A hierarchy of functional states in working memory

Paul S. Muhle-Karbe[1,2], Nicholas E. Myers[1,2,3], Mark G. Stokes[1,2,3]

[1]Department of Experimental Psychology, University of Oxford, UK

[2]Oxford Centre for Human Brain Activity, University of Oxford, UK

[3]These authors contributed equally

Corresponding authors:

Paul S. Muhle-Karbe

University of Oxford, Department of Experimental Psychology

OX2 6GG, Oxford, UK

paul.muhle-karbe@psy.ox.ac.uk

Mark G. Stokes

University of Oxford, Department of Experimental Psychology

OX2 6GG, Oxford, UK

mark.stokes@psy.ox.ac.uk

Abstract

Extensive research has examined how information is maintained in working memory (WM), but it remains unknown how WM is used to guide behaviour. We addressed this question using a combination of electroencephalography, pattern analyses, and cognitive modelling with a task that required maintenance of two WM items and flexible priority shifts between them. This enabled us to discern neural states coding for immediately and prospectively task-relevant items, and to examine how these states contribute to WM-based decisions. We identified two qualitatively different neural states: a functionally latent state encoded both items, was unrelated to performance on the current trial, but predictive of performance accuracy over longer time scales. In contrast, a functionally active state encoded only the immediately task-relevant item, and closely tracked the quality of evidence integration on the current trial. These results delineate a hierarchy of functional states whereby latent memories supporting general maintenance are transformed into active decision-circuits to guide WM-based behaviour.

Introduction

Working memory (WM) refers to the ability to maintain and manipulate information that is no longer available in the environment. It provides a flexible mental workspace that scaffolds many higher cognitive functions such as planning, reasoning, or cognitive control.[1,2] As a consequence, a longstanding and central theme in cognitive neuroscience has been to delineate the neural mechanisms that underpin WM.[3-6] This research has greatly advanced our understanding of mechanisms that enable the brain to maintain WM contents over delay periods, e.g., via persistent activity,[7-9] neuronal oscillations,[10,11] or activity-silent brain sates.[12-14] At the same time, however, this work has so far neglected a key aspect of WM, namely how its contents are used to guide flexible behaviour.

Existing evidence suggests that the use of memories is distinct from the mere maintenance of information.[15] For instance, patients with frontal lobe damage sometimes exhibit a phenomenon termed goal neglect, wherein instructed task components are disregarded during performance, even though they have been understood and remembered.[16,17] Studies in healthy humans furthermore suggest that WM items that will guide the next decision are in a privileged state that biases perception toward memory-matching stimuli, while other concurrently held items do not affect ongoing cognition.[18] Likewise, in neuroimaging studies, immediately relevant WM items can be robustly decoded from brain activity patterns, while other concurrently held items typically exhibit little or no decodability.[19-23] Such findings have commonly been thought to reflect the effect of attention that prioritises a single item within WM by amplifying the corresponding neural patterns.[24-27]

We have recently proposed a more complex account of WM-guided behaviour that does not rely on attentional selection alone, but instead entails a reconfiguration of representational formats from a purely mnemonic state into an action-ready state that is tuned for efficient task-dependent decision making.[15,28] According to this view, WM representations should exhibit qualitatively different functional attributes depending on their momentary task relevance. Items that are immediately relevant for guiding behaviour should be encoded in a functionally *active* state that is optimised for efficient task-dependent readout. By contrast, items that are not immediately task-relevant, but will become relevant later on, should be held in a functionally *latent* state that enables precise maintenance, but does not modulate cognition until it becomes relevant for guiding behaviour.

We tested this view using electroencephalography (EEG) and a novel task that allowed us to directly compare these two putative functional states. Participants were asked to hold two items in WM for an extended period of time, though at any given moment only one item was immediately task-relevant (active state), while the other item was maintained for later use (latent state). We leveraged a combination of multivariate decoding analyses and cognitive modelling to recover representations of active and latent WM items, and to link variance in their decoding strength with variance in

performance. Consistent with our proposal, we find that active and latent WM items are encoded in largely dissociable neural patterns. The overlap between those patterns reflects a latent mnemonic space that predicts decision accuracy at longer time scales but is unrelated to trial-wise variance in performance. By contrast, the unique pattern component coding for active items closely tracks trial-wise variance in the quality of evidence integration for WM-based decisions, and also explains performance costs that arise after priority shifts between WM items.

Results

Behaviour

50 healthy adults participated in two experiments (20 participants in experiment 1 and 30 participants in experiment 2) while undergoing EEG recordings. Participants performed a WM-based decision-making task that required maintenance of two items for a block of 16 trials, and flexible priority shifts between those items on a trial-wise basis. At the beginning of each block, participants were shown two orientated bars in blue and yellow colour, which served as WM items. During the block, each trial started with the presentation of an auditory cue that signalled which of the two WM items should be used as a reference for a forthcoming perceptual decision (see Figure 1 for illustration). Accordingly, on any given trial, only one of the two WM items was task-relevant (active item), while the other item had to be maintained for use later in the block (latent item). The cue was followed by a 700ms delay period until a randomly oriented Gabor patch was presented as target. In experiment 1, the target was presented centrally on the screen, while it was presented peripherally in experiment 2 (see Materials and Methods for details). Participants were given a maximum time window of 4s to classify the target via right-hand or left-hand button press as being tilted clockwise or counter-clockwise relative to the active memory item. Importantly, participants saw the memory items only during the encoding phase at the block start and did not receive response feedback. Accordingly, to solve the task, they had to maintain precise representations of both WM items throughout the block.

Participants were able to perform the task well above chance (experiment 1 = 82.3%; experiment 2 = 82.3%). As shown in Figure 1, performance was strongly modulated by the angular distance between the target orientation and the orientation of the active WM item. Reaction time (RT) decreased with greater distances (experiment 1: $F_{19} = 62.49$, $p < 0.001$; experiment 2: $F_{29} = 64.03$, $p < 0.001$), while accuracy increased (experiment 1: $F_{19} = 546.23$, $p < 0.001$; experiment 2: $F_{29} = 1452.79$, $p < 0.001$). By contrast, the angular distance between the target and the latent WM item had no reliable impact on RT (experiment 1: $F_{19} = 2.49$, $p = 0.069$; experiment 2: $F_{29} =2.125$, $p = 0.103$), and only a modest effect on accuracy (experiment 1: $F_{19} = 3.12$, $p = 0.033$; experiment 2: $F_{29} = 4.18$, $p = 0.050$). In addition, we observed a reliable effect of priority shifts, whereby performance was slower (experiment 1: F_{19} =

25.94, $p < 0.001$; experiment 2: $F_{29} = 75.40$, $p < 0.001$) and more error prone (experiment 1: $F_{19} = 8.68$, $p = 0.008$; experiment 2: $F_{29} = 5.80$, $p = 0.023$) on trials that incurred a shift in priority between the two WM items, relative to the previous trial, than on trials in which the item priority repeated (see supplementary figure 1).

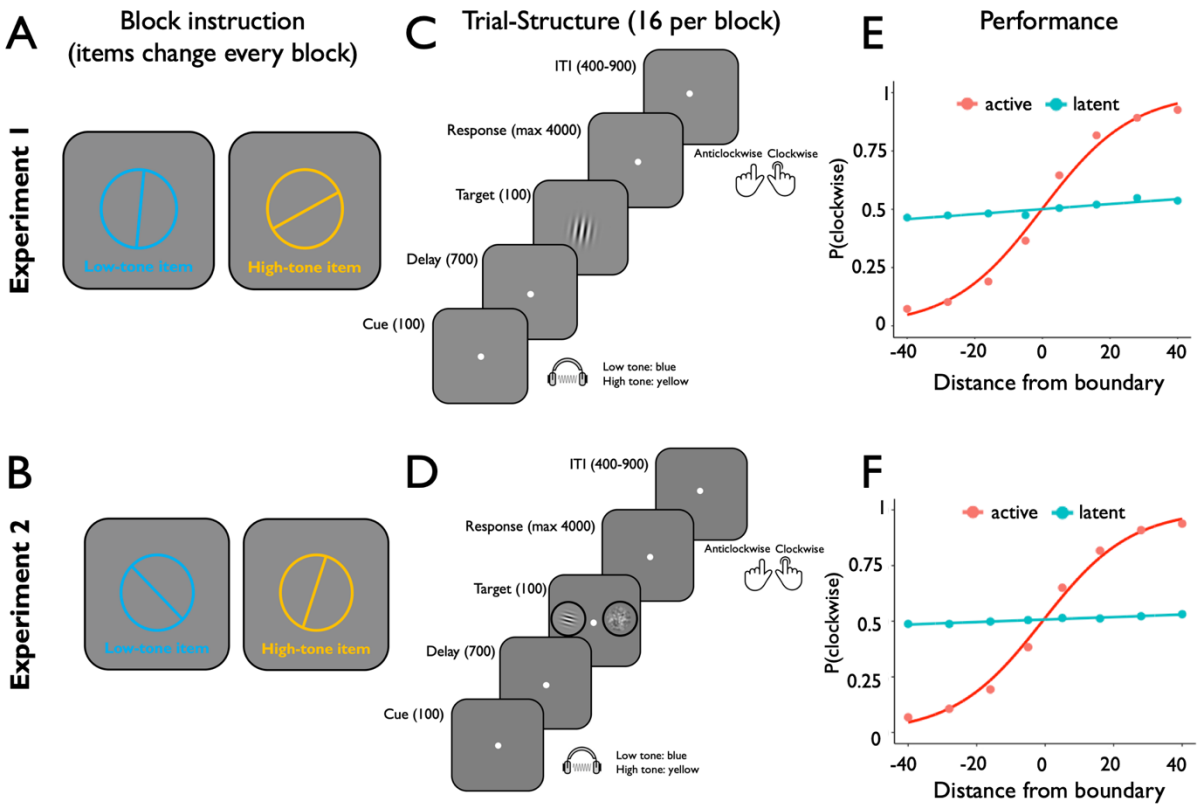


Figure 1: Illustration of the task design and behavioural results of experiment 1 (upper panel) and experiment 2 (bottom panel). ***A,B*** *At the beginning of each block, two randomly oriented bars were shown successively in blue and yellow colour. These two items served as WM items for the remainder of the block and were associated with a high-pitch tone and a low-pitch tone respectively (mapping counterbalanced across subjects).* ***C,D*** *During the block, each trial started with the presentation of an auditory cue that signals which WM item should be used as decision criterion on the current trial (active item), while the other item had to be maintained for later use in the block (latent item). After a brief delay, a randomly oriented gabor patch was presented as target, and participants were required to indicate if the target was oriented clockwise or anticlockwise relative to the active WM item. In experiment 1, the target was shown centrally on the screen (C). In experiment 2, the target was shown peripherally on the left or right side of the fixation dot, while a noise patch was presented on the other side (D). The noise patch matched the target in luminance, size, contrast, and eccentricity. Note that the location of the target was indicated to participants at the beginning of each block.* ***E,F*** *Probability of clockwise responses as a function of the angular difference between target stimulus and the currently active WM item (shown in red) and from the currently latent WM item (shown in blue). Data are shown in dots and the lines represent a fitted binomial cumulative distribution function. Responses were strongly modulated by the angular distance between the orientation of the target and the orientation of the active item, whereas the distance between the target and latent item had only minimal impact on performance (see Material and Methods and Results for details).*

Time-resolved decoding of task-variables

We conducted a series of multivariate pattern analyses to characterise the neural representations that underpin performance in our task. In a first step, we conducted a time-resolved decoding analysis to reveal the time courses within which three task variables were explicitly encoded in EEG sensor activity: 1) the orientation of the target stimulus, 2) the orientation of the WM item, separately for active and latent conditions, and 3) a decision variable that was calculated as the absolute distance between the orientation of the target and the orientation of the WM item, again separately for active and latent conditions. To recover information about these variables from EEG sensor activity with high temporal resolution, we computed Mahalanobis distances between the patterns of sensor activity that were evoked by different stimulus orientations, and measured the extent to which these distances reflected the underlying circular orientation space or linear decision variable space (see Figure 2 for illustration and Materials and Methods for details).

