

15

Abstract

16 Working memory (WM) is important to maintain information over short time periods to
17 provide some stability in a constantly changing environment. However, brain activity is
18 inherently dynamic, raising a challenge for maintaining stable mental states. To investigate
19 the relationship between WM stability and neural dynamics, we used electroencephalography
20 to measure the neural response to impulse stimuli during a WM delay. Multivariate pattern
21 analysis revealed representations were both stable and dynamic: there was a clear difference
22 in neural states between time-specific impulse responses, reflecting dynamic changes, yet the
23 coding scheme for memorized orientations was stable. This suggests that a stable
24 subcomponent in WM enables stable maintenance within a dynamic system. A stable coding
25 scheme simplifies readout for WM-guided behaviour, whereas the low-dimensional dynamic
26 component could provide additional temporal information. Despite having a stable subspace,
27 WM is clearly not perfect – memory performance still degrades over time. Indeed, we find
28 that even within the stable coding scheme, memories drift during maintenance. When
29 averaged across trials, such drift contributes to the width of the error distribution.

30

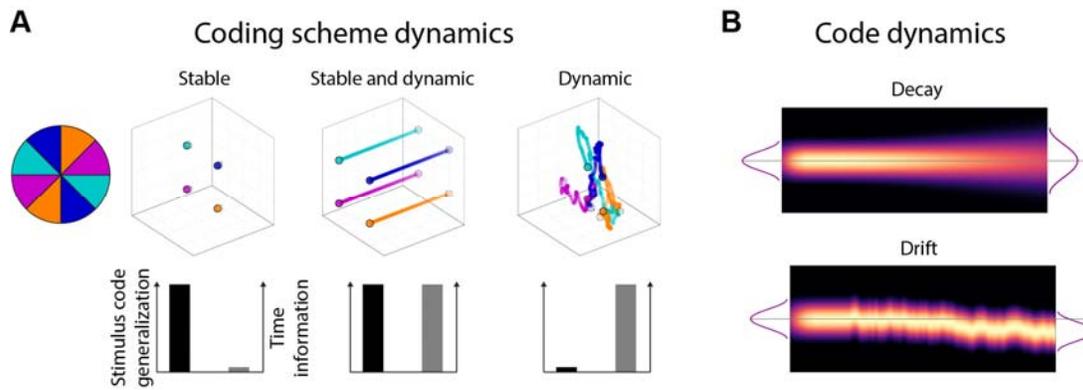
Introduction

31 Neural activity is highly dynamic, yet often we need to hold information in mind in a stable
32 state to guide ongoing behaviour. Working memory is a core cognitive function that provides
33 a stable platform for guiding behaviour according to time extended goals; however, it remains
34 unclear how such stable cognitive states emerge from a dynamic neural system.

35 At one extreme, WM could effectively pause the inherent dynamics by falling into a stable
36 attractor (e.g., 1,2). This solution has been well-studied, and provides a simple readout of
37 memory content irrespective of time (i.e., memory delay). However, more dynamic models
38 have also been suggested. For example, in a recent hybrid model, stable attractor dynamic
39 coexist with a low-dimensional, time varying component (3,4); see Fig. 1A for model
40 schematics). This permits some dynamic activity, whilst also maintaining a fixed coding
41 relationship of WM content over time (5). As in the original stable attractor model, the
42 coding scheme is stable over time, permitting easy and unambiguous WM read out by
43 downstream systems, regardless of maintenance duration (6). Finally, it is also possible to
44 maintain stable information in a richer dynamical system (e.g., 7). Although the relationship
45 between activity pattern and memory content changes over time, the representational
46 geometry could remain relatively constant (5). Such dynamics emerge naturally in a recurrent
47 network, and provide rich information about the previous input, and elapsed time (8), but
48 necessarily entail a more complex readout strategy (time-specific decoders or a high-
49 dimensional classifier that finds a high-dimensional hyperplane that separates memory
50 condition for all time points - (9)).

51 Although all models seek to account for stable WM representation, it is also important to note
52 that maintenance in WM is far from perfect. In particular, WM performance decreases over
53 time. (10), which could be ascribed to two different mechanisms (Fig. 1B). On the one hand,
54 the neural representation could simply degrade over time, either due to an overall decrease in
55 WM specific neural activity, or through a general broadening of the neural representation
56 (11). In this framework, the distribution of recall error reflects sampling from a broad
57 underlying distribution. On the other hand, the neural representation of WM content might
58 gradually drift along the feature dimension as a result of the accumulating effect of random
59 shifts due to noise (12). Even if the underlying neural representation remains sharp, variance
60 in the mean over trials results in a relative broad distribution of errors over trials.

61



62 **Figure 1.** Model predictions. (A) The relationship between the neural coding scheme of
63 orientations (colours) in WM over time. Left: A stable coding scheme within a stable
64 population. Middle: A stable coding scheme within a dynamic neural population. Right: A
65 dynamically changing coding scheme. (B) The fidelity of the population code in WM over
66 time. Top: The code decays and becomes less specific over time, leading to random errors
67 during read-out. Bottom: The code drifts along the feature dimension, leading to a still sharp,
68 but shifted code during read-out.

69 Computational modelling based on behavioural recall errors from WM tasks with varying set-
70 sizes and maintenance periods predict a drift for colours and orientations maintained in WM
71 (13,14). At the neural level, evidence for drift has been found in the neural population code in
72 monkey prefrontal cortex during a spatial WM task (15), where trial-wise shifts in the neural
73 tuning profile predicted if recall error was clockwise or counter-clockwise relative to the
74 correct location. Recently, a human fMRI study has found that delay activity reflected the
75 probe stimulus more when participants erroneously concluded that it matched the memory
76 item (16), which is consistent with the drift account.

77 Tracking these neural dynamics of non-spatial neural representations, which are not related to
78 spatial attention or motor planning, is not trivial in humans. Previously we found that the
79 presentation of a simple impulse stimulus (task-relevant visual input) presented during the
80 maintenance period of visual information in WM results in a neural response that reflects
81 non-spatial WM content (17,18). Here we extend this approach to track WM dynamics. In the
82 current study we developed a paradigm to test the stability (and/or dynamics) of WM neural
83 states and the consequence for readout by “pinging” the neural representation of orientations
84 at specific time-points during maintenance.

85 We found that the coding scheme remained stable during the maintenance period, even-
86 though maintenance time was coded in an additional low-dimensional axis. We furthermore
87 found that the neural representation of orientations drifts in WM. This was reflected in a shift
88 of the reconstructed orientation towards the end of the maintenance period that predicted
89 behaviour.

90 **Methods**

91 **Participants**

92 Twenty-six healthy adults (17 female, mean age 25.8 years, range 20-42 years) were included
93 in all analyses. Four additional participants were excluded during preprocessing due to
94 excessive eye-movements (more than 30% of trials contaminated). Participants received
95 monetary compensation (£10 an hour) for participation and gave written informed consent.
96 The experiment was approved by the Central University Research Ethics Committee of the
97 University of Oxford.

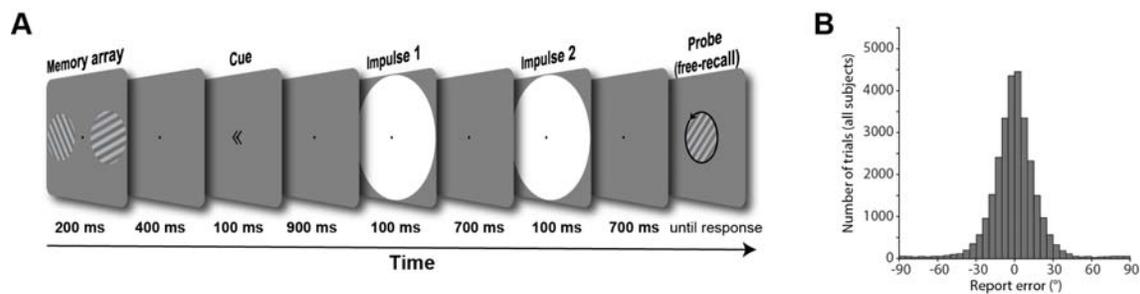
98 **Apparatus and stimuli**

99 The experimental stimuli were generated and controlled by Psychtoolbox (19), a freely
100 available Matlab extension. Visual stimuli were presented on a 23-inch (58.42 cm) screen
101 running at 100 Hz and a resolution of 1,920 by 1,080. Viewing distance was set at 64 cm. A
102 Microsoft Xbox 360 controller was used for response input by the participants.

