Cortical specialization for attended versus unattended working memory

Thomas B. Christophel[a,*], Polina Iamshchinina[a*], Chang Yan[a], Carsten Allefeld[a] & John-Dylan Haynes[a,b,c,d,e]

a Bernstein Center for Computational Neuroscience and Berlin Center for Advanced Neuroimaging and Clinic for Neurology, Charité Universitätsmedizin, corporate member of Freie Universität Berlin, Humboldt Universität zu Berlin, and Berlin Institute of Health, Berlin, Philippstraße 13, Haus 6, 10115, Germany

b Berlin School of Mind and Brain, Humboldt Universität, Berlin, Luisenstraße 56, Haus 1, Berlin, 10099, Germany

c Cluster of Excellence NeuroCure, Charité Universitätsmedizin, corporate member of Freie Universität Berlin, Humboldt Universität zu Berlin, and Berlin Institute of Health, Berlin, Charitéplatz 1, Hufelandweg 14, Berlin, 10117, Germany

d Department of Psychology, Humboldt Universität zu Berlin, Rudower Chaussee 18, Berlin, 12489, Germany

e SFB 940 Volition and Cognitive Control, Technische Universität Dresden, Zellescher Weg 17, 01069 Dresden, Germany

* These authors contributed equally to the current study; Please address correspondence to: tbchristophel@gmail.com, ++49 (0) 30 450 539352

Running Title: Decoding unattended working memory

Keywords: Working Memory, Visual Short-Term Memory, Attention, Activity-silent, fMRI, Multivariate Pattern Analysis

Acknowledgments: This work was funded by the Bernstein Computational Neuroscience Program of the German Federal Ministry of Education and Research BMBF Grant 01GQ0411, the Excellence Initiative of the German Federal Ministry of Education and Research DFG Grants GSC86/1-2009, KFO247, HA 5336/1-1 and JA 945/3-1/SL185/1-1.

Author Contributions: T.B.C., P.I. and J.D.H. designed the study. P.I. and C.Y. acquired data. T.B.C., P.I. and C.A. analyzed the data. T.B.C., P.I., C.A. and J.D.H. wrote the manuscript.

Abstract

Items held in working memory can be either attended or not, depending on their current behavioral relevance. It has been suggested that unattended contents might be solely retained in an activity-silent form. Here, instead, we demonstrate that encoding of unattended contents involves a division of labor. While visual cortex only maintains attended items, intraparietal areas and the frontal eye fields represent both attended and unattended items.

The short-term retention of sensory stimuli in working memory is fundamental to human cognition[1]. A wide range of primate electrophysiology and human imaging studies have reported content-selective brain signals that encode working memory contents across brief delays[2]. Such persistent stimulus-selective activity has been observed in multiple regions across the cortical sheet, including sensory, parietal, and frontal regions[2]. Recently, however, it has been postulated that working memory can be retained in an 'activity-silent state'[3–5]. In this notion, working memory contents are believed to be retained by changes in synaptic weights rather than neuronal firing[3,4]. In line with this, several studies have recently reported absence of persistent stimulus-selective activity when items are held in memory, but currently not behaviorally relevant[6–8]. Such currently non prioritized items are frequently referred to as 'unattended memory items' (UMIs) as opposed to 'attended memory items' (AMIs)[5]. These results suggest that attended memory items are retained actively while unattended memory items are retained in an activity-silent form.

However, the absence of content-selective signals for unattended items observed in prior work[6-8] might reflect a lack of sensitivity in the experimental procedures. For example, these studies used small numbers of subjects, they trained their classification models on attended items