As shown in Figure 3, information about the orientation of the target stimulus started to be encoded in EEG sensor patterns briefly after target onset and decoding remained significant for the remainder of the selected time window (experiment 1: first significant time point = 72ms after the target, cluster-corrected $p < 0.0001$; experiment 2: first significant time point = 68ms after the target, cluster corrected $p < 0.0001$). Active and latent WM items were both decodable, but we observed clear differences in their respective time courses. Decoding of active item started to rise during delay period and was further magnified after target onset (experiment 1: first cluster = 208-376ms after the cue; cluster corrected $p = 0.011$; second cluster = 400-2000ms after the cue, cluster corrected $p < .00001$; experiment 2: 420-2000ms after the cue, cluster-corrected $p < 0.0001$). In contrast, the latent item was not decodable during the delay period, but there was time window of significant decoding after the onset of the target stimulus (experiment 1: 192-528ms after the target, cluster corrected $p = 0.0005$, experiment 2: 272-316ms after the target, cluster corrected $p = 0.048$). Slightly later during the target period, we could decode a decision variable for the active item (experiment 1: 244-1000ms after the target, cluster corrected $p = 0.0002$; experiment 2: 388-1000ms after the target, cluster corrected $p = 0.001$), whereas no decision variable was decodable for the latent item. Notably, the decision variable signal was not encoded in electromyographic (EMG) activity recorded from the hands, whereas the behavioural choice (clockwise vs. counter-clockwise) was robustly encoded immediately before the execution of a response in both EEG and EMG activity (see Supplementary Figure 2). These results underscore that the decision variable did not index processes related to response execution. Finally, for all three task variables, we also validated that eye-movements could not explain the pattern of EEG decoding results (see Supplementary Figures 3 and 4).

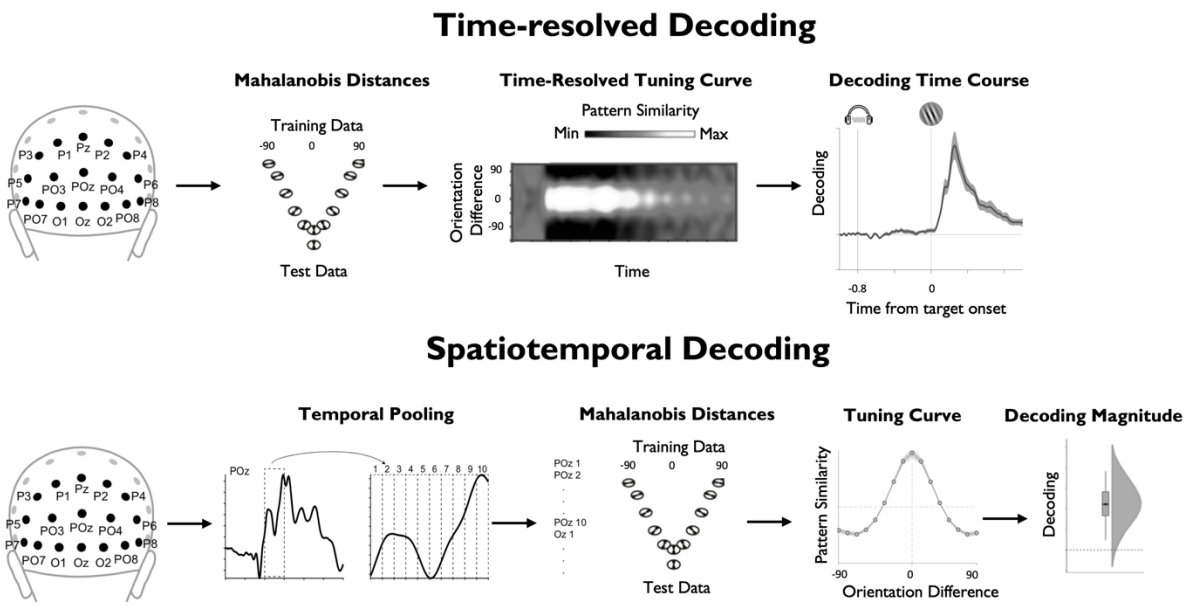


Figure 2: Illustration of time-resolved and spatiotemporal approaches that were used for the decoding of task variables. The upper panel illustrates the time-resolved decoding approach. Here, for each training set and time point, Mahalanobis distances are computed between activation patterns of posterior sensors of the test data and the training data (binned based on their orientation relative to the test data). After repeating this procedure for every training set and time point, the resulting distances were sign-reversed, so that positive scores reflect pattern similarity, and centred around the mean voltage across sensors for each time point. This was used to reconstruct a time-resolved population tuning curve that expresses the extent to which pattern similarity reflects the similarity of the underlying angular orientation space across time. We estimated the height of this tuning curve for each time point by convolving it with a cosine function (see Materials and Methods for details). The resulting vector served as index of time-resolved decoding accuracy. The bottom panel illustrates the spatiotemporal decoding approach. Here, the data entering the decoders were pooled over multiple time points, so as to exploit not only information that is encoded in spatial activation patterns, but also information that is encoded in their temporal structure. We focused on the time window from 100-400ms after target onset and treated individual trials as discrete events. Tuning curves were estimated based on mean-centred and sign-reversed Mahalanobis distances, as in the time-resolved approach, yielding a single index of decoding strength for each trial as final output.

Spatiotemporal Decoding

In a second step, we employed a complementary analysis approach to decode the same task variables from longer time-windows of stimulus-evoked brain activity. This approach exploits the dynamic temporal structure of event-related potentials by pooling multivariate information in time, thus capturing not only information encoded in spatial activation patterns, but also information encoded in the temporal unfolding of these patterns (see Figure 2 for illustration and Materials and Methods for details). Previous work has shown that such pooling over adjacent time points within trials can increase decoding sensitivity at the expense of temporal precision.[29,30] In keeping with these previous studies, we pooled data points over a time period from100-400ms after target onset and treated individual trials as discrete events. As shown in Figure 3, this approach clearly enhanced decoding sensitivity. Specifically, we could reliably decode the orientation of the target (experiment 1: $t_{19} = 11.585$, $p < 0.001$; experiment 2: $t_{29} = 11.301$, $p < 0.001$), the orientation of the active item (experiment 1: $t_{19} =$

5.599, $p < 0.001$; experiment 2: $t_{29} = 8.400$, $p < 0.001$) and the orientation of the latent item (experiment 1: $t_{19} = 4.232$, $p < 0.001$; experiment 2: $t_{29} = 2.253$, $p = 0.016$). Interestingly, and in contrast to the time-resolved analysis, we could decode a decision variable not only for the active item (experiment 1: $t_{19} = 5.922$, $p < 0.001$; experiment 2: $t_{29} = 5.233$, $p < 0.001$), but also for the latent item (experiment 1: $t_{19} = 2.627$, $p = 0.008$; experiment 2: $t_{29} = 1.889$, $p < 0.034$).

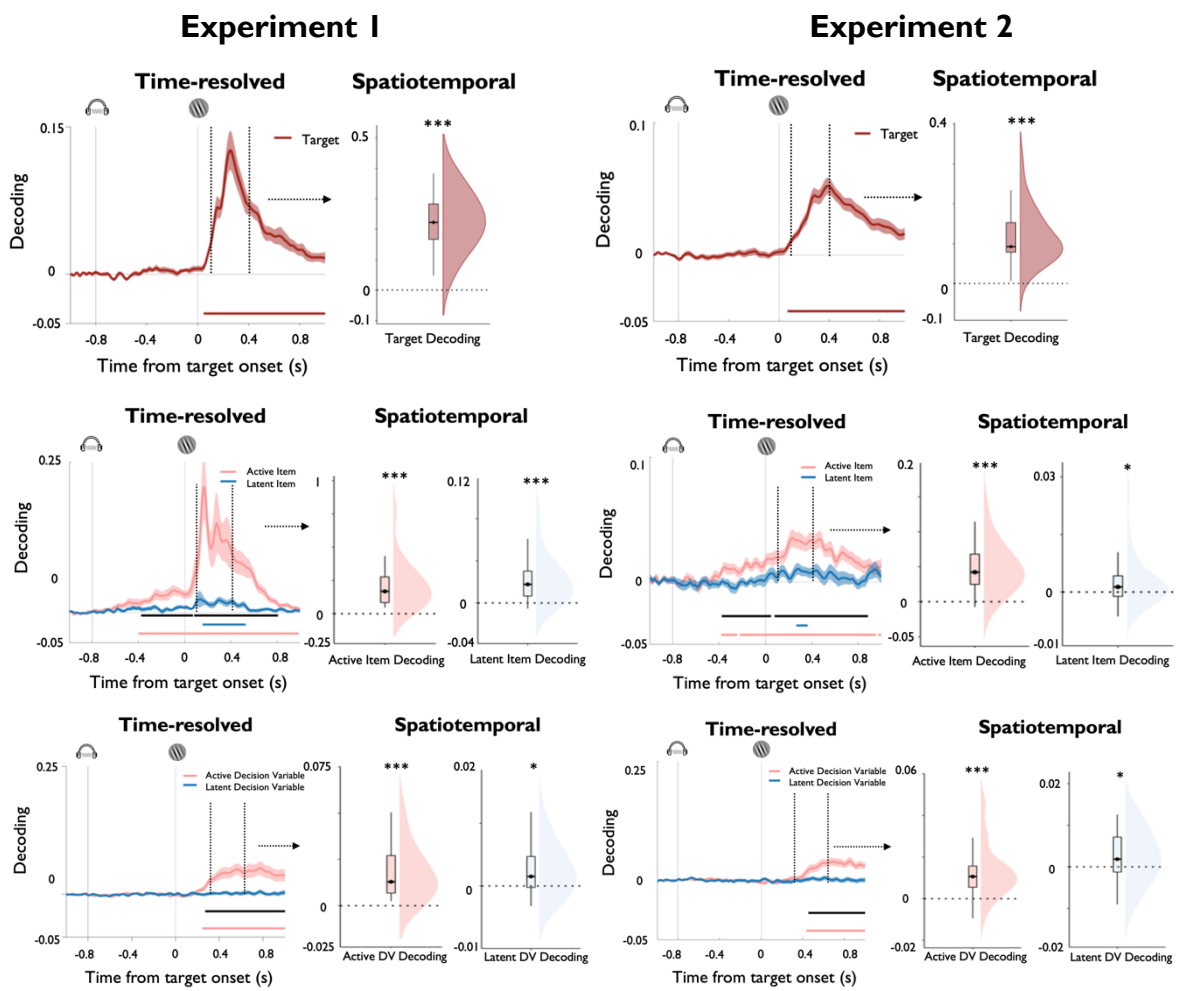


*Figure 3: Time-resolved and spatiotemporal decoding of task variables for experiment 1 (left side) and experiment 2 (right side). The upper panels display the decoding results for the orientation of the target stimulus, the middle panels display the decoding results for the orientation for the active and latent WM items, and the bottom panels display the decoding results for the absolute value of the decision variable, calculated separately with respect to the active and latent WM item. For each experiment and task variable, plots on the left display time-resolved decoding results from 200ms before cue onset until 800ms after target onset. Coloured lines represent cluster-corrected time-periods within which decoding was significantly greater than chance. Black lines in the middle and bottom rows indicate cluster-corrected time periods within which decoding strength of the active and latent WM items differed significantly. Shading indicates standard error of the mean. Plots on the right display results of the spatiotemporal decoding approach where data entered into the decoder were pooled within a time window from 100-400ms following target onset (see text for details). Boxplots: centre lines indicate the median, the box outlines indicate the 25th and 75th percentiles, and the whiskers indicate 1.5 times the interquartile range. Significance values: *** $p < 0.005$; ** $p < 0.01$; * $p < 0.05$.*

Regression analyses

(1) Active, but not latent, WM states predict trial-wise variance in task performance

We next conducted a series of regression analyses to establish the relevance of the decoded memory signals for task performance and test our hypothesis that active and latent items are encoded in qualitatively different functional states. To reiterate, we expected that the active item would be encoded in a decision-oriented state that directly predicts variance in performance, whereas the latent item should be held in a dormant state that has minimal impact on current performance. To test this prediction, we regressed log-transformed reaction time (log-RT) and accuracy against the trial-wise decoding strength of active and latent WM items using the spatiotemporal classifier. Decoding scores were normalised by subtracting the block average for the respective WM item from the trial-wise scores to isolate dynamic trial-wise fluctuations from more general differences in decoding accuracy between blocks (see section (3) for block-wise analyses). As shown in Figure 4, we observed that higher trial-wise decoding of the active item reliably predicted faster log-RT (experiment 1: t_{19} = -4.044, $p < 0.001$; experiment 2: t_{29} = -1.806, p = 0.041), and to some extent greater accuracy (experiment 1: t_{19} = 2.883, p = 0.005; experiment 2: t_{29} = 0.667, p = 0.255), whereas trial-wise variance in decoding of the latent item neither predicted log-RT (experiment 1: t_{19} = 1.699, p = 0.106; experiment 2: t_{29} = 0.973, p = 0.831) nor accuracy (experiment 1: t_{19} = -0.891, p = 0.384; experiment 2: t_{29} = 0.644, p = 0.525). These results support our prediction that active items are encoded in a prospective decision-oriented state that directly guides WM-based decisions, whereas latent items are held in a dormant state that has minimal impact on ongoing cognition.