103 A grey background (RGB = 128, 128, 128; 20.5 cd/m²) was maintained throughout the
104 experiment. A black fixation dot with a white outline (0.242°) was presented in the centre of
105 the screen throughout all trials. Memory items and the probe were sine-wave gratings
106 presented at 20% contrast, with a diameter of 8.51° and spatial frequency of 0.65 cycles per
107 degree, with randomised phase within and across trials. Memory items were presented at
108 6.08° eccentricity. The rotation of memory items and probe were randomized individually for
109 each trial. The impulse stimulus was a single white circle, with a diameter of 20.67°,
110 presented at the centre of the screen. The retro-cue was two arrowheads pointing right (>>) or
111 left (<<), and was 1.58° wide. A coloured circle (3.4°) was used for feedback. Its colour
112 depended dynamically on the precision of recall, ranging from red (more than 90 degrees
113 error) to green (0 degrees error). A pure tone also provided feedback on recall accuracy after
114 each response, ranging from 200 Hz (more than 90 degrees error) to 1,100 Hz (0 degrees
115 error).

116 Procedure

117 Participants participated in a free-recall, retro-cue visual WM task. Each trial began with the
118 fixation dot. After 1,000 ms the memory array was presented for 200 ms. After a 400 ms
119 delay, the retro-cue was presented for 100 ms, indicating which of the previously two items
120 would be tested, rendering the other item irrelevant. The first impulse stimulus was presented
121 for 100 ms, 900 ms after the offset of the retro-cue. After a delay of 700 ms, the second
122 impulse stimulus was presented for 100 ms. After another delay of 700 ms the probe was
123 presented. Participants used the left joystick on the controller with the left thumb to rotate the
124 orientation of the probe until it best reflected the memorized orientation, and confirmed their
125 answer by pressing the “x” button on the controller with the right thumb. Note that one
126 complete rotation of the joystick corresponded to 0.58 of a rotation of the probe. In
127 conjunction with the fact that the probe was randomly orientated on each trial, it was
128 impossible for participants to plan the rotation beforehand or memorize the direction of the
129 joystick instead of the orientation of the memory item. Accuracy feedback was given
130 immediately after the response where both the coloured circle and tone were presented
131 simultaneously. Each participant completed 1,100 trials in total, over a course of
132 approximately 135 minutes, including breaks. See Figure 2A for a trial schematic.



133

134 **Figure 2.** Trial schematic and behavioural results (A) Two randomly orientated grating
135 stimuli were presented laterally. A retro-cue then indicated which of those two would be
136 tested at the end of the trial. Two impulses (white circles) were serially presented in the
137 subsequent delay period. At the end of the trial a randomly oriented probe grating was
138 presented in the centre of the screen, and participants were instructed to rotate this probe until
139 it reflected the cued orientation. (B) Report errors of all trials across all subjects.

140 **EEG acquisition**

141 EEG was acquired with 61 Ag/AgCl sintered electrodes (EasyCap, Herrsching, Germany)
142 laid out according to the extended international 10–20 system and recorded at 1,000 Hz using
143 Curry 7 software (Compumedics NeuroScan, Charlotte, NC). The anterior midline frontal
144 electrodes (AFz) was used as the ground. Bipolar electrooculography (EOG) was recorded
145 from electrodes placed above and below the right eye and the temples. The impedances were
146 kept below 5 k Ω . The EEG was referenced to the right mastoid during acquisition.

147 **EEG preprocessing**

148 Offline, the EEG signal was re-referenced to the average of both mastoids, down-sampled to
149 500 Hz, and bandpass filtered (0.1 Hz high-pass and 40 Hz low-pass) using EEGLAB (20).
150 The continuous data was epoched relative to the memory array onset (-500 ms to 3,600 ms)
151 before independent component analysis (21) was applied. Components related to eye-blinks
152 were subsequently removed. The data was then epoched relative to memory array onset and
153 the two impulse onsets (0 ms to 400 ms), and trials were individually inspected. Trials with
154 saccadic eye movements, visually identified from the electrooculography, and trials with non-
155 archetypical artefacts, visually identified from the EEG, in the memory array epoch and in
156 either impulse epoch were removed from all subsequent analyses. Furthermore, trials where
157 the report error was 3 circular standard deviations from the participant's mean response error
158 were also excluded from EEG analyses to remove trials that likely represent complete
159 guesses (22). This lead to the removal of $M = 2.3\%$ ($SD = 1.2\%$) trials due to inaccurate
160 report trials, in addition to the $M = 3.52\%$ ($SD = 4.21\%$) and $M = 5\%$ ($SD = 5.2\%$) of trials
161 removed due to eye-movements and non-archetypical EEG artefacts from the memory array
162 and impulse epochs, respectively.

163 While MVPA on electrophysiological data is usually performed on each time-point
164 separately, taking advantage of the highly dynamic waveform of evoked responses in EEG by
165 pooling information multivariately over electrodes as well as time can improve decoding
166 accuracy, at the expense of temporal resolution (23,24). Since the previously reported WM-
167 dependent impulse response reflects the interaction of the WM state at the time of stimulation
168 and does not reflect continuous delay activity, we treat the impulse responses as discrete
169 events in the current study. Thus, the whole time window of interest relative to impulse
170 onsets (100 to 400 ms) from the 17 posterior channels was included in the analysis. The time
171 window was based on previous, time-resolved findings, which showed that the WM-

172 dependent neural response from a 100 ms impulse (as used in the current study) is largely
173 confined to this window (18). In the current study, instead of decoding at each time-point
174 separately, information was pooled across the whole time-window. The mean activity level
175 within each time window of each channel was first removed, thus normalizing the voltage
176 fluctuations and isolating the dynamic, impulse-evoked neural signal from more stable brain
177 states. The time-window was then down-sampled by taking the average every 10 ms, thus
178 resulting in 50 values per channel, each of which was treated as a separate dimension in the
179 subsequent multivariate analysis (850 in total). This data format was used on all subsequent
180 MVPA analyses, unless explicitly mentioned otherwise. The same approach over the same
181 time window of interest was used in our previous study (25).

182 **Orientation reconstruction**

183 We computed the mahalanobis distances as a function of orientation difference to reconstruct
184 grating orientations (18). The following procedure was performed separately for items that
185 were presented on the left and right side. Since the grating orientations were determined
186 randomly on a trial-by-trial basis and the resulting orientation distribution across trials was
187 unbalanced, we used a k-fold procedure with subsampling to ensure unbiased decoding.
188 Trials were first assigned the closest of 16 orientations (variable, see below) which were then
189 randomly split into 8 folds using stratified sampling. Using cross-validation, the train trials in
190 7 folds were used to compute the covariance matrix using a shrinkage estimator (26). The
191 number of trials of each orientation bin were equalized by randomly subsampling the
192 minimum number of trials in any bin. The subsampled trials of each angle bin were then
193 averaged. To pool information across similar orientations, the average bins were convolved
194 with a half cosine basis set raised to the 15th power (27–29). The mahalanobis distances
195 between each trial of the left-out test fold and the averaged and basis-weighted angle bins
196 were computed and mean-centred across the 16 distances to normalize. This was repeated for
197 all test and train fold combinations. To get reliable estimates, the above procedure was
198 repeated 100 times (random folds and subsamples each time), separately for eight orientation
199 spaces (0° to 168.75°, 1.40625° to 170.1563°, 2.8125° to 171.5625°, 4.2188° to 172.9688°,
200 5.625° to 174.375°, 7.0313° to 175.7813°, 8.4375° to 177.1875°, 9.8438° to 178.5938°, each
201 in steps of 11.25°). For each trial we thus obtained 800 samples for each of the 16
202 mahalanobis distances. The distances were averaged across the samples of each trial and
203 ordered as a function of orientation difference. The resulting “tuning curve” was summarized
204 into a single value (i.e., “decoding accuracy”) by computing the cosine vector mean of the

205 tuning curve (18), where a positive value suggests a higher pattern similarity between similar
206 orientations than between dissimilar orientations. The approach was the same for the
207 reanalysis of (17).

208 We also repeated the above analysis iteratively for a subset of electrodes in a searchlight
209 analysis across all 61 electrodes. In each iteration, the “current” as well as the closest two
210 neighbouring electrodes were included in the analysis (similar as in 30) The freely available
211 MATLAB extension fieldtrip (31) was used to visualise the decoding topographies. Note that
212 the topographies were flipped, such that the left represents the ipsilateral and the right the
213 contralateral side relative to stimulus presentation side.