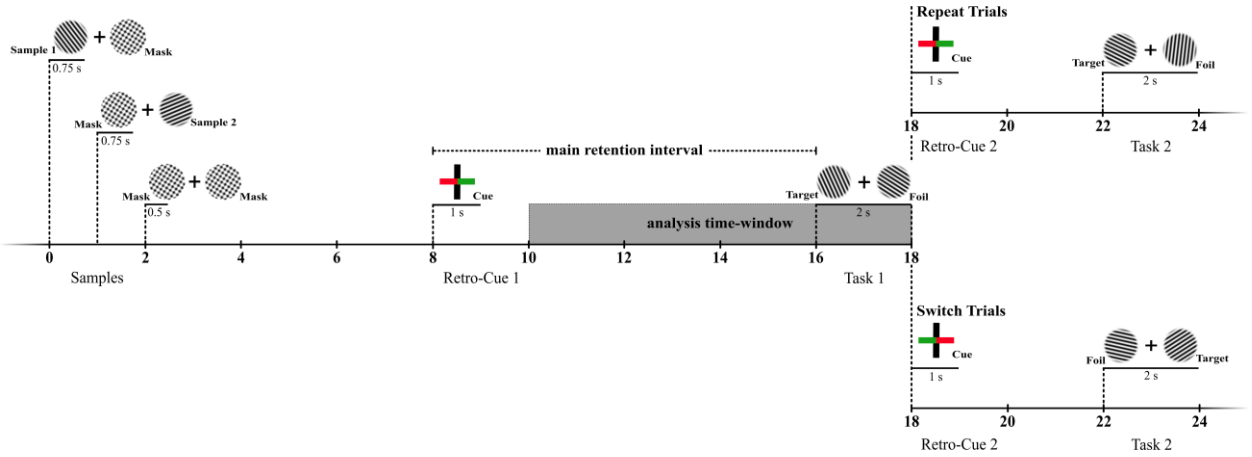


Figure 1. Two-stage orientation change discrimination task using retro-cue selection. In each trial, subjects are first presented with two sequential memory displays ('Sample 1', 'Sample 2'; see Fig. S1). The first presents one memory item (a Gabor patch of varying orientation) on one side and a plaid mask on the opposite side. The second display presents a second memory item on the other side, again accompanied by a plaid mask on the opposite side. The sequential presentation was chosen in order to avoid perceptual grouping of both memory items. The two memory displays were followed by a screen with plaid backward masks in the previous locations of the stimuli. 5.5 s later a first retro-cue (in red) indicated the side of the sample orientation that should be used for the first upcoming change discrimination task. This was followed by the main retention interval of 8 seconds. Following the main retention interval, participants viewed a Gabor presented at the cued (red) side and were required to judge whether it was rotated clockwise or counterclockwise to the cued sample. A random foil orientation was shown on the not-cued (green) side. This was followed by a second retro-cue that indicated either the same (i.e. the previously attended) memory sample ('Repeat Trials') or the other (previously unattended) memory item ('Switch Trials'). Following a short delay of 4 s participants were again probed with a test item and had to perform the same orientation judgement. Thus, to solve the task, subjects had to memorize both items during the main retention interval (following 'retro-cue 1'), but one item was prioritized (the attended memory item, AMI) over the other (the unattended memory item, UMI). MVPA analyses focused on the main retention period from 2 to 10 seconds after the retro-cue to account for hemodynamic delays.

in separate one-item tasks, and they only analyzed a limited set of voxels or electrodes, leaving it open whether unattended items might be represented in other brain areas or using an orthogonal neural code[9]. Here, we test directly whether brain regions in sensory but also parietal and frontal cortex contain memory representations during the delay phase for unattended stimuli. We acquired fMRI data from a large pool of subjects (N = 87) while they were memorizing orientation stimuli (see Fig. S1). We used a working memory design that allowed to separately identify representations of attended and unattended stimuli. In each trial, participants first memorized the

orientation of two gratings (see Fig. 1). After presentation of these stimuli a retro-cue indicated which of the two gratings would be tested in an upcoming change discrimination task following an extended delay, which is the main retention interval in our design. Then, after this memory test, a second retro-cue was shown that could select either the same or the other orientation for a second memory test. Such a two-stage retention task forces participants to maintain the orientations of both gratings until the second retro-cue, but prioritizes and thus directs attention to the first retro-cued item (AMI) while minimizing attention on the other item (UMI)[6-7].

We used a variant of multivariate pattern analysis (cvMANOVA, see Methods for details)[10] to identify which brain regions encoded the memorized orientations for attended and unattended items. The experiment was designed to optimize the ability to detect memory information in the main retention interval following the first retro-cue (see Fig. 1 & Online Methods for details). The analysis was conducted for stimuli in each hemifield separately in order to account for differences in retinotopic location. Our analysis focused on the set of regions where prior work indicated the presence of persistent stimulus-selective activity for orientations when attention was not manipulated[11–13]: Visual cortex (V1-V4), intraparietal sulcus (IPS0-5) and the frontal eye fields (see Fig. 2A).

In early visual cortex, we found reliable information about attended memory items (see Fig. 2B, one-tailed one-sample t-test; $t_{86} = 3.37$, $p = 0.000558$) whereas we found no significant information for unattended items (one-tailed one-sample t-test; $t_{86} = 0.19$, $p = 0.423091$, lower $CI^{95}_{(corrected)} = 0.01$). Information was also significantly higher for attended than for unattended items (two-tailed paired-sample t-test; $t_{86} = 2.65$, $p = 0.009467$, lower $CI^{95}_{(corrected)} = -0.012$, ΔD $CI^{95}_{(corrected)} = [0.007\ 0.048]$). This finding closely resembles previous reports that unattended memory items are not accompanied by delay-period information in perceptually driven brain

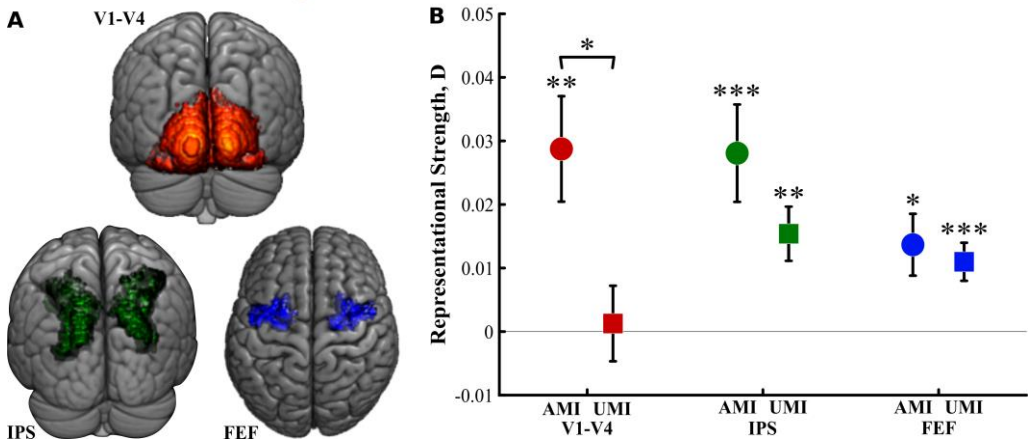


Figure 2. Representation of attended and unattended memory items. (A) Rendered representations of the human brain depicting the three main regions of interest (ROIs): Visual areas (V1-V4, in red), intraparietal areas (IPS0-5, in green) and the frontal eye fields (in blue). **(B)** Information about attended (AMI, circles) and unattended (UMI, squares) memory items as indicated by mean pattern distinctness D within each ROI. (n = 87 human subjects; error bars indicate SEM; tested using one-tailed one-sample t-tests & two-tailed paired-sample t-tests; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; all Bonferroni corrected for multiple comparisons)

regions[7]. Supplementary Figure 2 suggests that this attention effect is primarily driven by V1. It is worth noting that we cannot exclude that more sensitive methods might reveal information for unattended items also in visual cortex. Furthermore, whether either attended or unattended items can be decoded in the current study might depend on using a larger sample size than in prior work (see Fig. S3).