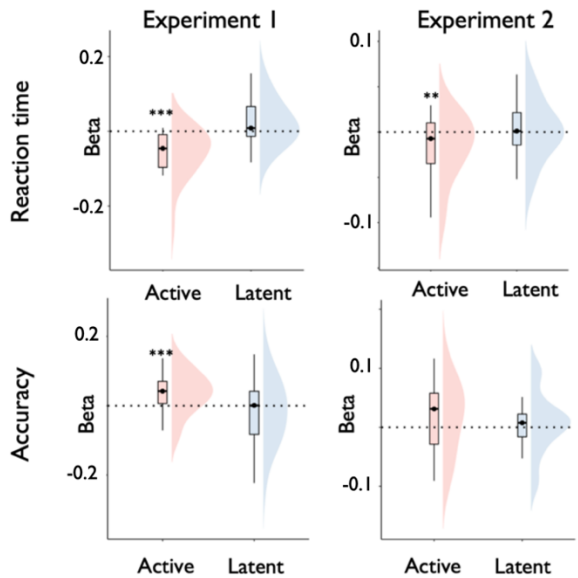


*Figure 4: Regression of task performance based on trial-wise variance in decoding strength of the active and latent WM items, separately for experiment 1 and experiment 2. The top row displays the distribution of regression weights across participants for the prediction of reaction time and the bottom row displays the for the prediction of accuracy. Boxplots: centre lines indicate the median, the box outlines indicate the 25th and 75th percentiles, and the whiskers indicate 1.5 times the interquartile range. Significance values: *** p < 0.005; ** p< 0.01; * p < 0.05.*

(2) Active WM items are simultaneously encoded in active and latent states

In a second step, we combined the foregoing analyses with a cross-item decoder that was trained on data sorted by the latent item but tested on data sorted by the active item (latent-to-active decoder). This approach aimed to test for overlap between the patterns coding for active and latent WM items, and to relate variance in this overlap to variance in performance. As shown in Figure 5, the cross-item decoder successfully classified active WM items with above-chance accuracy (experiment 1: $t_{19} = 4.972$, $p < 0.001$; experiment 2: $t_{29} = 3.067$, $p = 0.005$), though decoding was significantly weaker than with the regular decoder of the active item (experiment 1: $t_{19} = -5.456$, $p < 0.001$; experiment 2: $t_{29} = -8.081$, $p < 0.001$). This indicates that active and latent WM items were encoded in partly overlapping neural patterns and suggests two possibilities: firstly, active and latent WM states might be qualitatively similar with active states reflecting an amplified version of the same underlying state.[24-27] Secondly, both items might be encoded in latent states for general maintenance, but only the immediately task-relevant item is additionally encoded in active states for decision making (see Figure 5E for illustration). To adjudicate between these possibilities, we returned to the regression analyses, but now focussed on the shared patterns revealed by the latent-to-active cross-generalisation to isolate the putative latent component of the active item. If active and latent WM states are qualitatively similar, then the cross-decoder should exhibit similar links to performance as the regular decoder of the active item from the foregoing section. By contrast, if the overlap more specifically reflects the latent component for the active item, then it should not exhibit links with the quality of task performance, similar to the regular decoder of the latent item in the foregoing analysis. We observed that the latent-to-active decoder neither predicted log-RT (experiment 1: $t_{19} = -1.739$, $p = 0.098$; experiment 2: $t_{29} = -1.658$, $p = 0.108$) nor accuracy (experiment 1: $t_{19} = -0.643$, $p = 0.736$; experiment 2: $t_{29} = -0.818$, $p = 0.799$) on the current trial, suggesting that the overlap between patterns coding for the two WM items reflects a dormant mnemonic state that is unrelated to the active representation, and does not correlate with trial-wise variance in performance. Importantly, however, the cross-item decoding does provide meaningful variance for the current analysis, as we observed a significant link to behavioural accuracy at a block-wise timescale (see next section).

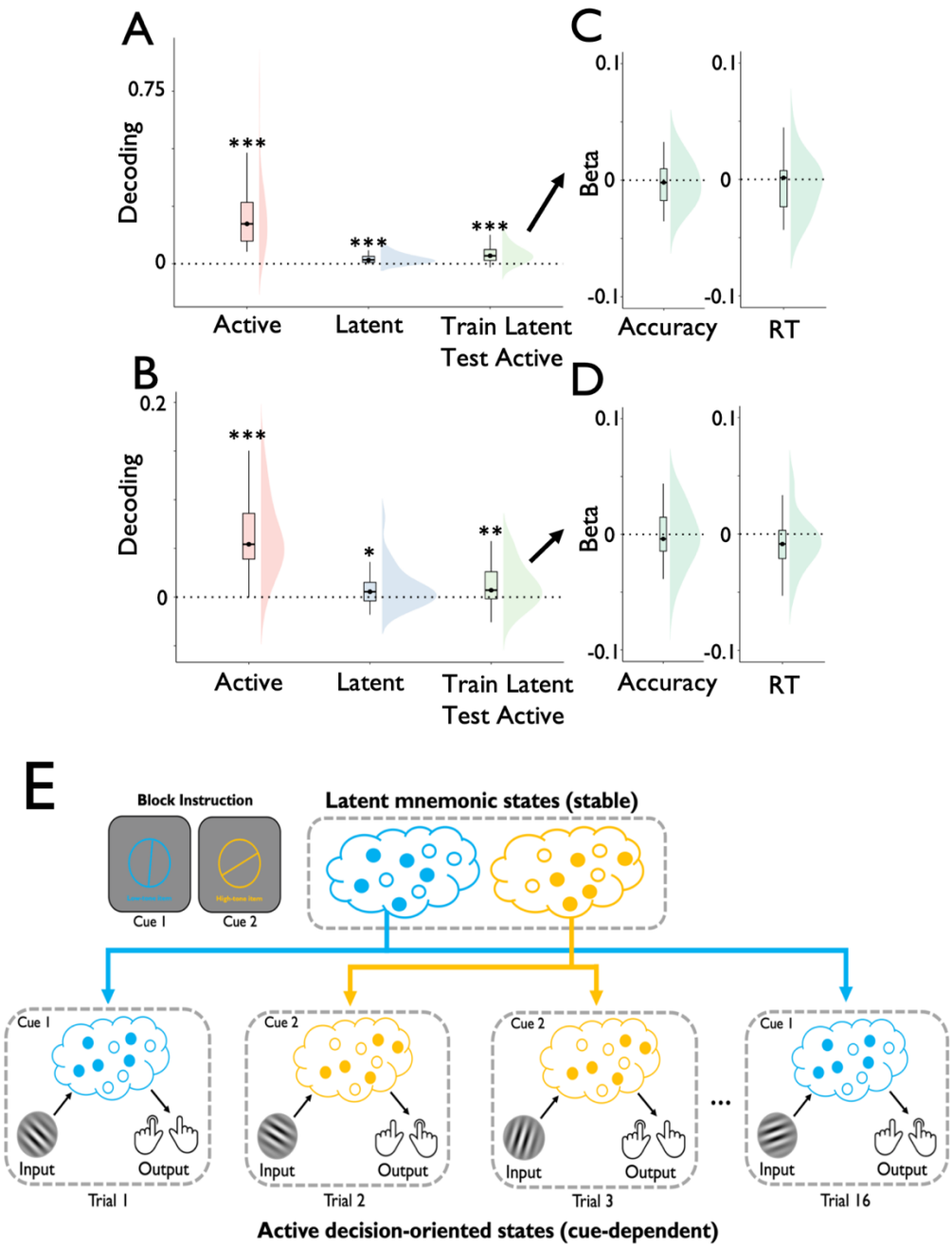


Figure 5: ***A, B*** *Cross-generalisation between active and latent WM items was tested using a decoder that was trained on the data from the latent WM item and tested on the data from the active WM item. The cross-item decoder classified active items with above chance accuracy in experiment 1 (A) and experiment 2 (B), but with a strong cost of cross-generalisation, suggesting that active and latent WM items are encoded in partly overlapping activity patterns.* ***C, D*** *Regression analyses established that the overlap in patterns coding for active and latent WM items does not predict variance in reaction time or accuracy. Plots display the distribution of regression weights across participants. Boxplots: centre lines indicate the median, the box outlines indicate the 25th and 75th percentiles, and the whiskers indicate 1.5 times the interquartile range.* ***E*** *These results are consistent with a hierarchical model of functional WM states wherein both WM items are encoded in latent mnemonic states that support general maintenance, but the immediately task-relevant item is additionally encoded in an active state that implements task-relevant transformations of sensory inputs into behavioural outputs. Within this model, latent states form the basic substrate of maintenance and can be flexibly converted into active states depending on task demands.*

(3) Both active and latent WM states track performance accuracy over longer time scales

After establishing that the decoding strength of latent WM states is unrelated to task performance on the current trial, we next tested the extent to which these signals nonetheless capture behaviourally relevant variance by examining the contribution of latent WM states to general maintenance. Although a strong representation of the latent item is not advantageous on the current trial, it is nonetheless critical for performance on other trials in the same block when priorities differ, and the currently latent item becomes active. Therefore, maintaining an accurate representation of the latent WM item reflects a minimal requirement for solving the task as a whole. Accordingly, the overall quality with which a WM item is encoded throughout the block when it is latent should track performance in the same block on those trials when the same item is active. We tested this assumption by regressing trial-wise log-RT and accuracy against the decoding strength of the two memory items averaged over the remaining trials of the block. As illustrated in Figure 9, for each trial we calculated the mean decoding score across all remaining trials of the same block on which the respective memory item was active, and across all remaining trials within the block on which the same item was latent. Average decoding scores were then used to predict trial-wise variance in performance. Assuming that block-wise decoding tracks the general quality with which memory items are encoded throughout the block, we reasoned that both active and latent decoding strength should scale positively with performance. As shown in Figure 6, we observed that block-wise decoding indeed predicted accuracy both for the active item (experiment 1: $t_{19} = 2.361$, $p < 0.014$; experiment 2: $t_{29} = 2.162$, $p = 0.020$) and for the latent item (experiment 1: $t_{19} = 1.819$, $p < 0.042$; experiment 2: $t_{29} = 2.224$, $p = 0.034$). Block-wise decoding of the active item moreover predicted log-RT (experiment 1: $t_{19} = -3.869$, $p < 0.001$; experiment 2: $t_{29} = -1.804$, $p = 0.041$), which was not the case for the latent item (experiment 1: $t_{19} = -1.197$, $p = 0.246$; experiment 2: $t_{29} = -1.411$, $p = 0.169$). We next repeated this analysis using the data from the latent-to-active cross-item decoder used in the previous section to validate our assumption that the pattern overlap between active and latent WM items reflects a functionally dormant mnemonic representation of the active item. Under this account, we expected the cross-item decoder to exhibit similar links with performance as the regular decoder of the latent item. In line with this prediction, we observed that that block-wise variance in decoding predicted accuracy (experiment 1: $t_{19} = 2.071$, $p = 0.026$; experiment 2: $t_{29} = 1.920$, $p = 0.032$), but not log-RT (experiment 1: $t_{19} = -0.424$, $p = 0.338$; experiment 2: $t_{29} = 0.495$, $p = 0.312$). Together, these results show that even though latent WM states are dormant on the current trial, they nonetheless support performance over longer timescales.

