NeuroScan, Charlotte, NC). The anterior midline frontal electrodes (AFz) served as the ground. Bipolar electrooculography (EOG) was recorded from electrodes placed above and below the right eye, and from electrodes placed to the left of the left eye and to the right of the right eye. The impedances of all electrodes were kept below 5 kΩ. Online, the EEG was referenced to the right mastoid and filtered using a 200 Hz low-pass filter.

EEG pre-processing

Offline, the data was re-referenced to the average of both mastoids, down-sampled to 500 Hz and band-pass filtered (0.1 Hz high-pass and 40 Hz low-pass) using EEGLAB [41]. The data was then epoched to the onset of the memory items and the impulse. In Experiment 1, the memory item epoch was from -200 ms to 1050 ms, relative to onset, and in Experiment 2 from -200 ms to 1200 ms. The impulse epochs were from -200 ms to 500 ms relative to onset in both experiments. Additionally, for the purpose of artefact rejection, which included the rejection of trials containing saccadic eye-movement prior to the time of interest (see below), the cue segment in Experiment 1 was also epoched (-200 ms to 1100 ms).

Subsequent artefact detection and trial rejection focused exclusively on the 17 posterior channels that were included in the analyses (P7, P5, P3, P1, Pz, P4, P6, P8, PO7, PO3, POz, PO4, PO8, O1, Oz, O2) and the EOGs. Each trial of each epoch was individually visually inspected for blinks, saccades and non-stereotyped artefacts. Trials from individual epochs were rejected from analyses involving that epoch if it contained any of the above-mentioned artefacts. Furthermore, impulse-epoch trials were also excluded from corresponding analyses if the EOG signal suggested that saccades occurred during any of the previous epochs of that trial. In Experiment 1 this exclusion procedure was applied to the cue-epoch as well. In Experiment 2, late impulse trials were also excluded if no response was registered for the preceding probe. For the decoding

analyses, each epoch was baselined using the average signal from -200 ms to 0 ms before stimulus onset. The multivariate data were also demeaned at each time-point by subtracting the average voltage for all posterior channels included in the analyses.

Time-frequency decomposition and lateralization analysis

In order to explore alpha power (8-12 Hz) lateralization [25,42], the spectral power from 6 to 16 Hz (in steps of 0.5 Hz) of the EEG signal was computed using Hanning tapers with time-windows of 5 cycles per frequency (in steps of 10 ms) using the MATLAB toolbox FieldTrip [43]. We included the whole experimental trial, ranging from 1000 ms before memory item onset until 1500 ms after (second) probe onset (-1000 to 4150 ms relative to memory items in Experiment 1, and -1000 to 5800 ms relative to memory items in Experiment 2). The power was log transformed, and lateralization was computed by subtracting the average power of the ipsilateral posterior electrodes from the average power of the contralateral posterior electrodes in relation to the cued memory item in Experiment 1 and to the early-tested item in Experiment 2 (P7, P5, P3, P1, PO7, PO3, O1 versus P8, P5, P6, P4, P2, PO8, PO4, O2).

Significant clusters of lateralization were determined using a cluster-corrected non-parametric sign-permutation test [44]. In both experiments, the whole trial was included in this analysis (-100 to 3150 ms relative to memory items onset in Experiment 1, and -100 to 4800 ms in Experiment 2).

Orientation decoding

To test whether the activity pattern of the posterior EEG channels of interest contained orientation-specific activity, we used the Mahalanobis distance [45] to compute the trial-wise distances between the full range of possible orientations, and quantify to what extent the computed distances adhere to the parametric circular space of the orientations [11]. This approach is an extension of the pairwise distance approach we used before [21] and is conceptually similar to the population tuning curve model [26].

The left and right presented items were decoded separately and independently within each participant and experimental session. All 17 posterior channels (see above) were used for all decoding analyses. The procedure followed a leave-one-trial-out cross-validation approach to compute the trial-wise decodability of the orientation of interest. The activity pattern of a single

test-trial at a particular time-point was compared to the pattern of all other trials at the same time-point. These were averaged into 12 orientation bins relative to the orientation of the test-trial, each containing trials with orientations within a range of 30° and centred around -75°, -60°, -45°, -30°, -15°, 0°, 15°, 30°, 45°, 60°, 75°, and 90°. The Mahalanobis distances between the test-trial and each orientation bin was computed using the covariance estimated from all trials excluding the test-trial using a shrinkage estimator[46]. To simplify visualization and interpretation, the 12 resulting distances were mean centred and the sign was reversed, resulting in a visual representation of a tuning curve. Higher values correspond to greater relative similarity between the test-trial and the averaged train-trials within a particular orientation bin, and lower values correspond to greater dissimilarity.