Regardless, if our analyses had focused exclusively on these visual brain regions we might have concluded that working memory representations for unattended stimuli are silent during the delay. Similarly, when we used no anatomical constraints on the voxels used by the classification algorithm but focused on voxels activated by perception of the samples (primarily found in sensory cortices, similar to Ref.[7]) we found information for attended items only (one-tailed one-sample t-test; AMI: $t_{86} = 4.51$, $p = 0.00001$, lower $CI^{95} = 0.024$; UMI: $t_{86} = 1.55$, $p = 0.061992$, lower $CI^{95} = -0.0008$; two-tailed paired-sample t-test; AMI versus UMI: $t_{86} = 2.49$, $p = 0.014547$, ΔD

$CI^{95}_{(corrected)}$ = [0.006 0.049]) replicating prior work[7]. However, when we focused our analysis on anterior regions (that were not selectively tested in prior work) the picture changed. In the intraparietal sulcus and the frontal eye fields, we found that both attended (one-tailed one-sample t-test; IPS: t_{86} = 3.66, p = 0.000216, lower $CI^{95}_{(corrected)}$ = 0.012; FEF: t_{86} = 2.53, p = 0.006667, lower $CI^{95}_{(corrected)}$ = 0.002) and unattended (one-tailed one-sample t-test; IPS: t_{86} = 3.24, p = 0.000848, lower $CI^{95}_{(corrected)}$ = 0.005; FEF: t_{86} = 3.81, p = 0.000129, lower $CI^{95}_{(corrected)}$ = 0.005) memory items were significantly represented by neural activity patterns (see Fig. 2B). We found no significant differences between information about attended and unattended items in these anterior regions (two-tailed paired-sample t-test; IPS: t_{86} = 1.52, p = 0.132059, ΔD $CI^{95}_{(corrected)}$ = [-0.004 0.033]; FEF: t_{86} = 0.11, p = 0.914863, ΔD $CI^{95}_{(corrected)}$ = [-0.011 0.014]). This pattern of results is reflected by significant differences in the modulation factor for attention (D_{AMI} / D_{UMI}; V1-V4: 23.96; IPS: 1.98; FEF: 1.05) in early visual areas as compared to IPS and FEF (bootstrap confidence intervals; V1-V4 – IPS: $CI^{95}_{(corrected)}$ = [3.1, $6.1 \cdot 10^5$], V1-V4 – FEF: $CI^{95}_{(corrected)}$ = [4.0, $4.7 \cdot 10^5$], IPS – FEF [-1.4, 6.2]).

In early visual cortex, the time-course of information closely resembles prior work[6,7] showing null-results for unattended items 2 s after cue onset, whereas more anterior areas appear to represent unattended items as late as 6-10 seconds after the cue (see Fig. S4). To explore this further, we asked whether similar brain activity patterns represented the remembered items in data recorded in the 2 seconds before the cue and in the time-period 6-10 seconds after the cue. We found such pattern stability xD (see methods for details) across time only in the intraparietal sulcus and the frontal eye fields (one-sided one-sample t-test; EVC: xD_{AMI} = 0.0020, t_{86} = 1.24, p = 0.1088; xD_{UMI} = 0.0017, t_{86} = 0.90, p = 0.1845; IPS: xD_{AMI} = 0.0063, t_{86} = 3.21, p = 0.0009; xD_{UMI}

= 0.0027, t_{86} = 1.51, p = 0.0678; FEF: xD_{AMI} = 0.0031, t_{86} = 2.70, p = 0.0042; xD_{UMI} = 0.0050, t_{86} = 3.64, p = 0.0002).

Finally, we tested whether voxels in the hemisphere contralateral to sample presentation carry more information about these memorized contents than ipsilateral voxels. In visual cortex, consistent with prior work[14], we found no evidence for lateralization (one-tailed paired-sample t-test; AMI: t_{86} = -1.13, p = 0.8696; UMI: t_{86} = -1.60, p = 0.9437). The frontal eye fields showed similar results (AMI: t_{86} = -0.11, p = 0.5447; UMI: t_{86} = 1.11, p = 0.1341). In intraparietal areas, however, we found more information regarding attended items in contra- versus ipsilateral voxels (AMI: t_{86} = 2.84, p = 0.0027; UMI: t_{86} = -2.1, p = 0.9823) and contralateral areas carried more information about attended then unattended items (one-tailed paired-sample t-test; t_{86} = 4.55, p = 0.000009).

Our results directly contradict the assertion that unattended working memory items are encoded solely in an activity-silent fashion. We cannot discern whether stimulus-selective persistent activity represents an active recurrent excitation network[15] or selective activity related to other potential forms of retention. Current computational models of retention via synaptic plasticity, for example, either (a) require neuronal firing as a means to uphold synaptic signals over longer periods of time[4] or (b) suggest a complimentary role of recursive activity and synaptic plasticity[16]. Selective changes in synaptic plasticity could even lead to purely epiphenomenal selective firing. Thus, the presence or absence of stimulus-selective persistent activity[2] (in spiking, LFP or BOLD activity) does not rule out synaptic contributions to working memory.