214 **Orientation code generalization**

215 To test cross-generalization between impulses, instead of training and testing within the same
216 time-window, the train folds were taken from impulse 1, and the test fold from impulse 2, and
217 vice versa. The analysis was otherwise exactly as described above.

218 To test cross-generalization between presented locations, the classifier was similarly trained
219 on trials where the item was presented on the left, and tested on the right, and vice versa.
220 Since left and right trials were independent trial sets, cross-validation does not apply.
221 However, to ensure a balanced training set, the number of trials of each orientation bin were
222 nevertheless equalized by subsampling (as described above), and this approach was repeated
223 100 times.

224 The cross-generalization of the orientation code between impulse onsets in (17) was tested
225 with the same analyses as the location cross-generalization described in the paragraph above:
226 The classifier was trained on the early onset condition, and tested on the late-onset condition,
227 and vice versa, while making sure that the training set is balanced through random
228 subsampling.

229 **Impulse/time and location decoding**

230 To decode the difference of the evoked neural responses between impulses, we used a leave-
231 one-out approach. The mahalanobis distances between the signals from a single trial from
232 both impulse epochs and the average signal of all other trials of each impulse epoch were
233 computed. The covariance matrix was computed by concatenating the trials of each impulse
234 (excluding the left-out trial). The average difference of same impulse distances were
235 subsequently subtracted from different impulse distances, such that a positive distance

236 difference indicates more similarity between same than different impulses. To convert the
237 distance difference into trial wise decoding accuracy, positive distance difference were
238 simply converted into “hits” (1) and negative into “misses” (0). The percentage of correctly
239 classified impulses were subsequently compared to chance performance (50%).

240 The presentation side and impulse onset (in (17)) was decoded using 8-fold cross-validation,
241 where the distance difference between different and same location/onset was computed for
242 each trial, which were then converted to “hits” and “misses”.

243 **Visualization of the spatial, temporal, and orientation code**

244 To explore and visualize the relationship between the location or impulse/time code and the
245 orientation code in state space (see Fig. 1A for different predictions), we used classical
246 multidimensional scaling (MDS) of the mahalanobis distances between the average signal of
247 trials belonging to one of four orientation bins (0° to 45° , 45° to 90° , 90° to 135° , 135° to
248 180°) and location (left/right) or time (impulse 1/impulse2).

249 For the visualization of the code across impulse/time, distances were computed separately for
250 left and right trials, before taking the average. Within each orientation bin, the data of half of
251 the trials were taken from impulse 1, and the data of the other half from impulse 2
252 (determined randomly). The number of trials within each orientation of each impulse were
253 equalized through random subsampling before averaging. The mahalanobis distances
254 between both orientation and impulses were then computed using the covariance matrix
255 estimated from all trials of both impulses. This was repeated 100 times (for each side),
256 randomly subsampling and splitting trials between impulses each time and then taking the
257 average across all iterations.

258 For the visualization of the code across space, the data of each trial were first averaged across
259 impulses. The number of trials of orientation bins (same as above) of each location were
260 equalized through random subsampling. The mahalanobis distances of the average of each
261 bin within each location condition were computed using covariance estimated from all left
262 and right trials. This was repeated 100 times, before taking the average across all iterations.

263 For the code across impulse onset/time visualization of the data from (17), the same
264 procedure as in the paragraph above was used, but instead of visualizing the stimulus code
265 between locations, it was visualized between impulse onsets (-30 ms, +30 ms).

266 **Relationship between behaviour and the neural representation of the WM item**

267 We were interested if imprecise reports that are clockwise (CW) or counter-clockwise (CCW)
268 relative to the actual orientation are accompanied by a corresponding shift of the neural
269 representation in WM (see Fig. 1B for model schematics). We used two approaches to test for
270 such a shift (Fig. 5A & 6A).

271 First, the trial-wise pattern similarities as a function of orientation differences (as obtained
272 from the orientation-reconstruction approach described above) were averaged separately for
273 all CW and CCW responses (Fig. 5A). Note that CW and CCW responses were defined
274 relative to the median response error within each orientation bin. This ensures a balanced
275 proportion of all orientations in CW and CCW trials, which is necessary to obtain meaningful
276 orientation reconstructions. It furthermore removes the report bias away from cardinal angles
277 in the current experiment (Suppl. fig. 1), similar to previous reports of orientation response
278 biases (32), and thus isolates random from systematic report errors.

279 We used another approach that exaggerates the potential difference between CW and CCW
280 trials and thus might be more sensitive to detect a shift. The data was first divided into CW
281 and CCW trials using the same within orientation bin approach as described above. The
282 classifier was then trained on CW trials, and tested on CCW trials, and vice versa (Fig. 6A).
283 The orientation bins in the training set were balanced through random subsampling, and the
284 procedure was repeated 100 times. Given an actual shift in the neural representation, the shift
285 magnitude of the resulting orientation reconstruction of this method should be doubled, since
286 both the testing data and the training data (the reference point) are shifted, but in opposite
287 directions.

288 To improve orientation reconstruction from the impulse epochs, the classifier was trained on
289 the averaged trials of both impulses, but tested separately on each impulse epoch
290 individually. While training on both impulses improved orientation reconstruction, in
291 particular for the second approach where only half of the trials are used for training, the shifts
292 in orientation representations as a function of CW/CCW reports are qualitatively the same
293 when training and testing within each impulse epoch separately (Fig. 5, 6, & Suppl. fig. 3).

294 **Statistical significance testing**

295 To test for statistical significance of average decoding at the group level, the sign of the data
296 of each participant was randomly flipped with a probability of 50% 100.000 times, and the
297 resulting null-distribution was used to calculate the p value of the null hypothesis (no

298 difference, chance decoding). Note that tests of within condition decoding (within
299 presentation location, impulse/onset) were one-sided, since only positive decoding is
300 plausible in those cases, whereas tests of cross-generalization between conditions were two-
301 sided, since negative decoding is theoretically plausible in those cases. Comparisons of
302 decodability between conditions/items were also two-sided.

303 The possible shift in representation towards the response was quantified and tested for
304 statistical significance at the group level. The circular mean of the shifted average tuning
305 curve (summarized such that a positive shift reflects a shift towards the response) was tested
306 against 0. The tuning curve of each subject was flipped left to right with 0.5 probability, such
307 that a subject's positively shifted tuning curve would then be negatively shifted, before
308 computing the circular mean of the resulting tuning curve averaged over all subjects 100,000
309 times. The resulting null distribution was used to obtain the p-value by calculating the
310 proportion of permuted tuning curves with circular means more positive than the actual
311 group-level circular mean. The test obtained p-value was one-sided, since we expected the
312 shift of the neural representation of the orientation to be towards the response.

313 **Code and data availability**

314 All data and custom Matlab scripts used to generate the results and figures of this manuscript
315 will be made available upon peer-reviewed publication.

316 **Results**

317 **Item and WM content-specific evoked responses during encoding and maintenance**

318 The neural response elicited by the memory array contained parametric information about the
319 presented orientations ($p < 0.001$, one-sided; Fig. 3, left).

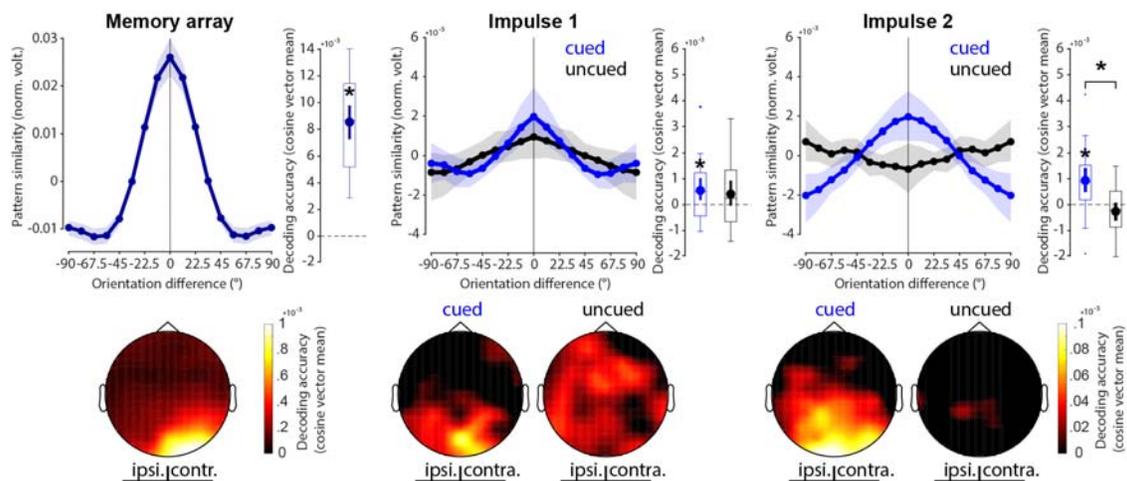
320 The first impulse response contained statistically significant information about the cued item
321 ($p = 0.008$, one sided), but not the uncued item, which failed to reach the statistical
322 significance threshold ($p = 0.057$, one-sided). The difference between cued and uncued item
323 decoding was not significant ($p = 0.694$, two-sided; Fig. 3, middle).