Next, the vector means of the tuning curves were computed [11]. First, the cosine of the centre of each orientation bin (θ) was rescaled to the range -180 to 180. It was then multiplied with the corresponding sign-reversed distances (d(θ)) before the mean of the resulting 12 values was taken, which made up the decoding accuracy (da).

Equation 1: $da = mean(d\ (\theta) \cos\ (2\ \theta))$

A high value reflects evidence for orientation tuning: the difference between the test-trial and train-trials with a similar orientation is smaller than between the test-trial and train-trials with different orientations. This procedure was repeated for all trials and all time-points. See Supplementary Information for the custom Matlab function used to decode orientations using Mahalanobis distance.

The decoding values were averaged over all trials, and smoothed over time with a Gaussian smoothing kernel (*SD* = 16 ms) for visualization and time-resolved significance testing.

Cluster-corrected sign-permutation significance tests were carried out within the memory items epoch (0 to 1050 ms in Experiment 1, 0 to 1200 ms in Experiment 2) and impulse epochs separately (0 to 500 ms in both experiments), in order to explore the significant decoding time-course. Additionally, to assess the overall decodability within an epoch, the decoding values were averaged over time (from 100 ms after stimulus onset until the end of the epoch) and then submitted to a two-sided permutation test.

<u>Relationship between behaviour and decoding</u>

The trial-wise average decoding scores after memory items presentation (100 to 1050 ms in Experiment 1, and 100 to 1200 ms in Experiment 2) and impulse presentation (100 ms to 500 ms) was median split. Non-response trials (to the early probe in Experiment 2) were excluded from this analysis. The average behavioural accuracies of high and low decoding trials were statistically compared using a two-sided permutation test.

Behavioural modelling

To further explore the relationship between WM task performance and trial-wise decoding, we modelled the behavioural performance as a function the difference in degrees between the orientation of the memory item and the probe using the following model that was fit to each participant separately [30].

Equation 2: $y = \lambda + \frac{(1-2\lambda)}{2} \times erfc(\frac{-\beta}{\sqrt{2}}(x-\alpha))$

where *erfc* is the complementary Gaussian error function, λ is the asymptote, β is the slope and α is the threshold/bias parameter. The modelling fitting was performed using the Palamedes Matlab toolbox [31]. The asymptote represents the guess rate, where a higher value reflects a higher probability that no information about the probed item is maintained in WM, resulting in a higher probability for mistakes even when the angular difference between the probe and the memory item is large. The slope is interpreted as the memory precision, where a high precision reflects a relatively high proportion of correct responses at small degree rotations between the probe and memory item. The asymptote and slope parameters were both unconstrained across the high and low decoding conditions. A single bias parameter was used, which was included (instead of fixing it at 0) because cumulative-likelihood tests [47] showed better model fits for all cases (Experiment 1: $n = 30$, $\chi^2(30) = 135.978$, $p < 0.001$; Experiment 2, $n = 19$, early accuracy: $\chi^2(19) = 215.351$, $p < 0.001$; late accuracy: $\chi^2(19) = 33.69$, $p = 0.02$).

The unconstrained model parameters (slope and asymptote) were subsequently compared between high and low decoding trials. Since the behavioural modelling was carried out as a direct follow up to the average accuracy effects observed in both experiments (two-sided tests), we had clear expectations about the directionality of the effects. For the positive relationship between decoding and accuracy observed for the cued item in Experiment 1 and both tests in Experiment 2, we expected that decoding should have a negative relationship with guess-rate

(i.e., lower guess-rate for high decoding) and/or a positive relationship with precision (higher precision for higher decoding), and vice versa for the negative accuracy effect of the uncued item in Experiment 1. Therefore, all tests of model parameter comparisons between high and low decoding trials were one-sided.