Critically, our results do provide support for a different hypothesis how attended and unattended items differ in their neural representation. One possibility is that sensory cortex

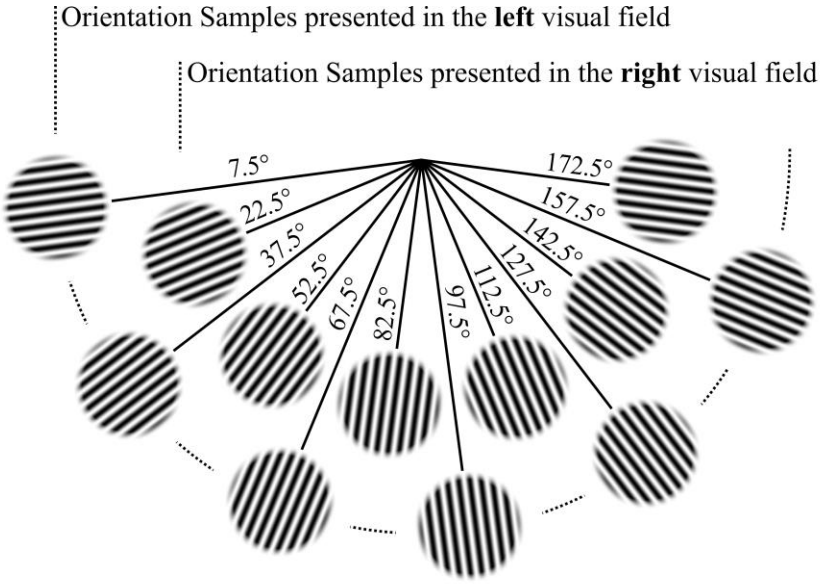
maintains a high-resolution representation of the currently attended memory item, whereas parietal cortex has a low-resolution representation of both attended and unattended items[2,9]. If that were the case one would expect working memory performance to be more accurate for previously attended items. In line with this, behavioral evidence shows that retention in an unattended state can result in impaired change detection[17], more guesses and non-target responses[18], stronger categorical biases[19] and less precision[18,19] as compared to attended memory items. The behavioral data from the current study are consistent with these findings (see Fig. S5). Our finding of selective recruitment of early visual cortex for the retention of attended memories could be the neural source of these behavioral benefits. In line with this, imaging evidence shows that the precision of neural representations in visual cortex during the delay period correlates with the behavioral precision during recall[20]. In the current study, the amount of information we found in early visual cortex for attended memory items correlated with the individual subjects' discrimination threshold (Pearson's linear correlation coefficient; $r = -0.3177$, $p = 0.0027$).

We thus propose that the attentional modulation of working memory is not implemented by switching from an activity-based to an activity-silent, synaptic code, or by increasing the level of selective activity for one representation globally across all areas. Rather the attentional prioritization in working memory might be realized by the selective recruitment of sensory representations which more precisely retain the information for an upcoming task.

References

1. Baddeley, A.D. (Clarendon Press: Oxford (UK), 1986).
2. Christophel, T.B., Klink, P.C., Spitzer, B., Roelfsema, P.R. & Haynes, J.-D. Trends Cogn. Sci. 21, 111–124 (2017).
3. Stokes, M.G. Trends Cogn. Sci. 19, 394–405 (2015).
4. Mongillo, G., Barak, O. & Tsodyks, M. Science 319, 1543–1546 (2008).
5. Larocque, J.J., Lewis-Peacock, J.A. & Postle, B.R. Front. Hum. Neurosci. 8, 5 (2014).

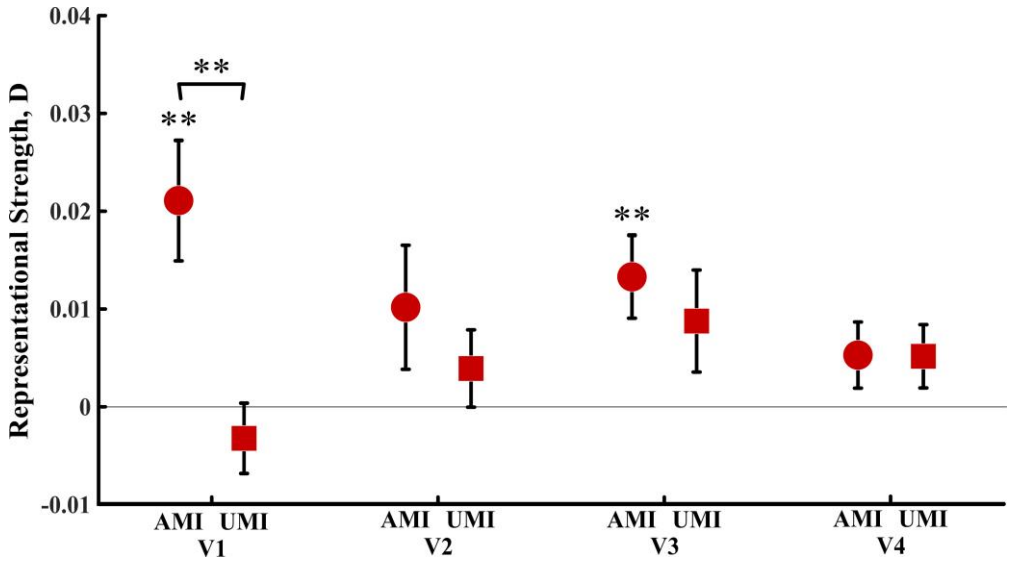
6. Lewis-Peacock, J.A., Drysdale, A.T., Oberauer, K. & Postle, B.R. J. Cogn. Neurosci. 24, 61–79 (2012).
7. LaRocque, J.J., Riggall, A.C., Emrich, S.M. & Postle, B.R. Cereb Cortex 27, 4881–4890 (2017).
8. Wolff, M.J., Jochim, J., Akyürek, E.G. & Stokes, M.G. Nat. Neurosci. 20, 864–871 (2017).
9. Olivers, C.N.L., Peters, J., Houtkamp, R. & Roelfsema, P.R. Trends Cogn. Sci. 15, 327–334 (2011).
10. Allefeld, C. & Haynes, J.-D. Neuroimage 89, 345–357 (2014).
11. Harrison, S.A. & Tong, F. Nature 458, 632–635 (2009).
12. Serences, J.T., Ester, E.F., Vogel, E.K. & Awh, E. Psychol. Sci. 20, 207–214 (2009).
13. Ester, E.F., Sprague, T.C. & Serences, J.T. Neuron 87, 893–905 (2015).
14. Ester, E.F., Serences, J.T. & Awh, E. J. Neurosci. 29, 15258–15265 (2009).
15. Hebb, D.O. (Wiley: Oxford, England, 1949).
16. Itskov, V., Hansel, D. & Tsodyks, M. Front. Comput. Neurosci. 5, 1–19 (2011).
17. Rerko, L. & Oberauer, K. J. Exp. Psychol. Learn. Mem. Cogn. 39, 1075–1096 (2013).
18. Emrich, S.M., Lockhart, H.A. & Al-Aidroos, N. J. Exp. Psychol. Hum. Percept. Perform. 43, 1454–1465 (2017).
19. Bae, G.-Y. & Luck, S. J. Vis. 16, 701–701 (2016).
20. Ester, E.F., Anderson, D.E., Serences, J.T. & Awh, E. J. Cogn. Neurosci. 25, 754–761 (2013).