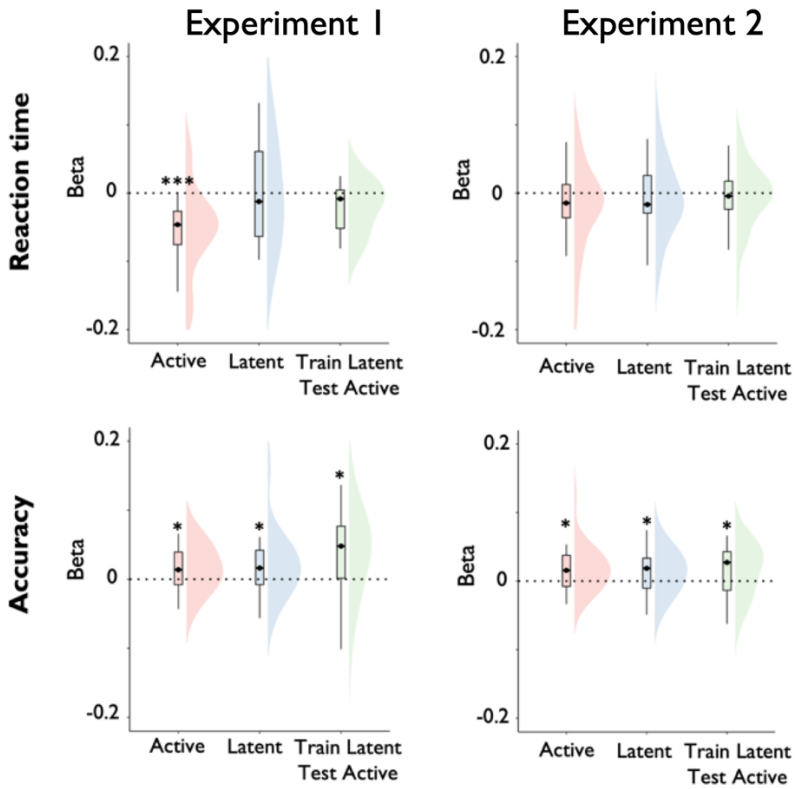


*Figure 6: Regression of task performance based on block-wise decoding accuracy of the active and latent WM items, separately for experiment 1 and experiment 2. The top row displays the distribution of regression weights across participants for the prediction of reaction time and the bottom row displays the same for the prediction of accuracy. Boxplots: centre lines indicate the median, the box outlines indicate the 25th and 75th percentiles, and the whiskers indicate 1.5 times the interquartile range. Significance values: *** $p < 0.005$; ** $p < 0.01$; * $p < 0.05$.*

(4) Similarity to the active state predicts interference from latent item after priority shifts

The previous sections established that latent WM items are encoded in a format that supports general maintenance, but without interfering with ongoing processing. However, at the behavioural level, we had observed two signatures of interference between WM items: (i) a modest but reliable effect of the angular distance between the latent item and the target stimulus, and (ii) a performance cost on trials demanding a priority shift between WM items relative to the previous trial. Given our previous finding that the strength of active, but not latent, WM states tracks fluctuations in performance, we next examined if interference between WM items may reflect the extent to which latent WM items are encoded in patterns resembling their active state. To this end, we combined the trial-wise regression analyses used above with a cross-item decoder that was trained on data sorted by the active item and tested on data sorted by the latent item (active-to-latent decoder). This decoder should reveal fluctuations in the extent to which latent items are encoded in patterns that resemble the item's active state. As shown in Figure 7, the active-to-latent cross-decoder successfully classified latent WM items with above chance accuracy (experiment 1: $t_{19} = 4.867$, $p < 0.001$; experiment 2: t_{29}

= 2.751, p = .010). Decoding magnitude was enhanced in comparison to the regular decoder of the latent item in experiment 1 (t_{19} = 2.74, p = 0.013), whereas the decoders did not differ in decoding magnitude in experiment 2 (t_{29} = 0.423, p = 0.675). Critically, higher trial-wise active-to-latent decoding reliably predicted slower log-RT (experiment 1: t_{19} = 2.235, p = 0.019; experiment 2: t_{29} = 2.781, p = 0.005) and to some extent reduced accuracy (experiment 1: t_{19} = -2.319, p = 0.016; experiment 2: t_{29} = -0.540, p = 0.297), corroborating that interference from latent WM items indeed reflects their similarity to the active state. We reasoned that these instances might be expressed most strongly on trials that demand priority shifts between the two items, and therefore repeated the foregoing analysis separately for trials that incurred a shift in priority relative to the previous trial (switch trials) and trials that did not (repetition trials). Consistent with our hypothesis, the negative prediction of log-RT was significant only on switch trials (experiment 1: t_{19} = 3.259, p = 0.002; experiment 2: t_{29} = 3.351, p = 0.002), but not on repetition trials (experiment 1: t_{19} = 1.127, p = 0.137; experiment 2: t_{29} = 0.648, p = 0.522), and the difference between trial types was significant (experiment 1: t_{19} = 2.034, p = 0.028; experiment 2: t_{29} = 1.725, p = 0.048). The regression of accuracy scores was neither significant for switch trials (experiment 1: t_{19} = -1.140, p = 0.134; experiment 2: t_{29} = -0.319, p = 0.134) nor for repetition trials (experiment 1: t_{19} = -1.680, p = 0.109; experiment 2: t_{29} = -0.752, p = 0.346) and it did not differ between trial types (experiment 1: t_{19} = 0.034, p = 0.974; experiment 2: t_{29} = -0.098, p = 0.923). Collectively, these results suggest that interference between WM items arises due to lingering neural patterns coding for previously active, but no longer immediately task-relevant, items.

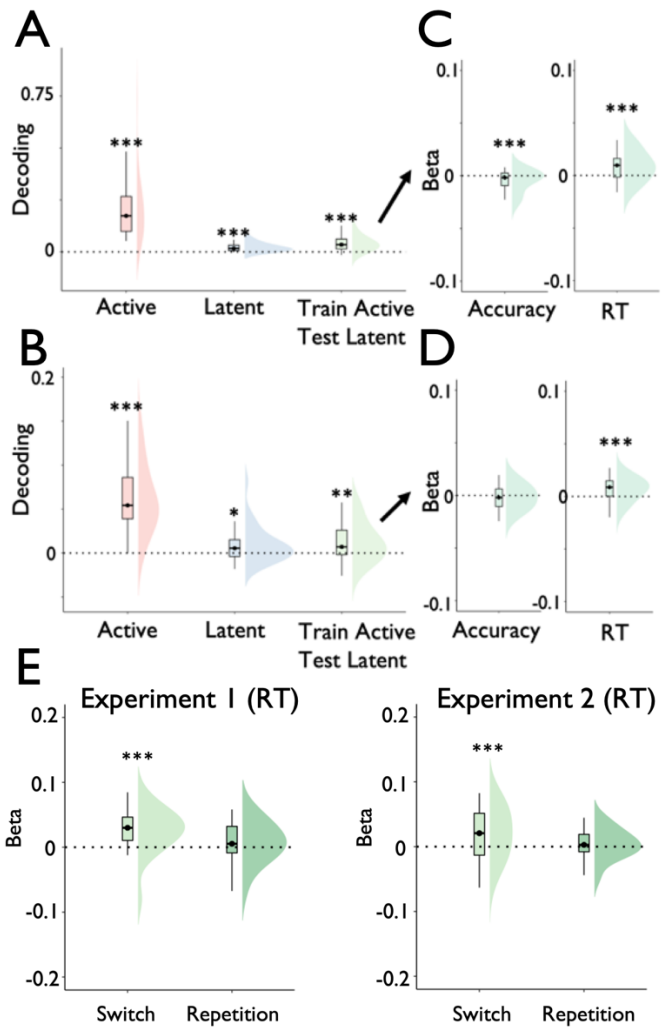


Figure 7: ***A, B*** *Cross-generalisation between active and latent WM items using a decoder that was trained on the data from the active WM item and tested on the data from the latent WM item. This cross-item decoder classified latent items with above chance accuracy in experiment 1 (A) and experiment 2 (B).* ***C, D*** *Trial-wise regression analyses established that the extent to which latent items were encoded in patterns resembling active states predicts behavioural indices of cognitive interference, namely reduced accuracy and slower reaction time on the current trial.* ***E*** *Reaction time effects were only observed on trials that required a priority shift between WM items relative to the previous trial (switch trials), but not on trials in which the item priority was repeated (repetition trials). Plots display the distribution of regression weights across participants. Boxplots: centre lines indicate the median, the box outlines indicate the 25th and 75th percentiles, and the whiskers indicate 1.5 times the interquartile range.*

Drift Diffusion Modelling

We conducted a final set of analyses to obtain more detailed insights into the mechanisms by which active WM states influenced decisions in our task and compare two different explanations for the link between decoding strength and performance (see Figure 8 for illustration): firstly, the strength of the active state may determine the ease with which sensory input is interpreted, and transformed into a decision variable.[31,32] This would be the case if the active WM item acts as a matched filter that directly

feeds in the accumulation of decision-related evidence (matched filter hypothesis). Secondly, the strength of the active state may determine the ease with which the items can be retrieved for decision making,[33] thus providing a head-start for decisions on trials with a strong representation of the active item (retrieval head-start hypothesis).

To evaluate these possibilities, we fit a set of drift diffusion models (DDMs) to our behavioural data and related variance in model parameters to variance in decoding. DDMs characterise decisions as the accumulation of noisy evidence between two competing options, whereby one option is chosen once a sufficient amount of evidence has been accumulated.[34] DDMs decompose performance in two-alternative choice tasks into decision parameters based on RT distributions and choice probabilities. Here, we focused on two parameters of the DDM: drift rate and non-decision time. The *drift rate* reflects the quality with which decision-relevant information is integrated, and scales negatively with categorisation difficulty (i.e., drift rates decrease with more difficult discriminations). The *non-decision time* reflects the time needed for processes that are not directly related to evidence accumulation such as the encoding of a stimulus or the execution of a response. To test the accounts outlined above, we fit a set of linear regression models predicting trial-wise changes in the two decision parameters based on trial-wise changes in decoding strength of the active memory item (see Materials and Methods section for details). The matched filter hypothesis predicts that stronger decoding is associated with higher drift rates, whereas the retrieval head-start hypothesis predicts that stronger decoding is associated with reduced non-decision time (see Figure 8). In support of the matched filter hypothesis, we observed reliable positive regression weights for predictions of drift rate in both experiments (experiment 1: $P_{Beta > 0} = 0.99$; experiment 2: $P_{Beta > 0} = 0.97$), whereas regression weights for the non-decision time were indistinguishable from zero (experiment 1: $P_{Beta < 0} = 0.46$; experiment 2: $P_{Beta < 0} = 0.54$).

We conducted an additional post-hoc analysis to follow up on this result. Here, we computed the EEG-based decision variable separately for trials with low and high decoding strength of the active WM item (based on median split), and compared the respective decoding time courses. As shown in Figure 8E, the signal began to rise around the same time for both classes of trials, but the rise was faster on trials with strong decoding of the active item, consistent with a change in the quality of evidence integration. Altogether, these results underscore that fluctuations in decoding strength of the active item track the efficiency with which sensory input is transformed into a task-dependent decision variable, but they do not affect processes preceding or following evidence integration.