324 The decodability of the cued item was also significant at the second impulse response ($p <$
325 0.001 , one-sided), while it was not of the uncued item ($p = 0.919$, one-sided). Notably, the
326 decodability of the cued item was significantly higher than that of the uncued item ($p =$
327 0.002 , two-sided; Fig. 3, right).

328 Overall, these results reflect previous findings (18) in that the impulse response reflects
 329 relevant information in WM, and that no longer relevant information leave no detectable trace
 330 in the WM network.

331 The decoding topographies highlight that most of the decodable signal came from posterior
 332 electrodes during both encoding and maintenance, and is therefore likely generated by the
 333 visual cortex. Notably, while contralateral electrodes showed unsurprisingly higher item
 334 decoding during encoding, this was not the case during maintenance in either impulse
 335 response (Fig. 2C bottom row).

336 -



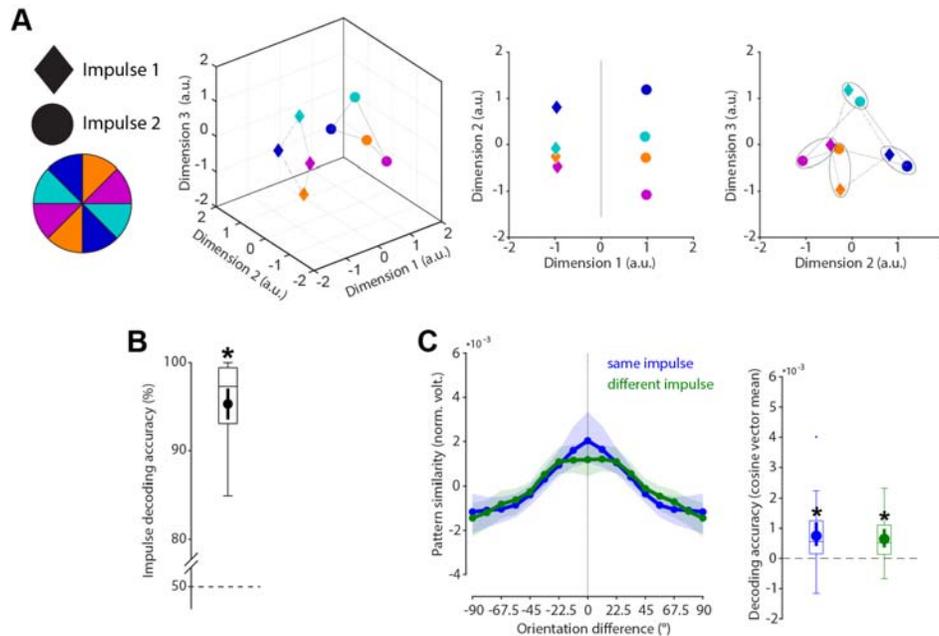
337 **Figure 3.** Decoding results. Top row: Normalized average pattern similarity (mean-centred,
 338 sign-reversed mahalanobis distance) of the evoked neural responses (100 to 400 ms relative to
 339 stimulus onset) as a function of orientation similarity, and decoding accuracy (cosine
 340 vector means of pattern similarities). Error shadings and error bars are 95 % C.I. of the mean.
 341 Centre lines of boxplots indicate the median; box outlines show 25th and 75th percentiles,
 342 and whiskers indicate 1.5x the interquartile range. Extreme values are shown separately
 343 (dots). Asterisks indicate significant decoding accuracies ($p < 0.05$, one-sided) or differences
 344 ($p < 0.05$, two-sided). Bottom row: Decoding topographies of the searchlight analysis.

346 Stable WM coding scheme in time

347 The relationship between orientations and impulses/time is visualized in state-space through
 348 MDS (Fig. 4A). While the first dimension clearly differentiates between impulses, the second
 349 and third dimensions code the circular geometry of orientations in both impulses, suggesting
 350 that while the impulse responses are different between impulses, the orientation coding
 351 schemes revealed by the impulse are the same. This is corroborated by significant decoding

352 accuracy of the impulse ($p < 0.001$, one-sided; Fig. 4B) on the one hand, but also significant
353 cross-generalization of the orientation code between impulses ($p < 0.001$, two-sided), which
354 was not significantly different from same-impulse orientation decoding ($p = 0.581$, two-
355 sided; Fig. 4C).

356



357

358 **Figure 4.** Cross-generalization of coding scheme between impulses. (A) Visualization of
359 orientation and impulse code in state-space. The first dimension discriminates between
360 impulses. The second and third dimensions code the orientation space in both impulses. (B)
361 Trial-wise accuracy (%) of impulse decoding. (C) Orientation decoding within each impulse
362 (blue) and orientation code cross-generalization between impulses (green). Error shadings
363 and error bars are 95 % C.I. of the mean. Centre lines of boxplots indicate the median; box
364 outlines show 25th and 75th percentiles, and whiskers indicate 1.5x the interquartile range.
365 Extreme values are shown separately (dots). Asterisks indicate significant decoding
366 accuracies or cross-generalization ($p < 0.05$).

367 It is not possible to conclude whether the difference between impulses is due to a neural
368 network that changes during the maintenance period over time, due to different stimulation
369 histories at the time of perturbation (i.e., the first impulse always preceded the second
370 impulse), or due to different WM operations at each impulse event (e.g. item selection at
371 impulse 1, response preparation at impulse 2).

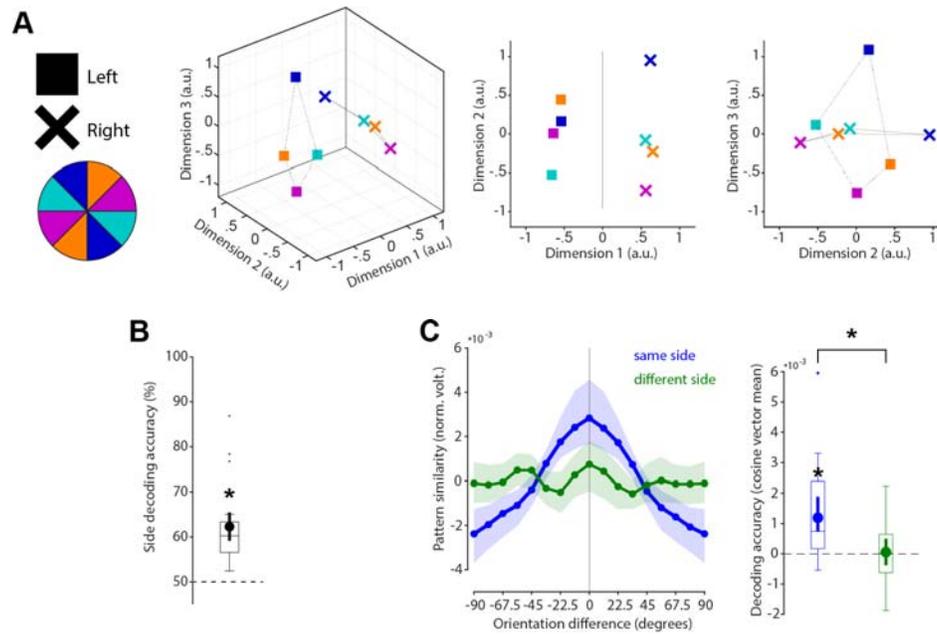
372 To rule out that the difference in impulse response reported above is not only due to
373 difference in stimulation history and changing WM operations, but also due to temporal
374 coding in the WM network, we reanalysed previously published data where a single impulse
375 stimulus was presented either 1,170 or 1,230 ms after the presentation of a single memory
376 item (17). The findings largely replicate the results reported above: State-space visualization
377 of impulse-onset and orientations shows the same circular geometry of the orientations at
378 each impulse onset, while also highlighting a separation of impulse onsets in state-space
379 (Suppl. fig. 2A). Decoding impulse-onset was significantly than from chance ($p = 0.005$, one-
380 sided; Suppl. fig. 2B). Cross-generalization of the orientation code between impulse-onsets
381 was significant ($p < 0.001$, two-sided), and did not significantly differ from decoding the
382 memorized orientation within the same impulse-onset ($p = 0.244$, two-sided; Suppl. fig. 2C).