Cross-temporal decoding

We also explored the cross-temporal dynamics of stimulus processing and maintenance as a function of item priority in Experiment 2, and cross-generalization between impulse and memory presentation epochs in both experiments. The decoding approach was the same as described above, except classifiers trained at each time point were tested at every other time point, resulting in 2-dimensional cross-temporal decoding matrices [48].

If the decoding patterns are stationary, it should not matter whether train/test is performed using the same time points. In contrast, decoding often appears dynamic: training and testing on the same time-points results in higher decoding scores than training and testing on different time-points (i.e., minimal cross-temporal generalization). We tested for this hallmark feature of dynamic coding using a non-parametric test used previously [27]. The decodability at each cross-temporal time-point $t_{x,y}$ was compared to the pair of decodabilities at the corresponding within time-points ($t_{x,x}$ and $t_{y,y}$) with two separate permutation tests. A significant difference in both was taken as evidence for dynamic coding. Time-points of significant dynamic coding were corrected for multiple comparisons using a two-dimensional cluster-based permutation test.

Significance testing

To determine statistical significance, we used the non-parametric sign-permutation test [44](with one exception, see ANOVA below), which does not make assumptions about the underlying distribution. Since the null hypotheses of all tests corresponded to no effect (i.e. no difference in power lateralization, no difference in decodability, etc.), the sign of the data of each participant was randomly flipped with a probability of 50% 50,000 times. The resulting distribution was used to derive at the *p*-value of the null-hypothesis that the mean effect was equal to 0. All tests were two-sided, unless otherwise stated.

For time-series and frequency data, the above procedure was repeated for each time-point and frequency (when applicable). To correct for multiple comparisons over time and/or frequencies,

a cluster-based permutation test was subsequently used using 50,000 permutations (5,000 for cross-temporal decoding, due to computer memory limitations), with a cluster-forming threshold and cluster significance threshold of $p < 0.05$. Tests concerning the average of specific time-windows (which includes decoding-behaviour relationships) were performed to test unique and independent hypotheses, therefore no correction applied. The sample size for all tests in Experiment 1 was $n = 30$, $n = 19$ in Experiment 2, and $n = 20$ in Experiment 3.

The 95 % confidence intervals of the error-bars were determined by bootstrapping from the corresponding data 50,000 times.

The boxplots used in our figures follow the standard conventions. The middle line represents the median, the box the first and third quartile, and the whiskers all data within 1.5 * interquartile range of the lower and upper quartile. Where appropriate, data points outside this range are displayed individually (small crosses).

A repeated measures ANOVA was used to analyse the behavioural data of Experiment 3. The normality and equal variances assumptions were tested with the Shapiro-Wilk test of normality and Mauchly's test of sphericity, respectively. Neither test provided evidence for assumption violations of the data.

Data availability

The data that support the finding of this study are publically available at http://datasharedrive.blogspot.co.uk/2017/03/dynamic-hidden-states-underlying.html. All necessary task/condition information has been provided within a self-contained format, as specified in the OECD Principles and Guidelines for Access to Research Data from Public Funding [49].

Code availability

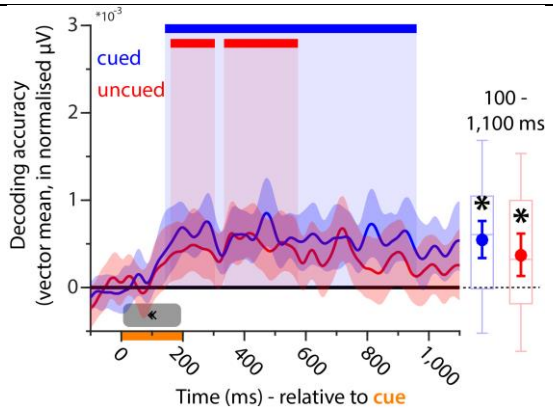
The custom Matlab orientation decoding function is provided with the paper (Supplementary Information). All complete custom Matlab routines used to generate the figures of this paper are available at http://datasharedrive.blogspot.co.uk/2017/03/dynamic-hidden-states-underlying.html

Methods References

40. Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis.* **10,** 433–436 (1997).