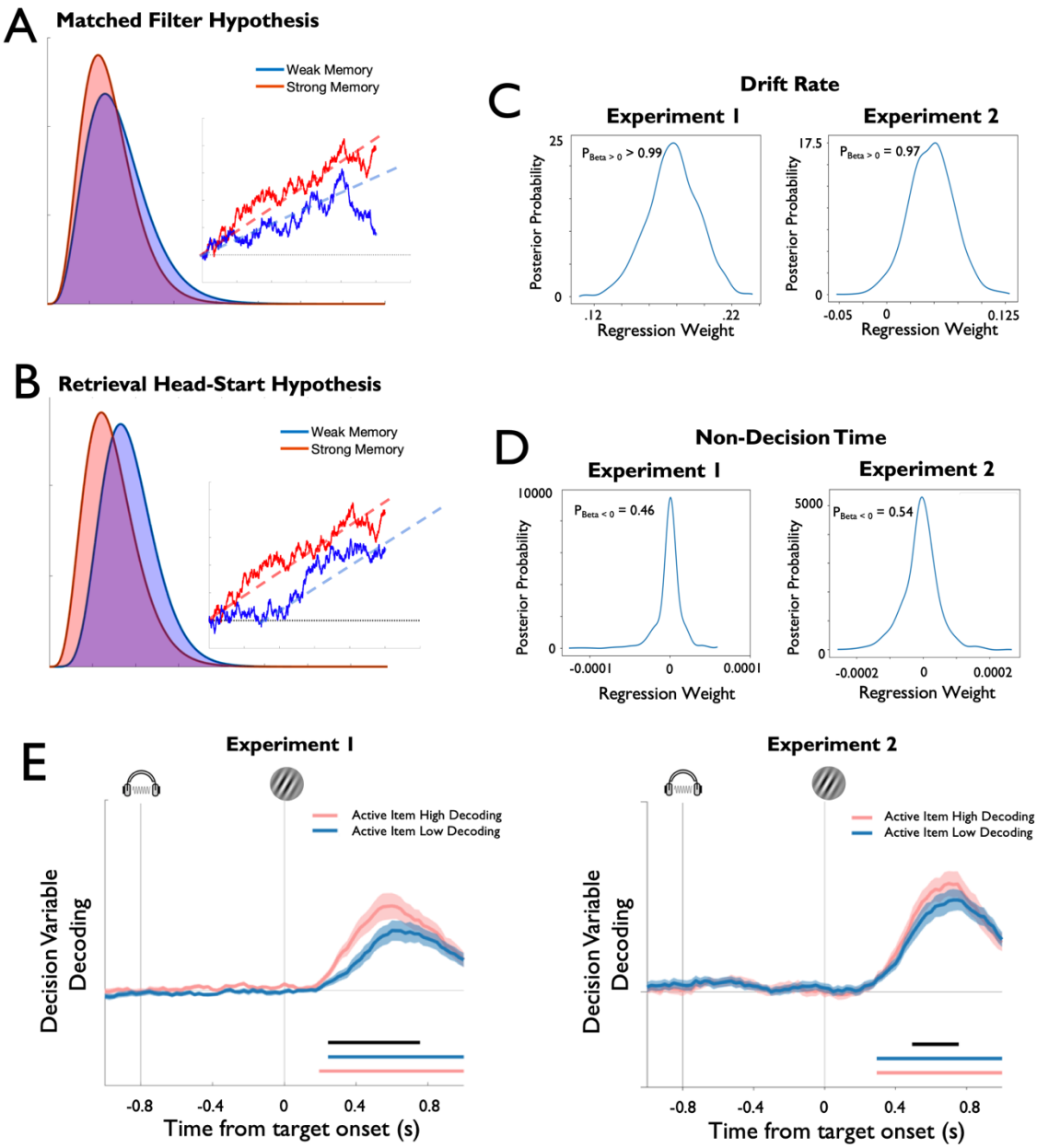


Figure 8: Drift diffusion modelling of task performance as a function of decoding strength of the active WM item. ***A, B,*** *Illustration of the two hypotheses under investigation that predict that trial-wise variance in decoding of the active item tracks either changes in drift rate (matched filter hypothesis) or non-decision time (retrieval head-start hypothesis). The plots display simulated RT distributions and exemplary single-trial diffusion patterns under the two accounts.* ***C, D*** *Results of the DDM regression analysis that predicted trial-wise changes in drift rate (**C**) and non-decision time (**D**) based on trial-wise changes in decoding strength of the active item. Plots display the posterior distribution of estimated regression weights, resulting from 5,000 iterations from which the initial 500 iterations were discarded as burn in. The significance of effects was evaluated by quantifying the amount of the posterior probability mass that was in the predicted direction (positive for drift rate and negative for non-decision time; see Materials and Methods and Results for details). As indicated on the plots, decoding of the active item reliably predicted changes in drift rate, but it was not associated with changes in non-decision time.* ***E*** *Decoding of the decision variable as a function of the decoding strength of the active WM item (based on a median split) corroborates the DDM results by showing that the decision variable signal begins to rise at the same time for both classes of trials, but it rises faster when decoding of the active WM item is strong.*

Discussion

In the present study, we compared different forms of WM storage using a task that required the maintenance of two items for an extended period of time as well as flexible priority shifts between them. Participants saw the items only during an initial encoding phase and did not receive response feedback throughout the block. Accordingly, while only a single item was task-relevant on any given trial, participants had to maintain precise representations of both items to be able to solve the task as a whole. EEG pattern analyses could recover representations of both items held in WM: decoding of the active item started to ramp up during the delay period soon after the item was cued. The latent item was not decodable during the delay period, but we observed a robust trace of both WM items in the target-evoked EEG response. Critically, the two items were encoded in two qualitatively different functional states that supported task performance in different ways: the strength of the latent state was unrelated to performance on the current trial, but predicted decision accuracy over longer, block-wise time scales, suggesting a role in the maintenance of prospectively relevant information over an intermediate time scale of minutes. By contrast, the strength of the active state predicted dynamic trial-by-trial changes in the quality of evidence integration, suggesting a critical role in guiding WM-based decisions over a sub-second time scale.

Our results provide direct empirical evidence for the proposal that the use of WM for behaviour not only involves selective internal attention to task-relevant items, but also a reconfiguration of latent mnemonic representations into active action-ready states.[28] Consistent with this view, a previous fMRI study by Muhle-Karbe et al. compared neural patterns coding for visual stimuli that were either encoded as a task rule for subsequent implementation or as a memorandum for subsequent recall.[15] Patterns critically depended on the anticipated readout demand, and were less time-stable during implementation, suggestive of a change in the underlying neural code. Our findings extend this work by establishing differential links between the strength of active and latent WM states and the success of task-dependent decision making. Variance in the latent state was predictive of performance accuracy at a block-wise time scale. This highlights the functional contribution of latent WM states to decision making by providing a storage format that protects memories from decay, while minimising interference with currently prioritised cognition. Such a coding scheme is critical for many complex and time-extended behaviours that incur multiple nested components and priority shifts among them.[35,36] Understanding how the brain structures cognition over such intermediate time scales will be key for developing more integrative theories of memory-guided behaviour,[37,38] and we anticipate that such research may reveal maintenance contributions of structures in the medial temporal lobe that have classically been associated with long-term memory.[39]

In contrast to the latent state, variance in the active state predicted dynamic trial-wise changes in performance, especially reaction time. Interestingly, these effects were specifically tied to changes in drift rate, while leaving other component processes reflected in the non-decision time unaffected. This suggests that active WM items act as matched filters that directly transform sensory information into task-dependent decision variables, e.g., by weighting the target-evoked patterns by mnemonic patterns coding for the active WM item.[31,32,40-42] At the same time, this result is at odds with proposals that attribute benefits of prioritisation within WM to the facilitation of retrieval processes, which may give decisions a head start, compared to decisions without prioritisation.[33] Notably, perceptual decisions have typically been studied using fixed decision boundaries, but varying levels of stimulus strength.[43] Under these conditions, changes in drift rate have typically been associated with changes in the noise level of the stimulus. Our results complement this literature and indicate that drift rate is sensitive not only to changes in the quality of decision-relevant sensory evidence, but also to changes in the quality of internal representations of the decision criterion. WM-guided decisions thus appear to involve sequential sampling to match sensory input with mnemonic templates,[44,45] analogously to what has been proposed for decisions based on contents of long-term memory.[46]

As reviewed by Myers et al.,[28] conceiving prioritisation within WM as a transition from latent mnemonic states toward active task-oriented states provides a new perspective on other established experimental phenomena such as the effects of retrospective cuing.[47] These effects have most commonly been explained by assuming a finite resource for WM maintenance that can be continuously distributed among different items.[48] From this perspective, a more selective focus on task-relevant items improves memory for cued items, but at the cost of degraded memory for uncued items. The account we propose offers an alternative explanation whereby retro-cue benefits reflect, at least in part, the reformatting of a purely mnemonic representation into a prospective action rule. From this perspective, capacity limitations also concern behavioural readout, in addition to basic maintenance, when assuming that only a single item can be encoded as an action rule at any given moment.[18] It also implies that prioritisation should not necessarily impede the maintenance of other concurrently held items, because once the selected item is encoded in a rule-like state, it is no longer necessary to sustain selective attention to the corresponding mnemonic representation. This could explain findings that beneficial retro-cue effects for cued items can sometimes prevail even when attention is diverted from the cued item,[49,50] or that they can occur in the absence of reliable performance costs for uncued items.[51,52] The distinction between active and latent WM states also accords well with cases of goal neglect, where WM contents remain accessible, even though they fail to exert control over behaviour.[16]

A central question for future research will be to delineate the precise neurophysiological mechanisms that underpin functionally active and latent WM states. One possibility is that functionally active and latent WM states directly map onto corresponding activity states: neurally active and 'activity-silent.' [6,53,54] Our results are consistent with this view, given that spontaneous delay period activity encoded only the active item, but we could identify traces of both items in the target-evoked EEG signal, consistent with an impulse response uncovering hidden neural states.[12-14,29,30] However, we note that our current study was not designed to test this specific question but instead focused on dissociating WM states functionally. Therefore, our results are agnostic to the precise neural implementation. It is theoretically possible for activity states, or activity-silent states, to support functionally active or latent cognitive states.[55] The only formal requirement is that their representational format differs qualitatively. This could be achieved by a division of labour between activity-based and activity-silent mechanisms, but also through qualitative differences in the exact patterns, or brain areas, for respective functional states.

Another important topic for future inquiry will be to compare the hierarchy of states we identified with different mnemonic coding schemes, especially the possibility that active and latent items could be represented in inverted activation patterns in the same cortical regions. Two recent fMRI studies that manipulated the priority status of WM items within individual trials observed negative cross-decoding between items,[23,56] suggestive of opponent coding. At first blush, our findings are inconsistent with this proposal, as we observed positive rather than negative cross-item decoding in our task. There are numerous differences between the two studies and ours, so we can only speculate about the source of this discrepancy. For example, the different brain imaging modalities may have captured different aspects of the underlying neural signals with EEG being more sensitive to rapidly unfolding stimulus-evoked signals and fMRI being more sensitive to tonic and regionally specific activation patterns.[57] We assume that orthogonal coding schemes for active and latent items confer general advantages over inverted schemes, as they permit robust maintenance of both items while minimising interference among them. However, task demands may alter the optimality of different formats. For instance, inverted representations might become advantageous when simultaneously maintained items are encoded in non-overlapping neural populations, which may facilitate suppression of patterns coding for latent items; or when readout demands require maximal disambiguation between WM items.[58,59] Therefore, more research will be necessary to establish the boundary conditions under which the brain employs orthogonal and inverted coding schemes for the maintenance of multiple items in WM.

In conclusion, our results delineate a hierarchical model of WM in which a single item is stored in a qualitatively different format to other concurrently held items. The prioritised format is functionally active and implements a task-relevant transformation of sensory input into decision-evidence, whereas

other items are stored in a functionally latent format that does not interact with ongoing processing. Importantly, prioritisation is highly flexible and dynamic, whereby latent states form the basic neural substrate for maintenance, and can be used to implement the active representation when needed.

Materials and Methods

Participants

20 healthy adults participated in experiment 1 (mean age = 28.1, age range = 18-37, 10 female, 1 left-handed) and 30 healthy adults participated in experiment 2 (mean age = 26.8, age range = 19-41, 14 female, 2 left-handed). Eleven participants took part in both experiments (among them authors PSMK and NEM). All subjects reported normal or corrected-to-normal vision and received a monetary compensation for participation. The study was approved by, and conducted in accordance with, the guidelines of the Central University Research Ethics Committee of the University of Oxford.

Apparatus

Stimulus presentation was controlled in Matlab using Psychtoolbox on a 22" monitor with a refresh rate of 100Hz. Unless reported otherwise, stimuli were shown in white font on a grey background (50% contrast, RGB = [127 127 127]) at a distance of approximately 60cm. Responses were given with the left and right index fingers on the "B" and "Y" buttons of a QWERTY keyboard. EEG data were collected using 61 channels that were distributed across the scalp according to the extended 10-20 positioning system. Data were collected at 1000 Hz using a NeuroScan SynAmps RT Amplifier and Curry 7 software. Impedances of all channels were kept below 5 kOhm. In both studies, eye-movements were recorded via electrooculography (EOG) using electrodes places above and below the left eye and to the left of the left eye and to the right of the right eye. In study 2, we additionally recorded eye-movements using a remote infrared eye-tracker (SR Research, Eyelink 1000) sampling both eyes at 1000Hz. We also recorded activity in the first dorsal interosseus muscle of the left and right hand via electromyography (EMG) at 1000Hz.