383 Overall, the results of the current study, as well as the reanalyses of (17) provide evidence for
384 a low-dimensional change over time, that can be revealed by perturbing the WM network at
385 different time-points (as predicted in (33)), while at the same time providing evidence for a
386 temporally stable coding scheme of WM content (3,4).

387 **Specific WM coding scheme in space**

388 As a counterpart to the stable coding scheme in time reported above, we explicitly tested if
389 the coding scheme is location specific (i.e., dependent on the previous presentation location
390 of the cued orientation). State-space visualization of cued item location and orientations
391 shows a clear separation between locations and no overlap in orientation coding between
392 locations (Fig. 5A). The cued location was significantly decodable from the impulse
393 responses ($p < 0.001$, one-sided; Fig. 5B). Cross-generalization of the orientation coding
394 scheme between cued item locations was not significant ($p = 0.403$, two-sided), and
395 significantly lower than same side orientation decoding ($p = 0.009$, two-sided; Fig. 5C).
396 These results reflect previous reports of spatially specific WM codes, even when location is
397 no longer relevant (34).

398



399

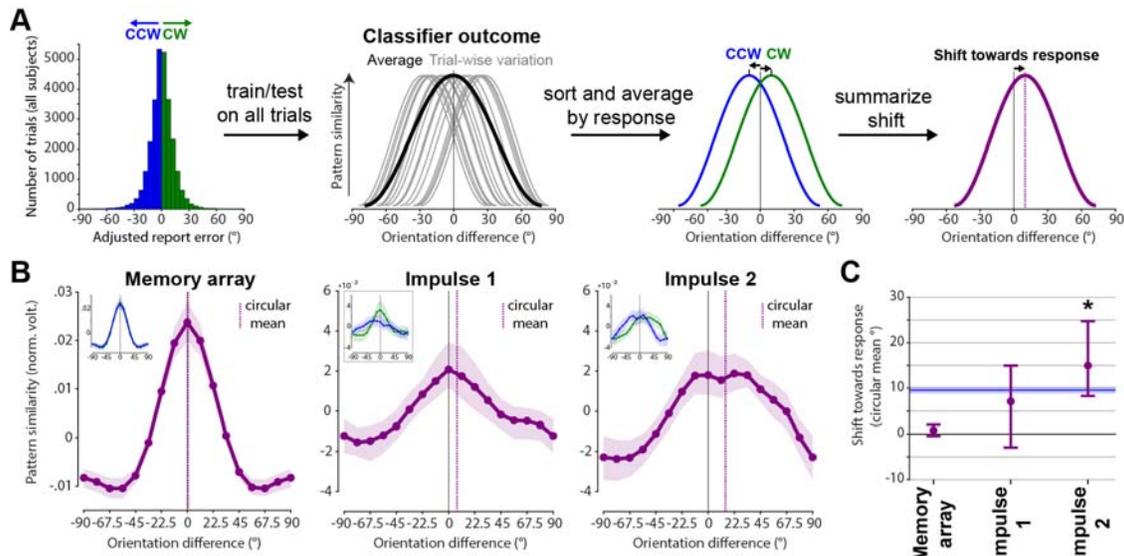
400 **Figure 5.** No cross-generalization of coding scheme between cued item locations during
 401 impulse responses **(A)** Visualization of orientation and item location code in state-space. The
 402 first dimension discriminates between item locations. The first and second dimensions code
 403 the orientation space, separately for WM items previously presented on the left or right side.
 404 **(B)** Trial-wise accuracy (%) of item location decoding. **(C)** Orientation decoding within each
 405 item location (blue) and orientation code cross-generalizing between different item locations
 406 (green). Error shadings and error bars are 95 % C.I. of the mean. Centre lines of boxplots
 407 indicate the median; box outlines show 25th and 75th percentiles, and whiskers indicate 1.5x
 408 the interquartile range. Extreme values are shown separately (dots). Asterisks indicate
 409 significant decoding accuracies and differences ($p < 0.05$).

410 **Drifting WM code**

411 The first approach to test for a possible shift of the neural representation towards the response
 412 averaged the trial-wise orientation tuning curves obtained from the cross-validated orientation
 413 reconstruction on all trials (see Methods and Fig. 6A).

414 No significant shift towards the response was evident during encoding/memory array
 415 presentation ($p = 0.117$, one-sided; Fig. 6B & C, left). No evidence for such a shift was found
 416 at impulse 1/early maintenance either ($p = 0.07$, one-sided; Fig. 6B & C, middle). However,
 417 the orientation tuning curve was significantly shifted towards the response at impulse 2/late
 418 maintenance ($p < 0.001$, one-sided; Fig. 6B & C, right).

419



420

421

Figure 6. Response-dependent averaging of trial-wise tuning curves demonstrates drift.

422

Schematic and results. **(A)** Testing for shift towards response by averaging trial-wise tuning

423

curves by CCW/CW responses. **(B)** Results of schematised approach in A. Orientation tuning

424

curves averaged by response such that a right-ward shift reflects a shift towards the response

425

(purple) at each event. Purple vertical lines show circular means of the tuning curves. Insets

426

show orientation tuning curves for CCW (blue) and CW (green) responses separately. Error

427

shadings are 95 % C. I. of the mean. **(C)** Group-level shifts towards the response (circular

428

mean) of each response-dependent tuning curve. Error-bars are 95 % C. I. of the mean. The

429

blue line and shading indicates the mean and 95 % C.I. of the absolute, bias-adjusted

430

behavioural response deviation.

431

The second approach to test for a possible shift of the neural representation towards the

432

response may be more sensitive since it trains the orientation classifier only on CCW trials,

433

and tests it on CW trials, and vice versa (see Methods and Fig. 7A), thus increasing any

434

response related shift by a factor of two.

435

This approach yielded similar results as the previous approach, though the shift magnitudes

436

are indeed larger. Neither the memory array presentation/encoding, nor impulse 1/early

437

maintenance showed a significant shift towards the response ($p = 0.124$, $p = 0.104$,

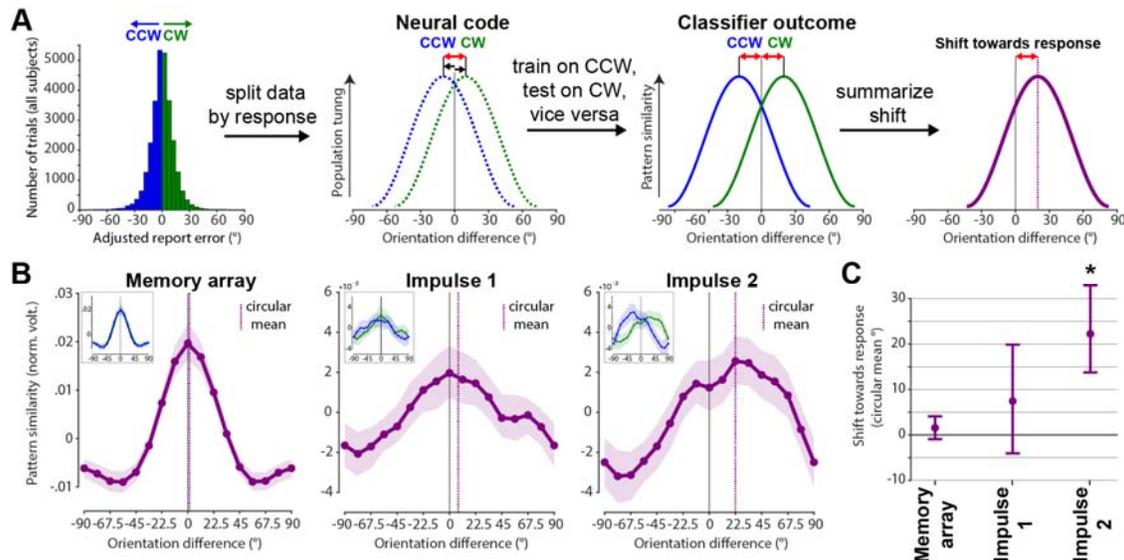
438

respectively, one-sided; Fig. 7, left & middle), while impulse 2/late maintenance did ($p <$

439

0.001 , one-sided; Fig. 7, right).