Supplementary Figure 1

Stimulus set.

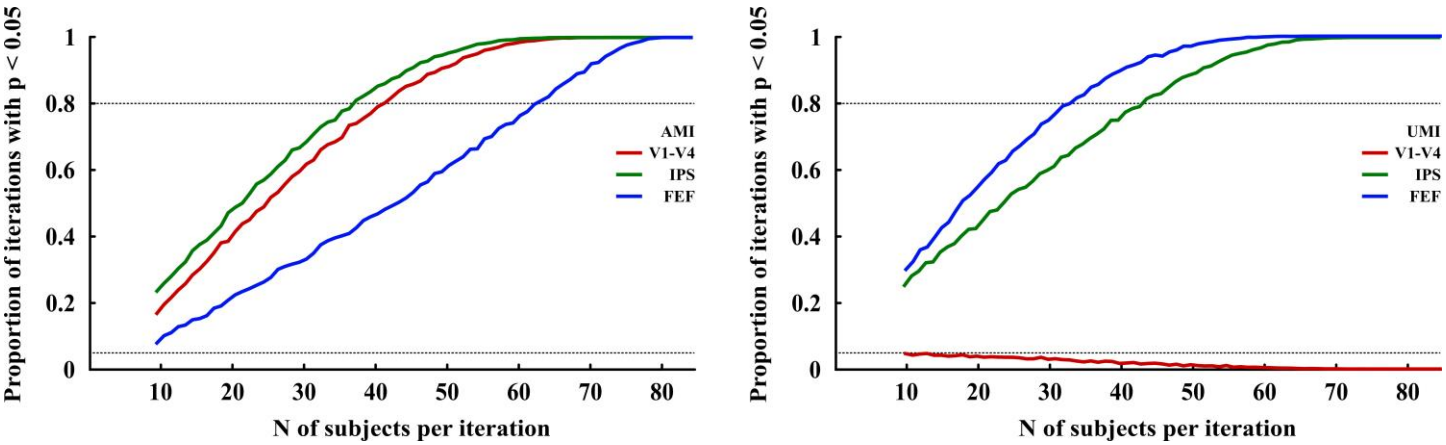
The main experiment used a set of twelve orientations, six on the left and six on the right. On each side the stimulus set consisted of orientations that were 30° apart, but the two sets were separated by 15° to experimentally decorrelate the two conditions.



Supplementary Figure 2

Representation of attended and unattended memory items in different regions of the visual cortex.

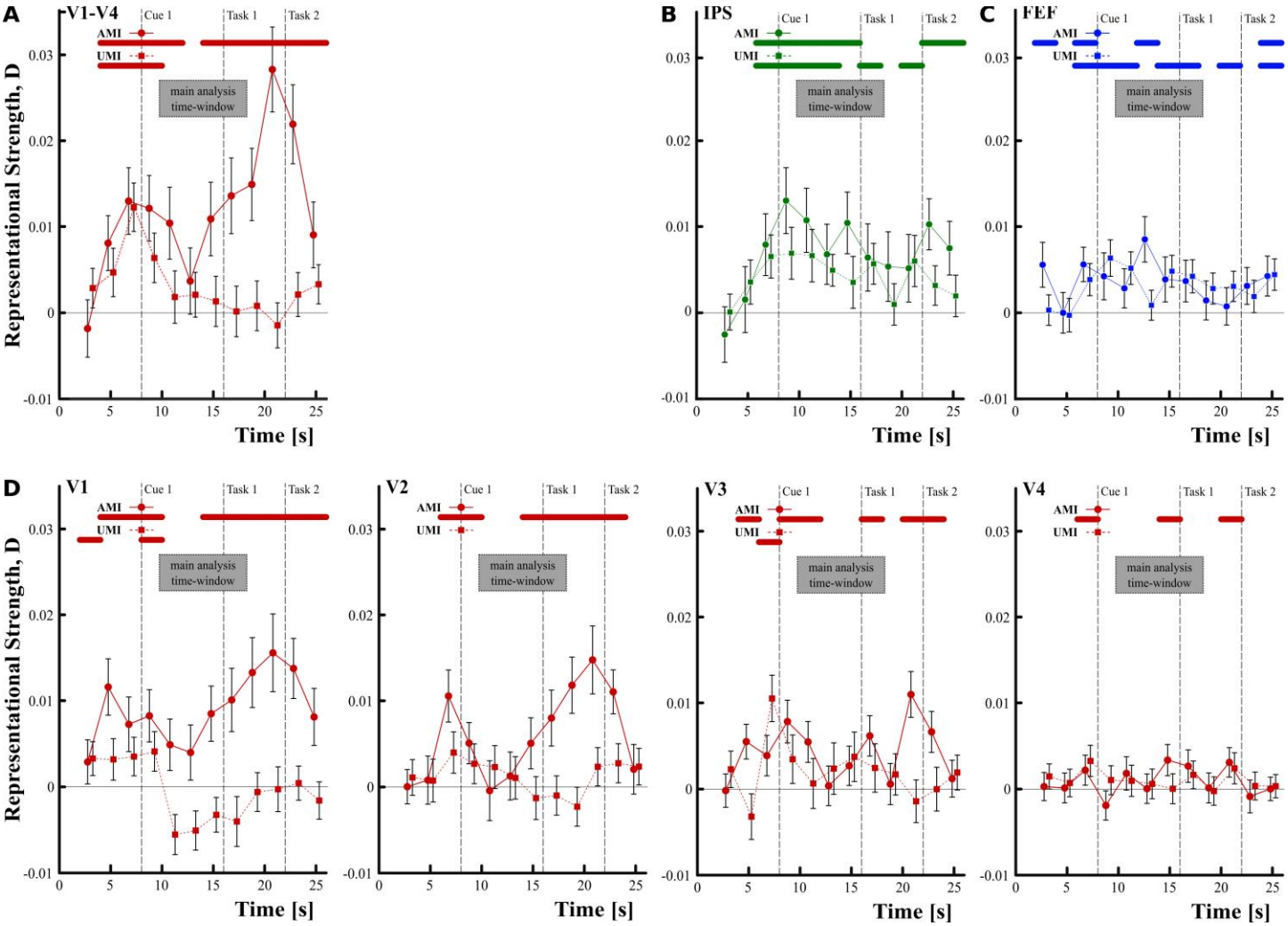
Information about attended (AMI, circles) and unattended (UMI, squares) memory items as indicated by mean pattern distinctness D within different regions of the visual cortex. (n = 87 human subjects; error bars indicate SEM; *: $p < 0.05$; **: $p < 0.01$; ***: $p < 0.001$; Bonferroni corrected for multiple comparisons). Results were tested against chance using one-tailed one-sample t-tests (V1AMI: $t86 = 3.34$, $p = 0.0006$; V1UMI: $t86 = -0.87$, $p = 0.81$; V2AMI: $t86 = 1.58$, $p = 0.059$; V2UMI: $t86 = 0.86$, $p = 0.19$; V3AMI: $t86 = 2.97$, $p = 0.002$; V3UMI: $t86 = 1.6$, $p = 0.057$; V4AMI: $t86 = 1.45$, $p = 0.076$; V4UMI: $t86 = 1.48$, $p = 0.071$) and for differences between AMIs and UMIs using two-tailed paired-sample t-tests (V1: $t86 = 3.42$, $p = 0.001$; V2: $t86 = 0.79$, $p = 0.43$; V3: $t86 = 0.61$, $p = 0.54$; V4: $t86 = 0.02$, $p = 0.98$).