Task design

The experimental task required the prolonged maintenance of two items in working memory and flexible priority shifts between those items to guide task-dependent decision making (see Figure 1 for an illustration). Overall, the task was broken into blocks each consisting of 16 trials. At the beginning of each block, participants were shown two orientated bars (6° visual angle in length and 0.25° in width, presented at the same location as subsequent stimuli) in blue (RGB = [25.5 25.5 204]) and

yellow (RGB = [204 204 25.5]) colour. These two bars served as memory items for the remainder of the block. For each participant, one colour was associated with a high-pitch tone and the other colour was associated with a low-pitch tone (mapping counterbalanced across participants). The two orientations were drawn randomly from a set of 16 possible orientations (spaced evenly at 11.25° intervals from 2.8125° to 171.5625°) with the sole constraints that the angular distance between the two items could not be identical or exactly orthogonal. Participants could encode the two items for a duration of their choosing and initiated the block via button press. Within the block, each trial started with a presentation of an auditory cue (pure sinusoidal tones, low tone: 440Hz, high tone: 880Hz, duration: 100ms including a 10ms ramp-up and 10ms ramp-down). The cue identity signalled which memory items should be used as a boundary for a forthcoming perceptual decision (active item), while the other item was maintained merely for later use in the block (latent item). The cue was followed by a 700ms delay period within which a black fixation dot was presented centrally on the screen (0.15° diameter). Thereafter, a randomly oriented Gabor patch was presented (orientation drawn a set of 16 possible orientations, spaced evenly at 11.25° intervals from 8.4375° to 177.1875°; patches had 6° diameter, 50% contrast, 1.75 cycles/°, random phase, Gaussian envelope with 1.2° SD). Participants were given a maximum time window of 4,000ms to classify the target via button press as being tilted clockwise or counter-clockwise relative to the active memory item. Targets were presented for 100ms and replaced by a fixation dot for the remainder of the response period. In experiment 1, the target was presented centrally on the screen on all trials. In experiment 2, the target was presented laterally at a distance of 6° from the screen centre. The side of target presentation (left vs. right) alternated predictably across blocks. A noise patch (Gaussian smoothed random white noise using a kernel with 0.13° SD, convolved with a Gaussian envelope with 1.2° SD) that matched the target stimulus in luminance, size, contrast, and eccentricity, was presented on the side of the screen at which no target appeared (see Figure 1 for illustration). Clockwise and counter-clockwise decisions were indicated with the right and left index fingers respectively. The response period was followed by a variable intertrial-interval (400-900ms, drawn from a truncated exponential distribution with mean 550ms) until the next trial started with the presentation of a cue. Importantly, participants received feedback about their performance only at the end of each block when their mean accuracy and response time from the preceding block was presented. Accordingly, they had to maintain precise representations of both memory items for the whole duration of the block, as they could not rely on trial-wise feedback to infer the orientation of an item if it was forgotten. Overall, participants completed 128 blocks, resulting in a total of 2048 trials and lasting approximately 2h.

Behavioural data analyses

Performance accuracy and reaction time was analysed using general linear models (GLMs). Initially, we tested to what extent performance was affected by the angular distance between the orientation of

the target and the orientation of the active WM item. We next repeated the analysis but this time testing for an effect of the distance between target orientation and the orientation of the latent WM item. To visualise these results, we fit the proportion of clockwise choices with a binomial cumulative distribution function using a GLM implemented in R with the ggplot2 and the psyphy packages. Finally, we fit another GLM to test for the presence of performance costs as a result of shifting priority within WM. To this end, we compared accuracy and RT between trials that incurred a shift in priority (switch trials) and trials that incurred priority repetitions.

Preprocessing

EEG data were initially re-referenced to the average of both mastoids. EEG, EOG, and EMG data were then down-sampled to 250Hz and bandpass filtered, using a high-pass filter of 0.1Hz and a low-pass filter of 45Hz. EEG channels with excessive noise were identified through visual inspection and replaced via interpolation using a weighted average of the surrounding electrodes. The continuous time series data was then divided into epochs, corresponding to the experimental trials starting 200ms before to the onset of the cue and terminating 1800ms after the onset of the target. Each trial was inspected visually for blinks, eye-movements, and non-stereotyped artefacts. Trials were rejected if they contained any of those artefacts during the delay and/or target period. Stereotyped artefacts outside those periods were subsequently removed from the data via independent component analysis. Unless stated otherwise, the data were baseline-corrected for the decoding analyses using the average signal from the time window of 200 to 50 ms before cue onset. Eye-tracking data were down-sampled to 250Hz and we identified and interpolated blinks using spline interpolation and a time window of ±100ms around the event.[61]

EEG decoding analyses

We conducted a series of multivariate pattern analyses to characterise the neural representations that underpin performance in our task.

Time-resolved decoding

In a first step, we conducted a *time-resolved decoding analysis* to reveal the time courses at which three variables of our task were explicitly encoded in EEG sensor activity: 1) the orientation of the target stimulus, 2) the orientation of the active and latent memory item, and 3) a decision variable that was calculated as the absolute distance between the orientation of the target and the orientation of the two memory items. To recover parametric information about these variables from EEG sensor activity with high temporal resolution, we computed Mahalanobis distances between the patterns of sensor

activity that were evoked by different stimulus orientations and measured the extent to which these distances reflected the underlying circular orientation space (or linear decision variable space). We used a leave-one-block-out cross validation procedure with the data from each block serving as test data once and all the remaining data serving as training data. Training data were subdivided into 16 bins, according to their orientation relative to the test data, and then averaged. Mahalanobis distances between the 16 average patterns in the training data and each test trial data were then computed using the noise covariance from training data. Noise covariance was calculated on the average residual data after subtracting orientation-specific mean activity from each trial with the corresponding orientation. Residuals were averaged over a time window from 0 to 1.8s relative to the onset of the auditory cue (and therefore encompassed the cue, delay, target, and decision phases). The covariance was calculated over the trials-by-sensors matrix of average residuals, using a shrinkage estimator.[62] This procedure was repeated for every block and for every time point within the trial using 4ms time bins. We normalised the resulting pattern distances by subtracting the mean across the 16 distances for each time point from each sensor's activity. To simplify interpretation, we also reversed their sign so that positive values reflected pattern similarity rather than distance. This procedure enabled us to compute an orientation tuning curve for each time point and trial, which expresses the extent to which pattern similarity decreases as a function of angular distance. To transform the 16-dimensional tuning curve into a one-dimensional index of decoding accuracy, we computed the cosine vector mean of each tuning curve by rescaling the cosine of the centre of each orientation bin to the range from -180 to 180 and multiplying it with the corresponding pattern similarities. The mean of the resulting 16 values served as index of decoding accuracy with positive values reflecting tuning of the EEG signal for stimulus orientation.[14,22] Decoding values were smoothed with a Gaussian kernel (SD = 24ms) for visualisation and significance testing. We tested for significance using one-sample t-tests against 0 at each time point, and corrected for multiple comparisons in time via cluster-based permutation testing using 10,000 permutations.[63] For the decoding of the memory items, cluster correction was applied for the whole trial. In contrast, for the decoding of the target and the decision variable, cluster correction was only applied for the target period (i.e., the 1800ms after target onset), because these variables could by definition only be encoded after the onset of the target stimulus. In keeping with previous work from our group,[14,29,30] decoding analyses were conducted only within posterior EEG sensors (P7, P5, P3, P1, P2, P4, P6, P8, PO7, PO5, PO3, POz, PO4, PO8, O1, Oz, O2). We selected these channels to be consistent with previous studies and because orientation signals are typically expressed most strongly in posterior regions.[14,32,64,65] Please note, however, that none of the reported results depend on this channel sub-selection (i.e., all reported effects remain significant when using all 61 channels). Moreover, please note that in experiment 2 decoding analyses were conducted separately for blocks with target presentation on the left and right side, and decoding accuracies were subsequently averaged across sides. The decision variable (absolute difference between memory item

and target) followed a discrete uniform distribution between 5.625° and 84.375°, unlike the other variables (memory items and target stimuli), which had a circular distribution. Our decoding approach therefore needed to be adjusted for decoding of the decision variable. Instead of reducing the dimensionality of the tuning curve using a cosine vector mean, we calculated the linear slope of the tuning curve.

Spatiotemporal decoding from extended stimulus evoked EEG signals

In a second step, we employed a complementary analysis approach to decode the same task variables from longer time-windows of stimulus-evoked brain activity. This approach exploits the dynamic temporal structure of event-related potentials by pooling multivariate information in time, thus capturing not only information encoded in spatial activation patterns, but also information that is encoded in the temporal unfolding of these spatial patterns. A previous study from our group has shown that such pooling over adjacent time points within trials can increase decoding sensitivity at the expense of temporal precision.[29] Consequently, the two decoding approaches capture complementary aspects of neural representations and should be considered mutually informative. In keeping with the previous study, we pooled EEG data from the 17 posterior channels within a time window of 100-400ms after target onset, thus treating single trials as discrete events. The mean activity at each sensor and time point was removed to normalise voltage fluctuations and isolate dynamic, stimulus-evoked brain states from brain activity that is stable across the chosen time period. To avoid having more data features (number of sensors X number of time points in the training data) than data samples (number of trials in the training set), we also down-sampled the signal at each sensor by a factor of 10 (i.e., from 250Hz to 25Hz). Following this preparation, we performed decoding analyses via Mahalanobis distances using the same procedure as detailed above. The 4D approach was used to decode the same task variables (target stimulus, memory items, decision variable, decision category), yielding a single vector of decoding accuracies for each subject and task variable. We evaluated the significance of results via one-sample t-test against chance level (one-tailed).

Regression analyses

We conducted a series of regression analyses to establish the behavioural relevance of the decoded memory signals and test our hypothesis that active and latent memory items are encoded in qualitatively different states. In a first step, we regressed log-transformed reaction time (log-RT) and accuracy against the trial-wise decoding strength of active and latent WM items. That is, for each trial and item, we calculated the decoding strength using the spatiotemporal approach described above, and normalised these scores by subtracting the block average to isolate across-trial fluctuations from

more sustained changes in decoding. We then used these normalised scores to predict trial-wise variance in performance with general linear models serving to predict log-RT and logistic regression models serving to predict accuracy. In a second step, we regressed log-RT and accuracy against the average block-wise decoding strength of each memory item. That is, we predicted performance on each trial (test trial) based on the average decoding strength across all remaining trials within the same block on which the respective memory item was active or based on the average decoding strength across all trials on which the respective memory item was latent (see Figure 9 for an illustration). We repeated both of these analyses with the output from cross-item decoding analyses where the pattern classifier was trained on the data from one item (active vs. latent) and tested on the data from other item. These analysis aimed to test for overlap in the neural patterns coding for active and latent WM items and to related variance in this overlap to variance in task performance.

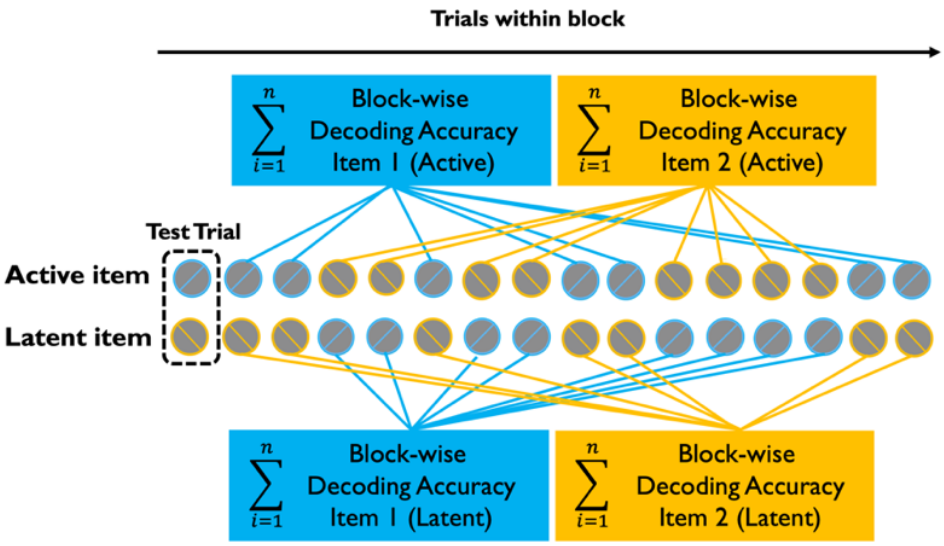


Figure 8: Illustration of the block-wise regression analysis approach. For each of the two WM items within the block, we calculated the average decoding accuracy over all trials on which the respective item task-relevant (top part) and over all the trials in which the respective item was task-irrelevant (bottom part). We then used the block-wise averages to predict trial-wise accuracy and log-RT using the average from all trials except the test trial.

Drift Diffusion Modelling

In the final set of analyses, we aimed to obtain more detailed insights into the way memory strength influenced WM-guided decisions in our task. To this end, we fit a set of drift diffusion models (DDMs) to our behavioural data and regressed variance in model parameters to variance in decoding. DDMs characterize decisions as the accumulation of noisy evidence between two competing options, whereby one option is chosen once a sufficient amount of evidence has been accumulated.[34] DDMs

decompose performance in two-alternative choice tasks into latent decision parameters based on RT distributions and choice probabilities. The most parsimonious version of the DDM has three parameters: drift rate, non-decision time, and decision threshold. Firstly, the *drift rate* reflects the quality or strength of decision-relevant information and scales negatively with categorisation difficulty (i.e., lower drift rate with more difficult discriminations). Secondly, the *non-decision time* is thought to reflect the time needed for processes that are not directly related to evidence accumulation such as the encoding of a stimulus or the execution of a response. Finally, the *decision threshold* (or boundary separation) reflects the amount of evidence that is needed to commit to a behavioural choice. This parameter is thought to be under the strategic control and to regulate speed-accuracy trade-offs (e.g., higher thresholds will result in slower but more accurate responses).