440



441

442

Figure 7. Response-dependent training and testing demonstrates drift. Schematic and results.

443

(A) Testing for shift towards response by first splitting the neuroimaging data into CW and

444

CCW data sets, and training on CW trials and testing on CCW trials, and vice versa. Given an

445

actual shift, the shift of the resulting orientation reconstruction will be doubled, since training

446

and testing data are shifted in opposite directions. (B) Results of schematised approach in A.

447

Average orientation tuning curves such that a rightward shift reflects a shift towards the

448

response (purple) at each event. Purple vertical lines show circular means of the tuning

449

curves. Insets show orientation tuning curves for CCW (blue) and CW (green) responses

450

separately. Error shadings are 95 % C. I. of the mean. (C) Group-level shifts towards the

451

response (circular mean) of each response-dependent tuning curve. Error-bars are 95 % C. I.

452

of the mean.

453

Note the reported results of shifts during impulse presentations were obtained by training the

454

classifier on both impulses, but testing it on each impulse separately. This was done to

455

improve power (as explained in Methods). This improved orientation reconstruction

456

particularly for the latter shift-analysis where the classifier is trained on only half the trials

457

(CW trials only or CCW trials only). However, the same analyses based on training (and

458

testing) within each impulse epoch separately yielded qualitatively similar results (no

459

significant shifts at impulse 1 in either approach, significant shifts at impulse 2 in both

460

approaches; Suppl. fig. 3).

461

Discussion

462 In the present study, we investigated the neural dynamics of WM by probing the coding
463 scheme over time, as well as drift in the actual memories. The neural response to impulse
464 stimuli in this non-spatial WM paradigm enabled us to show that the coding scheme of
465 parametric visual feature (i.e., orientation) in WM remained stable during maintenance,
466 reflected in the significant cross-generalization of the orientation decoding between early and
467 late impulses (Fig. 4). However, memories drift within this stable coding scheme, leading to a
468 bias in memories (Figs. 6 and 7).

469 This is consistent with previous reports of a stable subspace for WM maintenance (4,5), and
470 provides evidence for a time-invariant coding scheme for orientations maintained in WM.
471 However, more dynamic schemes have also been reported. For example, during the early
472 transition between encoding and maintenance (35,36). At the extreme end, some have
473 proposed that WM could be maintained in a dynamical system, where activity evolves along
474 a complex trajectory in neural state space (e.g. 37). Although this complicates readout
475 (discrimination boundaries at one time point do not generalise to other time-points), such
476 coding schemes evolve naturally from recurrent neural networks. Moreover, such dynamics
477 also provide additional information, such as elapsed time. In the current study, we find
478 evidence for a hybrid model (3,4): stable decoding of WM, despite dynamic activity over
479 time.

480 Specifically, while there was no cost of cross-generalizing the orientation code between
481 impulses, there was nevertheless a clear difference in the neural pattern between them,
482 suggesting that a separate dynamic neural pattern codes the passage of time. A reanalysis of
483 the data of a previously published study (17) confirmed these results, suggesting that the low-
484 dimensional dynamics code for time per se (rather than impulse number). The significant
485 decodability of impulse onset shows that the WM network changes during the maintenance
486 even within 60 ms, resulting in distinct neural impulse responses at different time-points
487 providing evidence for a neural time-code. Importantly, the low-dimensional representation
488 of elapsed time is orthogonal to the mnemonic subspace, allowing WM representations to be
489 stable. This hybrid of stable and dynamic representations may emerge from interactions
490 between dynamic recurrent neural networks and stable sensory representations (3).

491 Our index of WM-related neural activity was based on an impulse response approach that we
492 previous developed to measure WM-related changes in the functional state of the system,

493 including ‘activity-silent’ WM states (17,18,38,39). For example, activity states during
494 encoding could result in a neural trace in the WM network through short-term synaptic
495 plasticity resulting in a stable code for maintenance, whereas the time-dimension could be
496 represented in its gradual fading (33,40–42). The stable WM-content coding scheme could
497 also be achieved by low-level activity states that self-sustain a stable code through recurrent
498 connections, a key feature of attractor models of WM (1,43), while dynamic activity patterns
499 are coded in an orthogonal subspace that represents time. While we did not explicitly
500 consider tonic delay activity, it is nonetheless possible that the impulse responses also reflect
501 non-linear interactions with low-level, persistent activity states that are otherwise difficult to
502 measure with EEG. Therefore, we cannot rule out a contribution of persistent activity in the
503 stable coding scheme observed here.

504 We also found evidence that the orientation code itself drifts along the orientation dimension,
505 predicting recall errors. While there was no bias in the neural orientation representation at
506 either encoding or early maintenance, the second impulse towards the end of the maintenance
507 period revealed a code that was shifted towards the direction of response error. This pattern
508 of results is consistent with the drift account of WM, where neural noise leads to an
509 accumulation of error during maintenance, resulting in a still sharp, but shifted (i.e. slightly
510 wrong) neural representation of the maintained information (1,14). While previous
511 neurophysiological recordings from monkey PFC found evidence for drift for spatial
512 information (15), we could demonstrate a shifting representation that more faithfully
513 represents non-spatial WM content that is unrelated to sustained spatial attention or motor
514 preparation, by using lateralized orientations in the present study.

515 Bump attractors have been proposed as an ideal neural mechanism for the maintenance of
516 continuous representations (i.e. space, orientation, colour), where a specific feature is
517 represented by the persistent activity “bump” of the neural population at the feature’s location
518 along the network’s continuous feature space. Neural noise randomly shifts this bump along
519 the feature dimension, while inhibitory and excitatory connections maintain the same overall
520 level of activity and shape of the neural network (44,45). Random walk along the feature
521 dimension is thus a fundamental property of bump attractors, and has been found to explain
522 neurophysiological findings (15). Typically, this is considered within the framework of
523 persistent working memory, however transient bursts of activity could also follow similar
524 attractor dynamics (46,47). For example, the temporary connectivity changes of the
525 memorized WM item may indeed slowly dissolve and become coarser, periodic activity

526 bursts may keep this to a minimum, by periodically reinstating a sharp representation.
527 However, since this refreshing depends on the read-out of a coarse representation, the
528 resulting representation may be slightly wrong and thus shifted. This interplay between
529 decaying silent WM-states that are readout and refreshed by active WM-states also predicts a
530 drifting WM code, without depending on an unbroken chain of persistent neural activity.

531 Moreover, the representational drift does not necessarily have to be random. Modelling of
532 report errors in a free recall colour WM task suggests that an increase of report errors over
533 time may be due to separable attractor dynamics, with a systematic drift towards stable colour
534 representations, resulting in a clustering of reports around specific colour values, in addition
535 to random drift elicited by neural noise (13). The report bias of oblique orientations seen in
536 the present study could be explained by a similar drift towards specific orientations, which
537 would predict an increase of report bias for longer retention periods. However, clear
538 behavioural evidence for such an increase in systemic report errors of orientations is lacking
539 (10). In the present study we isolated random from systematic errors, both as a
540 methodological necessity, but also to be able to conclude that any observed shift is due to
541 random errors. Thus, while a systematic drift towards specific orientations might be possible,
542 the shift in representation reported here is unrelated to it.

543 Our results suggest that maintenance in WM is dynamic, although the fundamental coding
544 scheme remains stable over time. Low-dimensional dynamics could provide a valuable
545 readout of elapsed time, whilst allowing for a time-general readout scheme for the WM
546 content. We also show that drift within this stable coding scheme could explain loss of
547 memory precision over time.

548

Acknowledgments

549 This research was in part funded by a James S. McDonnell Foundation Scholar Award
550 (220020405) and an ESRC grant (ES/S015477/1) to MGS, and by the NIHR Oxford Health
551 Biomedical Research Centre. The Wellcome Centre for Integrative Neuroimaging is
552 supported by core funding from the Wellcome Trust (203139/Z/16/Z). The views expressed
553 are those of the authors and not necessarily those of the National Health Service, the National
554 Institute for Health Research or the Department of Health. EGA is in part funded by an Open
555 Research Area grant (464.18.114). We would like to thank N.E. Myers and D. Trübutschek
556 for helpful comments.