Supplementary Figure 3

Estimation of the probability of significance of the multivariate pattern analyses with a reduced number of subjects.

Probabilities were estimated using a bootstrapping procedure with 10000 iterations for each of the main multivariate pattern analyses, each area and each possible N ranging from 10 to 86. In each iteration, N values of D were drawn from the whole sample of 87 human subjects and a one-tailed one-sample t-test was testing against chance-level (D = 0) at an uncorrected threshold of $p < 0.05$. Please note that these estimates do not easily generalize to studies using a different stimulus set, different numbers of repetitions or an otherwise different experimental paradigm. Please further note that the qualitative difference in reliability of information in FEF for unattended over attended items is a result of a difference in variance.



Supplementary Figure 4

Representation of attended and unattended memory items across time.

Information about attended (AMI, circles) and unattended (UMI, squares) memory items as indicated by mean pattern distinctness D on a time-point by time-point basis. Vertical lines indicate the onset of the first cue and the two tasks. Grey panels demark the time-points used for the main analysis shown in Figure 2. (n = 87 human subjects; error bars indicate SEM; Colored lines on the top part of the figure indicate time points with significant above-chance information at $p < 0.05$ for AMI [top] and UMI [bottom]; tested using one-tailed one-sample t-tests).

Online methods

Participants

89 healthy right-handed human subjects (42 female; mean age: 26.8, SEM ±0.4) with normal or corrected-to-normal vision were recruited for the current study. We based our estimate of the necessary sample size on our prior work on attended working memories[21-24] as effect sizes for unattended memories are unknown. We substantially increased the N relative to these studies, to account for putative reductions in effect size for unattended items and because of a generally lowered trial count. Data acquisition for two subjects was aborted during the experiment upon the request of the participants. This was prior to any stage of data analysis. We thus analyzed data from 87 participants (41 female; mean age: 26.8, SEM ±0.4). Subjects gave informed consent and the study was approved by the local ethics committee (Ethics committee, Department of Psychology, Humboldt University, Berlin). Data collection and analysis were not performed blind to the conditions of the experiment. Detailed information on the experimental design can also be found in the 'Life Sciences Reporting Summary' published alongside this article.

Procedure and design

In an MRI scanner, subjects performed a two-stage delayed change discrimination task[6,7] using Gabor grating stimuli. During this task, subjects memorized two gratings and were twice instructed using retro-cues to attend to one or the other for an upcoming change detection task. Critically, this paradigm results in a situation after the first cue where one orientation sample is prioritized (the attended memory item, AMI) while the other sample is remembered but with lower priority (the unattended memory item, UMI).

In each trial, participants viewed two sine-wave sample grating stimuli shown consecutively left and right to the center of the screen in random order (0.8° off center, for 750 ms, ISI 250 ms). During grating presentation the side opposite to the grating was occupied with Gabor plaid stimuli and the same plaids were shown as masks after stimulus presentation (for 500 ms). The onset of these masks was followed after 5500 ms with by retro cue 1 (presented for 1000 ms). For this cue, one of the handles of the fixation cross turned red while the opposite handle turned green. Participants were instructed that the sample grating which had been shown on the red side of the fixation cross was to be used for the upcoming change detection task.

This first retro-cue was followed by a prolonged delay (7000 ms), the main retention interval, during which participants were expected to maintain orientation representations of both sample gratings with the cued orientation prioritized over the other. This delay is the critical time-period of interest during which neural representations of prioritized contents (attended memory items, AMI) and not-prioritized information ('unattended' memory items, UMI) can be distinguished. During the change discrimination task that followed ('task 1', 2000 ms), participants had to report whether a target grating presented on the cued side was rotated clockwise or counterclockwise relative to the cued sample grating previously presented on the same side. On the opposite side of the screen a randomly rotated foil grating was shown.