After observing that trial-wise variance in decoding of the active, but latent, WM item tracked variance in task performance (see Results section), the central aim of this analysis was to identify which decision parameter could account for this benefit. This enabled us to compare two hypotheses: (1) the matched filter hypothesis and (2) the retrieval head-start hypothesis. Under the matched filter hypothesis, variance in decoding strength should track the ease with which sensory input is interpreted and transformed into a decision variable. This view predicts that variance in decoding across trials should be positively associated with variance in drift rate (see Figure 8). By contrast, under the retrieval head-start hypothesis, variance in decoding strength should track the accessibility of the active WM item for decision making. This view predicts that variance in decoding across trials should be negatively associated with variance in non-decision time (see Figure 8).

To test these accounts, we conducted a regression analysis using a hierarchical Bayesian version of the DDM implemented in the HDDM python toolbox under version 0.6.[66] HDDM applies Markov Chain Monte Carlo sampling methods to estimate posterior distributions over DDM parameters. We fit a set of linear regression models using the patsy library to determine the relation between trial-wise fluctuation in decoding strength for both active and latent items and trial-wise fluctuation in drift rate and non-decision time respectively. We used performance accuracy (correct vs. incorrect) as response bounds rather than response identity (clockwise vs. counter-clockwise) because we aimed to establish links between decoding strength and performance, and had no reason to suspect a response bias. Separate analyses were run for drift rate and non-decision time. Each analysis contained 5000 samples from which the first 1000 were discarded as burn-in. We evaluated results statistically via the posterior distribution of the estimated regression weights. Effects were considered significant if at least 95% of the posterior probability mass was above or below zero, and in the predicted direction (positive for drift rate and negative for non-decision time; see explanation above).

Acknowledgement

This research was supported by the Wellcome Trust (award 210849/Z/18/Z to PSMK and award 201409/Z/16/Z/ to NEM), Linacre College Oxford (PSMK), University College Oxford (NEM), the Lucy Halsall Foundation (PSMK), the Research Foundation Flanders (award 12R8817N to PSMK), the Biotechnology and Biological Sciences Research Council (award BB/M010732/1 to MGS), the James S McDonnel Foundation (award 220020405 to MGS), and the NIHR Oxford Health Biomedical Research Centre. The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust (203139/Z/16/Z). The authors would like to thank Dan Bang and Annika Boldt for advice on the drift diffusion modelling, Michael Wolff for support with Figure 2, and John Duncan and Sanjay Manohar for comments on previous versions of the manuscript.

References

1. Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.
2. Oberauer, K. (2009). Design for a working memory. *Psychology of Learning and Motivation*, *51*, 45-100.
3. Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *Trends in Cognitive Sciences*, *7*(9), 415–423.
4. Goldman-Rakic, P. S. (1995). Cellular basis of working memory. *Neuron*, *14*(3), 477–485.
5. Miller, E. K., Lundqvist, M., & Bastos, A. M. (2018). Working Memory 2.0. *Neuron*, *100*(2), 463–475.
6. Stokes, M. G. (2015). 'Activity-silent' working memory in prefrontal cortex: A dynamic coding framework. *Trends in Cognitive Sciences*, *19*(7), 394–405.
7. Fuster, J.M., Alexander, G.E. (1971). Neuron activity related to short-term memory. *Science*, *173*, 653–654.
8. Murray, J. D., Bernacchia, A., Roy, N. A., Constantinidis, C., Romo, R., & Wang, X.-J. (2017). Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proceedings of the National Academy of Sciences*, *114*(2), 394–399.
9. Wimmer, K., Nykamp, D. Q., Constantinidis, C., & Compte, A. (2014). Bump attractor dynamics in prefrontal cortex explains behavioral precision in spatial working memory. *Nature Neuroscience*, *17*(3), 431–439.
10. Buschman, T. J., Denovellis, E. L., Diogo, C., Bullock, D., & Miller, E. K. (2012). Synchronous Oscillatory Neural Ensembles for Rules in the Prefrontal Cortex. *Neuron*, *76*(4), 838–846.
11. Lundqvist, M., Rose, J., Herman, P., Brincat, S. L., Buschman, T. J., & Miller, E. K. (2016). Gamma and Beta Bursts Underlie Working Memory. *Neuron*, *90*(1), 152–164.
12. Rose, N. S., LaRocque, J. J., Riggall, A. C., Gosseries, O., Starrett, M. J., Meyering, E. E., & Postle, B. R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. *Science*, *354*(6316), 1136–1139.
13. Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic Coding for Cognitive Control in Prefrontal Cortex. *Neuron*, *78*(2), 364–375.
14. Wolff, M. J., Jochim, J., Akyürek, E. G., & Stokes, M. G. (2017). Dynamic hidden states underlying working-memory-guided behavior. *Nature Neuroscience*, *20*(6), 864–871.
15. Muhle-Karbe, P. S., Duncan, J., De Baene, W., Mitchell, D. J., & Brass, M. (2017) Neural coding for instruction-based task sets in human frontoparietal and visual cortex. *Cerebral Cortex, 27*, 1891-1905.

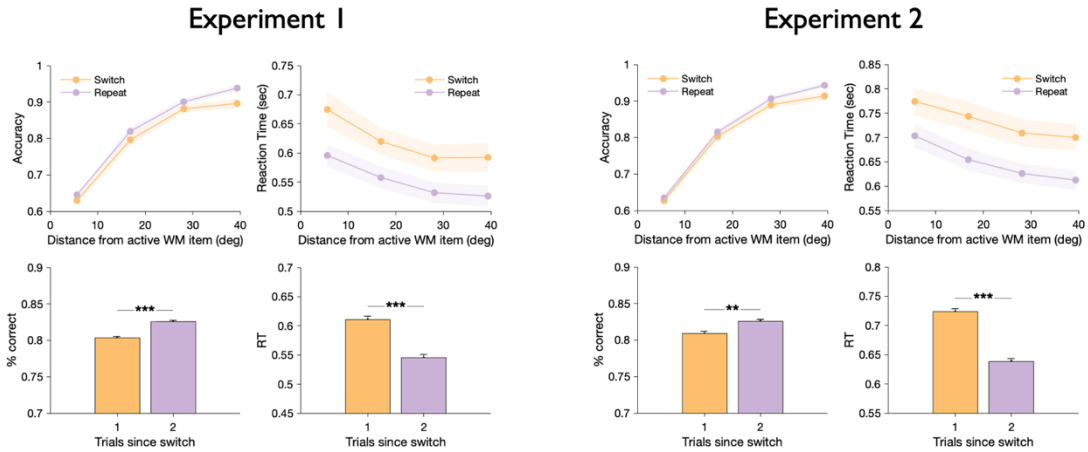
16. Duncan, J., Parr, A., Woolgar, A., Thompson, R., Bright, P., Cox, S., Bishop, S., & Nimmo-Smith, I. (2008). Goal neglect and Spearman's g: Competing parts of a complex task. *Journal of Experimental Psychology: General*, *137*(1), 131–148.
17. Bhandari, A., & Duncan, J. (2014). Goal neglect and knowledge chunking in the construction of novel behaviour. *Cognition*, *130*(1), 11–30.
18. Olivers, C. N. L., Peters, J., Houtkamp, R., & Roelfsema, P. R. (2011). Different states in visual working memory: When it guides attention and when it does not. *Trends in Cognitive Sciences*, *15*, 327-334.
19. LaRocque, J. J., Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2013). Decoding Attended Information in Short-term Memory: An EEG Study. *Journal of Cognitive Neuroscience*, *25*(1), 127–142.
20. Lewis-Peacock, J. A., Drysdale, A. T., Oberauer, K., & Postle, B. R. (2012). Neural Evidence for a Distinction between Short-term Memory and the Focus of Attention. *Journal of Cognitive Neuroscience*, *24*(1), 61–79.
21. Lewis-Peacock, J. A., Drysdale, A. T., & Postle, B. R. (2015). Neural Evidence for the Flexible Control of Mental Representations. *Cerebral Cortex*, *25*(10), 3303–3313.
22. Sprague, T. C., Ester, E. F., & Serences, J. T. (2016). Restoring Latent Visual Working Memory Representations in Human Cortex. *Neuron*, *91*(3), 694–707.
23. van Loon, A. M., Olmos-Solis, K., Fahrenfort, J. J., & Olivers, C. N. (2018). *Current and future goals are represented in opposite patterns in object-selective cortex. ELife, 7,* e38677.
24. Chun, M. M. (2011). Visual working memory as visual attention sustained internally over time. *Neuropsychologia*, *49*(6), 1407–1409.
25. Kiyonaga, A., & Egner, T. (2013). Working memory as internal attention: Toward an integrative account of internal and external selection processes. *Psychonomic Bulletin & Review*, *20*(2), 228–242.
26. LaRocque, J. J., Lewis-Peacock, J. A., & Postle, B. R. (2014). Multiple neural states of representation in short-term memory? It's a matter of attention. *Frontiers in Human Neuroscience*, *8*.
27. Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356.
28. Myers, N. E., Stokes, M. G., & Nobre, A. C. (2017). Prioritizing Information during Working Memory: Beyond Sustained Internal Attention. *Trends in Cognitive Sciences*, *21*(6), 449–461.
29. Wolff, M. J., Jochim, J., Akyürek, E. G., Buschman, T. J., & Stokes, M. G. (2019). Drifting codes within a stable coding scheme for working memory. *PLoS Biology, 18(3), e3000625.*

30. Wolff, M. J., Kandemir, G., Stokes, M. J., & Akyürek, E. G. (2019). Unimodal and bimodal access to sensory working memories by auditory and visual impulses. *Journal of Neuroscience*, *40(3),* 1194–1219.
31. Hayden, B. Y., & Gallant, J. L. (2013). Working Memory and Decision Processes in Visual Area V4. *Frontiers in Neuroscience*, *7*.
32. Myers, N. E., Rohenkohl, G., Wyart, V., Woolrich, M. W., Nobre, A. C., & Stokes, M. G. (2015). Testing sensory evidence against mnemonic templates. *ELife*, *4*, e09000.
33. Souza, A. S., Rerko, L., & Oberauer, K. (2016). Getting more from visual working memory: Retro-cues enhance retrieval and protect from visual interference. *Journal of Experimental Psychology: Human Perception and Performance*, *42 (6),* 890-910.
34. Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, *20*(4), 873–922.
35. Donoso, M., Collins, A. G. E., & Koechlin, E. (2014). Foundations of human reasoning in the prefrontal cortex. *Science*, *344*(6191), 1481–1486.
36. Holroyd, C. B., Ribas-Fernandes, J. J. F., Shahnazian, D., Silvetti, M., & Verguts, T. (2018). Human midcingulate cortex encodes distributed representations of task progress. *Proceedings of the National Academy of Sciences*, *115*(25), 6398–6403.
37. Hasson, U., Chen, J., & Honey, C. J. (2015). Hierarchical process memory: memory as an integral component of information processing. *Trends in cognitive sciences*, *19*(6), 304-313.
38. Kamiński, J. (2017). Intermediate-Term Memory as a Bridge between Working and Long-Term Memory. *The Journal of Neuroscience*, *37*(20), 5045.
39. Olsen, R. K., Nichols, E. A., Chen, J., Hunt, J. F., Glover, G. H., Gabrieli, J. D. E., & Wagner, A. D. (2009). Performance-Related Sustained and Anticipatory Activity in Human Medial Temporal Lobe during Delayed Match-to-Sample. *Journal of Neuroscience*, *29*(38), 11880–11890.
40. David, S. V., Hayden, B. Y., Mazer, J. A., & Gallant, J. L. (2008). Attention to Stimulus Features Shifts Spectral Tuning of V4 Neurons during Natural Vision. *Neuron*, *59*(3), 509–521.
41. Merrikhi, Y., Clark, K., Albarran, E., Parsa, M., Zirnsak, M., Moore, T., & Noudoost, B. (2017). Spatial working memory alters the efficacy of input to visual cortex. *Nature Communications*, *8*(1), 1-10.
42. Sugase-Miyamoto, Y., Liu, Z., Wiener, M. C., Optican, L. M., & Richmond, B. J. (2008). Short-Term Memory Trace in Rapidly Adapting Synapses of Inferior Temporal Cortex. *PLoS Computational Biology*, *4*(5), e1000073.
43. Hanks, T. D., & Summerfield, C. (2017). Perceptual Decision Making in Rodents, Monkeys, and Humans. *Neuron*, *93*(1), 15–31.