557

References

- 558 1. Compte A, Brunel N, Goldman-Rakic PS, Wang X-J. Synaptic Mechanisms and
559 Network Dynamics Underlying Spatial Working Memory in a Cortical Network Model.
560 Cereb Cortex. 2000 Sep 1;10(9):910–23.
- 561 2. Wang X-J. Synaptic reverberation underlying mnemonic persistent activity. Trends in
562 Neurosciences. 2001 Aug 1;24(8):455–63.
- 563 3. Bouchacourt F, Buschman TJ. A Flexible Model of Working Memory. Neuron. 2019 Jul
564 3;103(1):147-160.e8.
- 565 4. Murray JD, Bernacchia A, Roy NA, Constantinidis C, Romo R, Wang X-J. Stable
566 population coding for working memory coexists with heterogeneous neural dynamics in
567 prefrontal cortex. PNAS. 2017 Jan 10;114(2):394–9.
- 568 5. Spaak E, Watanabe K, Funahashi S, Stokes MG. Stable and Dynamic Coding for
569 Working Memory in Primate Prefrontal Cortex. J Neurosci. 2017 Jul 5;37(27):6503–16.
- 570 6. Cueva CJ, Marcos E, Saez A, Genovesio A, Jazayeri M, Romo R, et al. Low
571 dimensional dynamics for working memory and time encoding. bioRxiv. 2019 Jan
572 31;504936.
- 573 7. Barak O, Sussillo D, Romo R, Tsodyks M, Abbott LF. From fixed points to chaos:
574 Three models of delayed discrimination. Progress in Neurobiology. 2013 Apr
575 1;103:214–22.
- 576 8. Romo R, Brody CD, Hernández A, Lemus L. Neuronal correlates of parametric working
577 memory in the prefrontal cortex. Nature. 1999 Jun;399(6735):470.
- 578 9. Druckmann S, Chklovskii DB. Neuronal Circuits Underlying Persistent Representations
579 Despite Time Varying Activity. Current Biology. 2012 Nov 20;22(22):2095–103.
- 580 10. Rademaker RL, Park YE, Sack AT, Tong F. Evidence of gradual loss of precision for
581 simple features and complex objects in visual working memory. Journal of
582 Experimental Psychology: Human Perception and Performance. 2018;44(6):925–40.

- 583 11. Barrouillet P, Camos V. Developmental Increase in Working Memory Span: Resource
584 Sharing or Temporal Decay? *Journal of Memory and Language*. 2001 Jul 1;45(1):1–20.
- 585 12. Kinchla RA, Smyzer F. A diffusion model of perceptual memory. *Perception &*
586 *Psychophysics*. 1967 Jun 1;2(6):219–29.
- 587 13. Panichello MF, DePasquale B, Pillow JW, Buschman TJ. Error-correcting dynamics in
588 visual working memory. *Nature Communications*. in press;
- 589 14. Schneegans S, Bays PM. Drift in Neural Population Activity Causes Working Memory
590 to Deteriorate Over Time. *J Neurosci*. 2018 May 23;38(21):4859–69.
- 591 15. Wimmer K, Nykamp DQ, Constantinidis C, Compte A. Bump attractor dynamics in
592 prefrontal cortex explains behavioral precision in spatial working memory. *Nat*
593 *Neurosci*. 2014 Mar;17(3):431–9.
- 594 16. Lim PC, Ward EJ, Vickery TJ, Johnson MR. Not-so-working Memory: Drift in
595 Functional Magnetic Resonance Imaging Pattern Representations during Maintenance
596 Predicts Errors in a Visual Working Memory Task. *Journal of Cognitive Neuroscience*.
597 2019 May 21;1–15.
- 598 17. Wolff MJ, Ding J, Myers NE, Stokes MG. Revealing hidden states in visual working
599 memory using electroencephalography. *Front Syst Neurosci*. 2015
- 600 18. Wolff MJ, Jochim J, Akyürek EG, Stokes MG. Dynamic hidden states underlying
601 working-memory-guided behavior. *Nature Neuroscience*. 2017 Jun;20(6):864–71.
- 602 19. Kleiner M. Visual stimulus timing precision in Psychtoolbox-3: Tests, pitfalls solutions
603 [Internet]. 2010 [cited 2016 Jul 12]. Available from: [http://www.neuroschool-tuebingen-](http://www.neuroschool-tuebingen-ena.de/fileadmin/user_upload/Dokumente/neuroscience/AbstractbookNeNa2010u.pdf)
604 [ena.de/fileadmin/user_upload/Dokumente/neuroscience/AbstractbookNeNa2010u.pdf](http://www.neuroschool-tuebingen-ena.de/fileadmin/user_upload/Dokumente/neuroscience/AbstractbookNeNa2010u.pdf)
- 605 20. Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial
606 EEG dynamics including independent component analysis. *Journal of Neuroscience*
607 *Methods*. 2004 Mar;134(1):9–21.
- 608 21. Hyvarinen A. Fast and robust fixed-point algorithms for independent component
609 analysis. *IEEE Transactions on Neural Networks*. 1999 May;10(3):626–34.

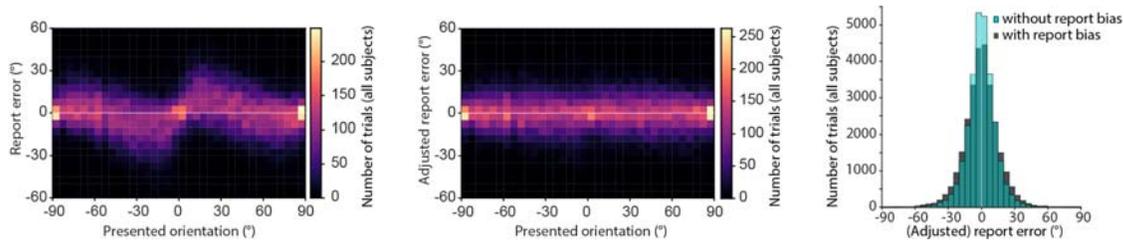
- 610 22. Fritsche M, Mostert P, de Lange FP. Opposite Effects of Recent History on Perception
611 and Decision. *Current Biology*. 2017 Feb 20;27(4):590–5.
- 612 23. Grootswagers T, Wardle SG, Carlson TA. Decoding Dynamic Brain Patterns from
613 Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series
614 Neuroimaging Data. *J Cogn Neurosci*. 2017 Apr;29(4):677–97.
- 615 24. Nemrodov D, Niemeier M, Patel A, Nestor A. The Neural Dynamics of Facial Identity
616 Processing: Insights from EEG-Based Pattern Analysis and Image Reconstruction.
617 *eNeuro*. 2018 Jan 1;5(1):ENEURO.0358-17.2018.
- 618 25. Wolff MJ, Kandemir G, Stokes MG, Akyurek EG. Impulse responses reveal unimodal
619 and bimodal access to visual and auditory working memory. *bioRxiv*. 2019 Apr
620 30;623835.
- 621 26. Ledoit O, Wolf M. Honey, I shrunk the sample covariance matrix. *The Journal of*
622 *Portfolio Management*. 2004;30(4):110–9.
- 623 27. Myers NE, Rohenkohl G, Wyart V, Woolrich MW, Nobre AC, Stokes MG. Testing
624 sensory evidence against mnemonic templates. *eLife*. 2015 Dec 14;4:e09000.
- 625 28. Serences JT, Saproo S. Computational advances towards linking BOLD and behavior.
626 *Neuropsychologia*. 2012 Mar 1;50(4):435–46.
- 627 29. Brouwer GJ, Heeger DJ. Decoding and reconstructing color from responses in human
628 visual cortex. *J Neurosci*. 2009 Nov 4;29(44):13992–4003.
- 629 30. Ede F van, Chekroud SR, Stokes MG, Nobre AC. Concurrent visual and motor selection
630 during visual working memory guided action. *Nature Neuroscience*. 2019
631 Mar;22(3):477.
- 632 31. Oostenveld R, Fries P, Maris E, Schoffelen J-M, Oostenveld R, Fries P, et al. FieldTrip:
633 Open Source Software for Advanced Analysis of MEG, EEG, and Invasive
634 Electrophysiological Data, FieldTrip: Open Source Software for Advanced Analysis of
635 MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and*
636 *Neuroscience, Computational Intelligence and Neuroscience*. 2010 Dec 23;2011,
637 2011:e156869.