41. 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).
42. Schneider, D., Mertes, C. & Wascher, E. The time course of visuo-spatial working memory updating revealed by a retro-cuing paradigm. *Sci. Rep.* **6,** 21442 (2016).
43. Oostenveld, R. *et al.* FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data, FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Comput. Intell. Neurosci. Comput. Intell. Neurosci.* **2011, 2011,** e156869 (2010).
44. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164,** 177–190 (2007).
45. De Maesschalck, R., Jouan-Rimbaud, D. & Massart, D. L. The Mahalanobis distance. *Chemom. Intell. Lab. Syst.* **50,** 1–18 (2000).
46. Ledoit, O. & Wolf, M. Honey, I shrunk the sample covariance matrix. *J. Portf. Manag.* **30,** 110–119 (2004).
47. Claessens, P. M. E. & Wagemans, J. A Bayesian framework for cue integration in multistable grouping: Proximity, collinearity, and orientation priors in zigzag lattices. *J. Vis.* **8,** 33–33 (2008).
48. King, J.-R. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn. Sci.* **18,** 203–210 (2014).
49. Pilat, D. & Fukasaku, Y. OECD Principles and Guidelines for Access to Research Data from Public Funding. *Data Sci. J.* **6,** OD4-OD11 (2007).

Supplemental Information

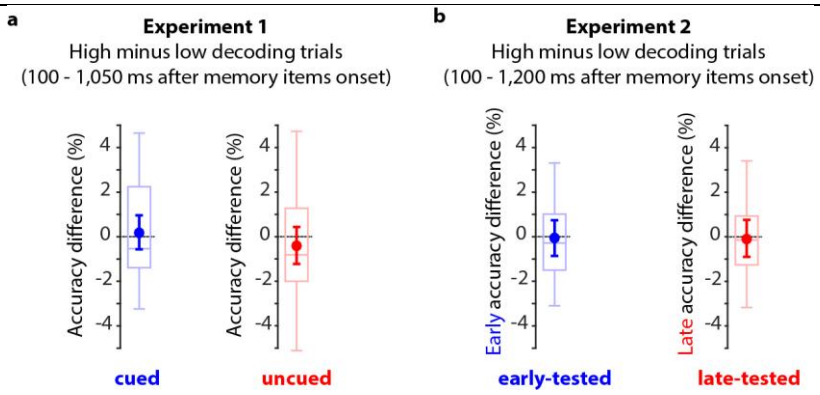


Supplementary Figure 1

Cue-specific item decoding time-course (Experiment 1).

The cue-specific neural response showed robust decoding for the cued (n = 30, cluster: 142 to 960 ms, p < 0.001, corrected; average: p < 0.001) and uncued item (n = 30, clusters: 158 to 304 ms, p = 0.035, corrected, 328 to 574 ms, p = 0.006, corrected, average: p = 0.006).

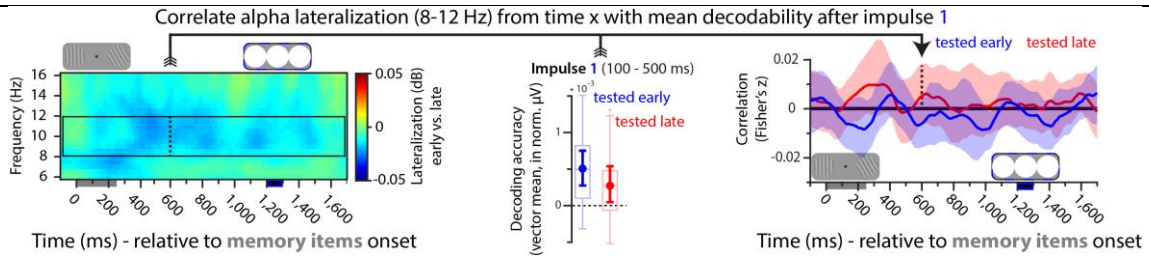
Error shading is the 95 % C.I. of the mean. The boxplots and superimposed circles with error-bars (mean and 95 % C.I. of the mean) represent average decoding from 100 to 1,100 ms after cue onset. Significant average decoding is marked by an asterisk (permutation test, n = 30, p < 0.05).



Supplementary Figure 2

Testing the relationship between item decoding during the memory item epoch and WM accuracy.