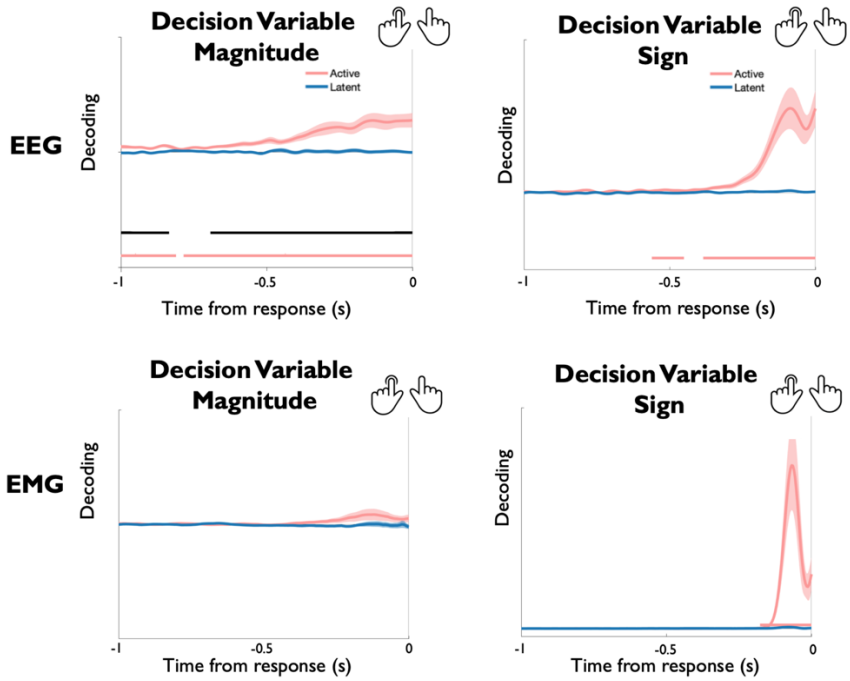
44. Pearson, B., Raskevicius, J., Bays, P. M., Pertzov, Y., & Husain, M. (2014). Working memory retrieval as a decision process. *Journal of Vision*, *14*(2), 2–2.
45. Shepherdson, P., Oberauer, K., & Souza, A. S. (2018). Working memory load and the retro-cue effect: A diffusion model account. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(2), 286–310.
46. Wagner, A. D., Shannon, B. J., Kahn, I., & Buckner, R. L. (2005). Parietal lobe contributions to episodic memory retrieval. *Trends in Cognitive Sciences*, *9*(9), 445–453.
47. Griffin, I. C., & Nobre, A. C. (2003). Orienting Attention to Locations in Internal Representations. *Journal of Cognitive Neuroscience*, *15*(8), 1176–1194.
48. Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature neuroscience*, *17*(3), 347-356.
49. Hollingworth, A., & Maxcey-Richard, A. M. (2013). Selective maintenance in visual working memory does not require sustained visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(4), 1047–1058.
50. Rerko, L., Souza, A. S., & Oberauer, K. (2014). Retro-cue benefits in working memory without sustained focal attention. *Memory & Cognition*, *42*(5), 712–728.
51. Gunseli, E., van Moorselaar, D., Meeter, M., & Olivers, C. N. L. (2015). The reliability of retro-cues determines the fate of noncued visual working memory representations. *Psychonomic Bulletin & Review*, *22*(5), 1334–1341.
52. Myers, N. E., Chekroud, S. R., Stokes, M. G., & Nobre, A. C. (2018). Benefits of flexible prioritization in working memory can arise without costs. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(3), 398–411.
53. Kamiński, J., & Rutishauser, U. (2019). Between persistently active and activity-silent frameworks: Novel vistas on the cellular basis of working memory. *Annals of the New York Academy of Sciences*, nyas.14213.
54. Manohar, S. G., Zokaei, N., Fallon, S. J., Vogels, T. P., & Husain, M. (2019). Neural mechanisms of attending to items in working memory. *Neuroscience & Biobehavioral Reviews*, *101*, 1–12.
55. Stokes, M. G., Muhle-Karbe, P. S., & Myers, N. E. (under review). Theoretical distinction between functional states in working memory and their corresponding neural states. *Visual Cognition.*
56. Yu, Q., & Postle, B. R. (2018). *Different states of priority recruit different neural codes in visual working memory* [Preprint]. Neuroscience. https://doi.org/10.1101/334920
57. Itthipuripat, S., Sprague, T. C., & Serences, J. T. (2019). Functional MRI and EEG Index Complementary Attentional Modulations. *The Journal of Neuroscience*, *39*(31), 6162–6179.

58. Geng, J. J., DiQuattro, N. E., & Helm, J. (2017). Distractor probability changes the shape of the attentional template. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(12), 1993–2007.
59. Scolari, M., & Serences, J. T. (2009). Adaptive Allocation of Attentional Gain. *Journal of Neuroscience*, *29*(38), 11933–11942.
60. Webster, R. B., Anthony, S. E., & Scheirer, W. J. (2018). Psyphy: A psychophysics driven evaluation framework for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *41*(9), 2280-2286.
61. van Ede, F., Chekroud, S. R., & Nobre, A. C. (2019). Human gaze tracks attentional focusing in memorized visual space. *Nature Human Behaviour*, *3*(5), 462–470.
62. Ledoit, O., & Wolf, M. (2004). Honey, I Shrunk the Sample Covariance Matrix. *Journal of Portfolio Management*, *30*, 110–119.
63. Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *16å4*(1), 177–190.
64. Cichy, R. M., Ramirez, F. M., & Pantazis, D. (2015). Can visual information encoded in cortical columns be decoded from magnetoencephalography data in humans? *NeuroImage*, *121*, 193–204.
65. Kok, P., Mostert, P., & de Lange, F. P. (2017). Prior expectations induce prestimulus sensory templates. *Proceedings of the National Academy of Sciences*, *114*(39), 10473–10478.
66. Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*.

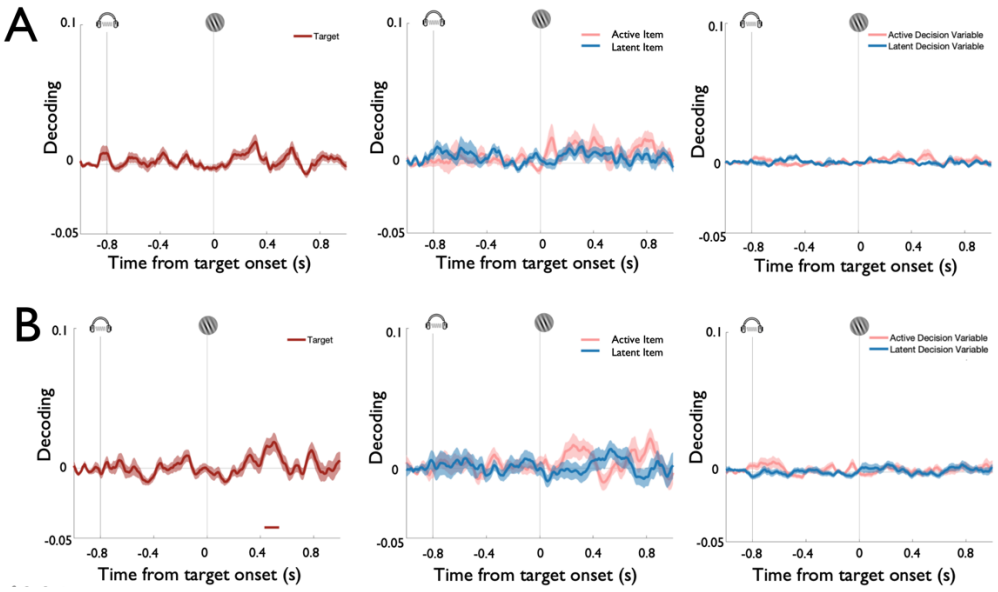
Supplementary Material



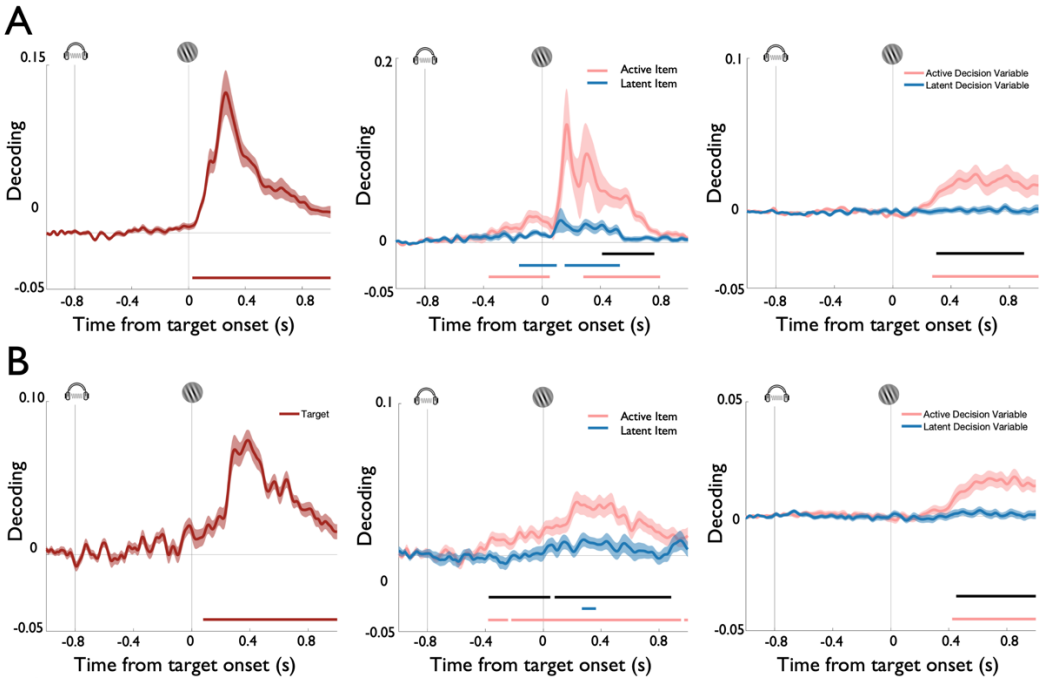
*Supplementary Figure 1: Behavioural effects of priority shifts between WM items were observed in experiment 1 (left plots) and experiment 2 (right plots) whereby performance was less accurate and slower on the first trial after a switch (bottom panels). These effects were observed across all angular distances between the target orientation and the orientation of the active WM item (upper plots). ** $p < 0.01$, *** $p < 0.005$. Shadings and error bars indicate standard error of the means.*



Supplementary Figure 2: Response-locked decoding of decision variable magnitude (i.e., the absolute angular distance between target orientation and WM item orientation – equivalent to the decision variable in the main article) and decision variable sign (i.e., a binary variable indicating whether the target was oriented clockwise or counter-clockwise relative to the WM item – equivalent to the behavioural choice). Decoding was conducted separately using the EEG channels (upper plots) and the EMG channels (lower plots) and shown separately for experiment 1 (left plots) and experiment 2 (right plots). The decision variable magnitude signal could only be decoded from EEG channels and ramped up gradually prior to overt responses. In contrast, the decision variable sign could be decoded from EEG and EMG activity and exhibited a sharp signal increase immediately preceding overt responses. These results demonstrate that the decision variable magnitude signal does not index processes of response execution.



Supplementary Figure 3: Time-resolved decoding of task variables using EOG channels in experiment 1 (A) and the high-resolution eye-tracker in experiment 2 (B). These data show that eye-channels did not contain reliable information about the task variables that were decoded from EEG sensor patterns. Each plot displays time-resolved decoding results from 200ms before cue onset until 800ms after target onset. Coloured lines represent cluster-corrected time-periods within which decoding was significantly greater than chance. Shading indicates standard error of the mean.



Supplementary Figure 4: To rule out that the signals decoded from EEG activity were not caused by subtle eye-movements we also regressed decoding time series obtained with the eye-channels against the respective time series obtained with the EEG. Plots indicate the residual EEG signal after regressing out variance that could be explained by eye-channel signals in experiment 1 (A) and experiment 2 (B). No qualitative changes were observed compared to the original EEG decoding analysis (see Figure 3 in the main text), corroborating that eye-movements did not explain the observed decoding results.