- 638 32. Pratte MS, Park YE, Rademaker RL, Tong F. Accounting for stimulus-specific variation
639 in precision reveals a discrete capacity limit in visual working memory. *Journal of*
640 *Experimental Psychology: Human Perception and Performance*. 2017;43(1):6–17.
- 641 33. Buonomano DV, Maass W. State-dependent computations: spatiotemporal processing in
642 cortical networks. *Nature Reviews Neuroscience*. 2009 Feb;10(2):113–25.
- 643 34. Pratte MS, Tong F. Spatial specificity of working memory representations in the early
644 visual cortex. *Journal of Vision*. 2014 Mar 1;14(3):22–22.
- 645 35. Wasmuht DF, Spaak E, Buschman TJ, Miller EK, Stokes MG. Intrinsic neuronal
646 dynamics predict distinct functional roles during working memory. *Nature*
647 *Communications*. 2018 Aug 29;9(1):3499.
- 648 36. Cavanagh SE, Towers JP, Wallis JD, Hunt LT, Kennerley SW. Reconciling persistent
649 and dynamic hypotheses of working memory coding in prefrontal cortex. *Nature*
650 *Communications*. 2018 Aug 29;9(1):3498.
- 651 37. Maass W, Natschläger T, Markram H. Real-time computing without stable states: a new
652 framework for neural computation based on perturbations. *Neural Comput*. 2002
653 Nov;14(11):2531–60.
- 654 38. Stokes MG. ‘Activity-silent’ working memory in prefrontal cortex: a dynamic coding
655 framework. *Trends in Cognitive Sciences*. 2015 Jul;19(7):394–405.
- 656 39. Masse NY, Yang GR, Song HF, Wang X-J, Freedman DJ. Circuit mechanisms for the
657 maintenance and manipulation of information in working memory. *Nature*
658 *Neuroscience*. 2019 Jun 10;1.
- 659 40. Nikolić D, Häusler, Stefan, Singer, Wolf, Maass, Wolfgang. Temporal dynamics of
660 information content carried by neurons in the primary visual cortex. *Advances in Neural*
661 *Information Processing Systems*. 2007;19:1041–1048.
- 662 41. Nikolić D, Häusler S, Singer W, Maass W. Distributed Fading Memory for Stimulus
663 Properties in the Primary Visual Cortex. *PLOS Biol*. 2009 Dec 22;7(12):e1000260.
- 664 42. Zucker RS, Regehr WG. Short-Term Synaptic Plasticity. *Annu Rev Physiol*. 2002 Mar
665 1;64(1):355–405.

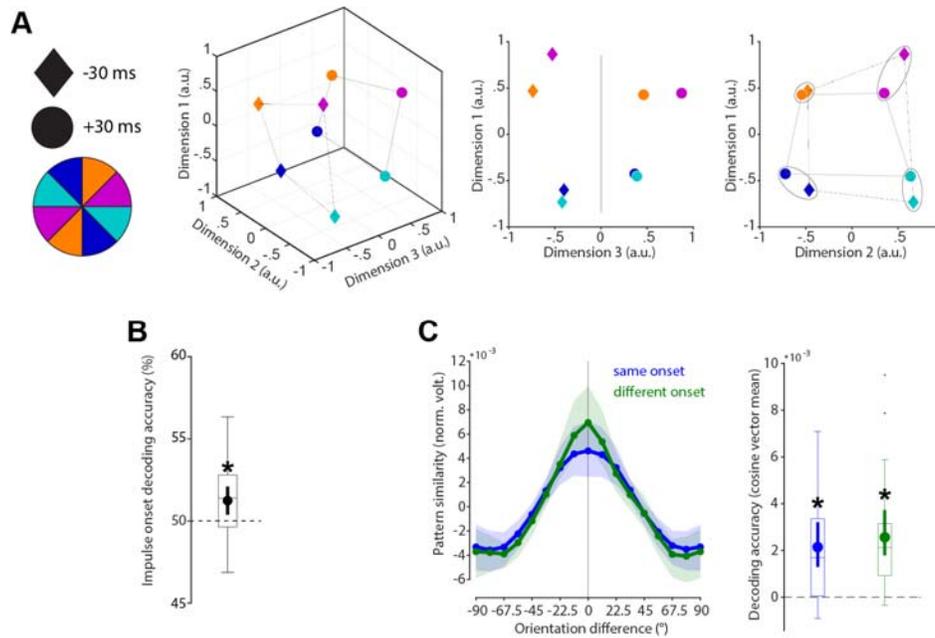
- 666 43. Chaudhuri R, Fiete I. Computational principles of memory. *Nature Neuroscience*. 2016
667 Mar;19(3):394–403.
- 668 44. Amari S. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biol*
669 *Cybern*. 1977 Jun 1;27(2):77–87.
- 670 45. Brody CD, Romo R, Kepecs A. Basic mechanisms for graded persistent activity:
671 discrete attractors, continuous attractors, and dynamic representations. *Current Opinion*
672 *in Neurobiology*. 2003 Apr 1;13(2):204–11.
- 673 46. Lundqvist M, Herman P, Lansner A. Theta and Gamma Power Increases and
674 Alpha/Beta Power Decreases with Memory Load in an Attractor Network Model.
675 *Journal of Cognitive Neuroscience*. 2011 Mar 31;23(10):3008–20.
- 676 47. Mongillo G, Barak O, Tsodyks M. Synaptic Theory of Working Memory. *Science*. 2008
677 Mar 14;319(5869):1543–6.
- 678

679

Appendix

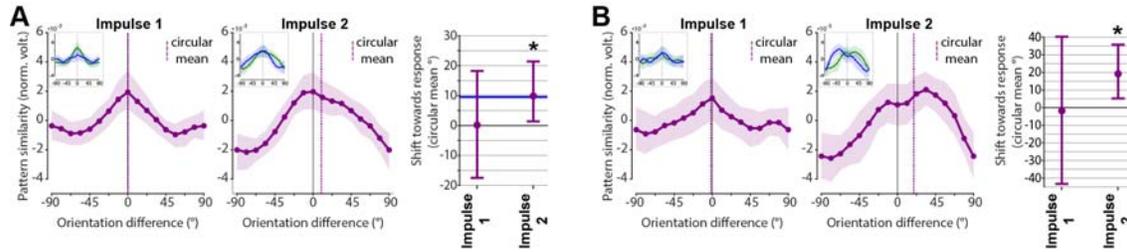


680 **Supplemental figure 1.** Report-bias of orientations. Participants showed a bias, exaggerating
681 the tilt of oblique orientations, manifesting itself as a repulsion from the cardinal axes (0 and
682 90 degrees; *left*), similar to previous reports (32). To ensure an unbiased estimate of a
683 possible shift in our analysis, and to isolate random from systematic errors, the report bias
684 was removed by subtracting the median error within 11.25 degree orientation bins (*middle*).
685 By removing orientation-specific error, the resulting error distribution is narrower (*right*).
686 Clockwise and counter-clockwise reports were defined as positive and negative reports
687 relative to this “adjusted”, unbiased, report error.



688

689 **Supplemental figure 2.** Cross-generalization of coding scheme between impulse onsets in
 690 reanalyses of (17). **(A)** Visualization of orientation and impulse-onset code in state-space.
 691 The third dimension discriminates between impulse-onsets. The first and second dimensions
 692 code the orientation space in both impulses. **(B)** Trial-wise accuracy (%) of impulse-onset
 693 decoding. **(C)** Orientation decoding within each impulse-onset (blue) and orientation code
 694 cross-generalizing between impulse-onsets (green). Error shadings and error bars are 95 %
 695 C.I. of the mean. Centre lines of boxplots indicate the median; box outlines show 25th and
 696 75th percentiles, and whiskers indicate 1.5x the interquartile range. Extreme values are
 697 shown separately (dots). Asterisks indicate significant decoding accuracies or cross-
 698 generalization ($p < 0.05$).



699

700

Supplemental figure 3. Within impulse training and testing to estimate drift. **(A)** Response-

701

dependent averaging of trial-wise tuning curves (Fig. 6A). Shift towards response: Impulse 1:

702

$p = 0.4918$; Impulse 2: $p = 0.022$, one-sided. **(B)** Response-dependent training and testing

703

(Fig. 7A). Shift towards response: Impulse 1: $p = 0.545$; Impulse 2: $p = 0.009$, one-sided.

704

Same convention as Fig. 6B-C and Fig. 7B-C.