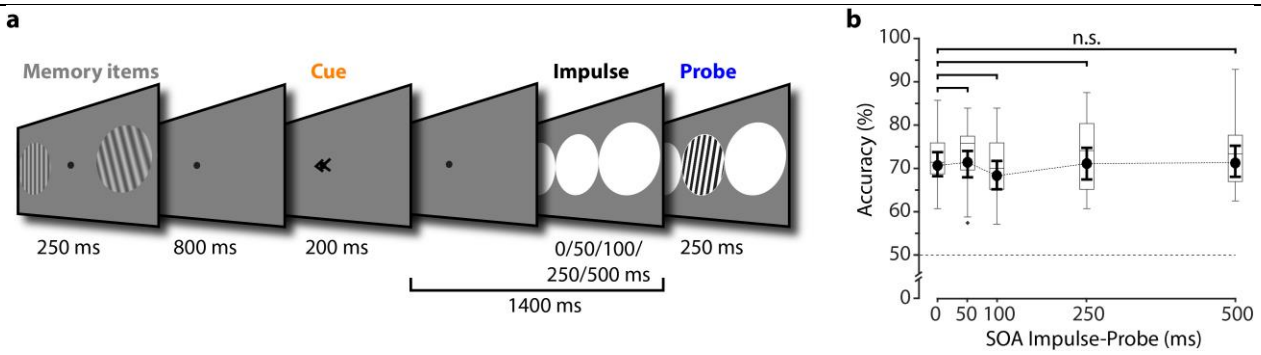
a. Task accuracy difference between high and low decoding trials of the cued (blue; n = 30, p = 0.673) and uncued (red; n = 30, p = 0.344) item during the memory items epoch (average decoding from 100 to 1,050 ms relative to memory items onset) in Experiment 1. **b.** Early accuracy difference between high and low decoding trials of the early-tested item (blue; n = 19, p = 0.865), and late accuracy difference between high and low decoding trials of the late-tested (red; n = 19, p = 0.978) during the memory items epoch (average decoding from 100 to 1,200 ms relative to memory items onset) in Experiment 2. Circles and error-bars superimposed on the boxplots represent mean and 95% C.I. of the mean.



Supplementary Figure 3

Testing the relationship between alpha-lateralization and item decoding after the first impulse in Experiment 2.

Both attended and unattended memory items were decodable after the first impulse in Experiment 2; however, it remains possible that participants sometimes attended to the less-relevant item, contributing to decoding on some trials. To consider this possibility, we test whether the impulse-specific WM item decoding after impulse 1 presentation covaries with trial-wise fluctuations in spatial attention. Spatial attention was indexed by alpha-power lateralization relative to the location of the early-tested item of each time-point (left, also see Figure 4c and corresponding results), and trialwise item decodability was estimated 100-500ms after impulse 1 onset (middle panel). The correlation time-course (right), where each time-point represents the mean correlation of the averaged item decoding (100 – 500 ms after impulse 1) with the alpha-lateralization of that time-point, shows no evidence for a relationship between item decoding and alpha-lateralization for any time-point (permutation test, n = 19, early-tested item: all $p > 0.058$; late-tested item: all $p > 0.148$, uncorrected). Therefore, we find no evidence that the impulse-response varies with the focus of attention, even on a trial-wise basis. Error shadings are 95% C.I. of the mean. Circles and error bars superimposed on the boxplots represent mean and 95% C.I. of the mean. Data points outside the 1.5 * interquartile range are marked as crosses in the boxplots.



Supplementary Figure 4

Task schematic and results of behavioural experiment.

a. Two memory items were presented, and participants were instructed to memorize both. A retro-cue indicated which item would be tested at the end of the current trial. The impulse stimulus was presented at varying delays (or not at all) and stayed on screen until the probe was presented. Participants indicated whether the probe was rotated clockwise or anti-clockwise relative to the orientation of the cued item. **b.** Behavioural performance as a function of impulse-probe SOA. None of the uncorrected paired comparisons between the no-impulse condition (SOA 0 ms) and the other SOA conditions reached significance (permutation test, n = 20). Circles and error bars superimposed on the boxplots represent mean and 95% C.I. of the mean. Data points outside the 1.5 * interquartile range are marked as crosses in the boxplots.