This first task was followed by a second cue ('retro-cue 2', 1000 ms), an additional delay (3000 ms) and a second task ('task 2', 2000 ms). The grating cued for the second task could either be the same as for the first task (repeat trials) or the other grating (switch trials). The switch probability was 50%. This second task ensured that an item not prioritized for the first task needed to be retained until the second cue was presented because subjects did not know in advance whether the item might be relevant later. Please note that our study was designed to maximize sensitivity

in the first delay period. For this reason, the second delay was chosen very short. Thus, we did not perform an analysis of a potential re-instatement of information in the 50% switch trials when a stimulus feature switches from unattended to attended[6,7]. This was further motivated by recent proposals that evidence of re-instatement is not conclusive evidence of silent working memory[25]. The intertrial interval was either 2000 ms (50% of trials), 4000 ms (33.3%) or 6000 ms (16.7%). A fixation cross (width 0.2°) remained on screen throughout the experiment.

Prior to the main experiment in the MRI scanner (2-4 days in advance), participants took part in a training session outside the scanner using a conventional LCD display. The training comprised four experimental runs with shorter delays (1500 ms and 2500 ms for the first and the second delay periods, respectively) and using fully randomized sample orientations. Subjects were instructed not to use verbal labels for memorizing the stimuli. In the MRI scanner, the experimental paradigm was presented on a NordicNeuroLab Monitor (70,5 cm width) and subjects viewed the screen via a mirror. Stimulus presentation was controlled using Psychtoolbox[26].

The stimuli used were sine-wave Gabor gratings (5.7° size, phase randomized, spatial frequency: 1.8 cycles/degree, 3.06° away from the screen center) with varying orientations. Importantly, sample orientations presented (see Fig. S1) in the left visual field were drawn from a different pool (7.5°, 37.5°, 67.5°, 97.5°, 127.5°, 157.5°) than orientations on the right (22.5°, 52.5°, 82.5°, 112.5°, 142.5°, 172.5°) which allowed us to decorrelate items shown in the left and right hemifield and thereby AMIs and UMIs (see ref. [27]). Gabor plaid stimuli (5.7° size) consisting of two random but orthogonal orientations (phase randomized, spatial frequency: 1.8 cycles/degree) were used as masks.

For the change discrimination tasks, targets were rotated clockwise and counter-clockwise on an equal number of trials. The extent of rotation of the test grating was initially set to 20° and adjusted using a staircase procedure to generate a consistently challenging task and avoid ceiling effects. For each correct response in a given trial (0-2), the difference between test and sample orientation was reduced by 0.5° making change discrimination harder. Reversely, the difference was increased by 2° for each incorrect response, thus making them easier to differentiate. Changes to this discrimination threshold were only applied after the end of a given trial and the same levels were used for 'task 1' and 'task 2' to allow for comparisons between the tasks. The adjustment of started during training and continued throughout the fMRI experiment.

There were 4 scanning runs of 48 trials each. We used a within-subject 2 (retro-cue 1: left vs. right) by 2 (retro-cue 2: switch vs. repeat) design. Each of the 12 orientations (6 for each side) had to be memorized in 8 trials per run. The pairing of orientations on the left and right was fully randomized to allow statistically independent analyses of attended and unattended orientations. The assignment of attention conditions to orientation conditions was fully randomized in the first 39 participants. Based on theoretical considerations[28] we then decided to slightly modify the randomization so that in the second set of 48 participants each attention condition is associated with exactly the same number of trials for each orientation. We found no significant differences between the groups in our main analyses. For the change discrimination tasks, the rotation of the test orientation was randomized with respect to all other conditions with equal frequency of clockwise and counterclockwise rotations for each of the two tasks. The temporal order of conditions was fully randomized within each run.

Overall, the experiment occupied 90 min per participant. In order to decrease effects of long-term memory, a short 5-minute run was presented between the 2nd and 3rd experimental

block where participants performed the task with random sample orientations while anatomical scans were acquired. After the experiment, participants filled out a questionnaire covering the strategies participants used for memorization.

Data acquisition

fMRI data were collected with Siemens 3-Tesla TIM-Trio MR tomograph located at the Berlin Center for Advanced Imaging (Charité-Universitätsmedizin). Within each of the four runs, we recorded 663 T2*-weighted gradient-echo echo-planar images (EPI, 33 slices, 3 × 3 × 3 mm resolution, 0.6 mm gap, descending order, FoV = 192 mm, TR = 2000 ms, TE = 30 ms, flip angle = 80°). The onset of each trial was locked to the onset of the acquisition of an image to minimize variation due to slice acquisition onsets. Slices were aligned parallel to the anterior and posterior commissures and covered the whole neocortex. In addition, a high-resolution T1-weighted image was acquired (192 sagittal slices, 1 mm thickness, RT = 1900 ms, TE = 2.52 ms, flip angle = 9°, FOV = 256 mm).

fMRI preprocessing

Functional imaging data was analyzed using SPM12[29] and cvMANOVA[10]. After conversion to NIfTI format, the functional data were motion corrected and the anatomical image was coregistered to the first image of the BOLD time series. No normalization into a standard space was performed and we applied no Gaussian smoothing to the data prior to performing the multivariate analyses to preserve the fine-scaled spatial structure of the fMRI data. Using a similar reasoning, we abstained from using slice-time correction during preprocessing and avoided any other temporal filtering to retain the temporal precision of the stimulus-locked time-series (see 'Data Acquisition').

Anatomical regions-of-interest (ROI)

We aimed to identify information about attended and unattended memory items within brain regions previously found to carry information about memorized gratings[11–13]. For this, anatomical probability maps of retinotopic areas[30] in visual cortex (V1- V4), the intraparietal sulcus (IPS0-5) and the frontal eye fields (Fig. 2A) were backward transformed into participants' native space using unified segmentation[31]. These maps were thresholded to exclude voxels with a probability to be part of a given area that is lower than 0.1. In a post-hoc exploratory analysis, we investigated the information content of V1 to V4, separately. Please note that an analysis based on individual retinotopic maps might have had higher sensitivity to detect weak effects in areas beyond V1. For the main analyses, these anatomical masks were collapsed across the left and the right hemisphere. Separate masks for the left and right hemispheric portions of these areas were used to investigate the lateralization of representations in a post-hoc exploratory analysis. Finally, we also performed a post-hoc exploratory analysis without any anatomical preselection of voxels to compare our results to previous studies[7].

Univariate analyses of sample related activity

To estimate BOLD activity during the trial a GLM with seven regressors was designed using hemodynamic response functions (HRF) time-locked to the onsets of the following events and adjusted by their duration: sample grating onsets (1st regressor), first and second cue onsets (2nd-3rd), each of the delay period onsets separately modeled taking into account their respective durations (4th: prior to the first cue; 5th: following the first cue; 6th: following the second cue) and both discrimination task onsets merged into one regressor (7th). To identify voxels that responded to our grating stimuli, we generated *t*-maps contrasting the sample grating onsets (irrespective of orientation) against the implicit baseline of the model. Then, individual subject-level t-maps were

overlaid with each of the ROIs created for each participant. To avoid an arbitrary selection of n voxels for each ROI, we generated 25 versions of every ROI, each representing the n voxels with the highest sample related activity. For this, we varied n between 20 and 500 voxels in steps of 20 voxels. We only considered voxels that were positively activated ($t > 0$). These ROIs with varying sizes were later used in a nested cross-validation to choose an optimal voxel number for each ROI which at the same time avoiding overfitting.

Analyses of multivariate pattern distinctness using cvMANOVA

The goal of the multivariate pattern analyses was to test whether retinotopic areas in visual, parietal and frontal cortex have representations of the remembered attended and unattended orientations. For this, we used a recently developed technique for multi-voxel pattern analysis, cross-validated MANOVA[10]. cvMANOVA constitutes a variant of multivariate analyses of variance[32] that can be used to quantify differences in BOLD response patterns[24,33]. The method is comparable to more common classifier-based 'decoding' analyses[11-14,20–24,34-43] but has a number of advantages: It avoids binary classification in favor of a continuous measure of patterned differences, performs a parameter-free analysis based on a probabilistic model of the data (the multivariate general linear model) and results in an interpretable multivariate effect size (explained variance). Moreover, since D is a cross-validated version of a likelihood-ratio statistic, it can be expected to be more sensitive than classification accuracy (cf. Fig. 3d in Ref.[10]). Please note that prior work employed data from one-item tasks (i.e. including only attended items) for classifier training to avoid training on ambiguous two-item data where the representations of two items might overlap[6,7]. This analysis potentially biases the results in favor of attended memory items. Prior work failed to identify information for either attended or unattended items when only using data from two-item tasks[7], possibly due to a lack of power. Here, to avoid biasing our results, we

only recorded data in two-item tasks and significantly increased the N to counteract the lowered power when training on such data.

Here, we used cvMANOVA to ask whether activity patterns in our ROIs carried information about the memorized contents. As a first step, a multivariate general linear model (MGLM) using finite impulse response (FIR) functions was used to estimate memory-related activity in each voxel. Two first level models were estimated for samples presented in the left and the right visual field respectively. For each of the 6 orientations per side, we used 12 regressors to model the entire 24 seconds of each trial in 2-second time bins (i.e. the length of the TR) spanning the time between onset of the masks (after sample presentation) until two seconds after the end of the trial (only a subset of this time was in the actual decoding analysis time window, see below and Fig. 1). This set of regressors was modeled for each orientation in two different conditions: when a particular orientation was selected by the cue and when it was not (24 regressors per sample grating). BOLD activity in each voxel was fitted with this set of FIR regressors separately for each of the four runs. Thus, we had 144 regressors per run for the left and the right side separately (factors: 'time point' [12] × 'attended/unattended' [2] × 'orientation' [6]) collapsing across repeat and switch trials and a constant regressor to model the run mean. Parameter estimates and residuals from these two models were then used in the region-based cvMANOVA.

Within each model, to test the effect of orientation identity, we used two contrast matrices which (with respect to a single time point) had the forms

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \text{for attended items and } \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & -1 \end{pmatrix},$$

for unattended items, where the 12 rows correspond to the 12 regressors for attended and unattended orientations.

The columns correspond to the five partial contrasts (comparing orientations 1 vs. 2, 2 vs. 3, 3 vs. 4, 4 vs. 5, and 5 vs. 6) together defining the effect of orientation. We analyzed the four time-points following the presentation of the first cue (i.e. FIR bins 5-8, when attended and unattended items are differentiated) with a shift of two seconds to account for hemodynamic delay. For this, the two contrast matrices were replicated and zero padded. This can be exemplified for attended memory items ('A' = 1) as follows:

$$
\begin{array}{ccc}
T & O & A \\
1 & 1 & 1 \\
2 & 1 & 1 \\
3 & 1 & 1 \\
4 & 1 & 1 \\
5 & 1 & 1 \\
6 & 1 & 1 \\
7 & 1 & 1 \\
8 & 1 & 1 \\
9 & 1 & 1 \\
10 & 1 & 1 \\
11 & 1 & 1 \\
12 & 1 & 1 \\
1 & 2 & 1 \\
2 & 2 & 1 \\
3 & 2 & 1 \\
4 & 2 & 1 \\
5 & 2 & 1 \\
6 & 2 & 1 \\
\vdots & \vdots & \vdots
\end{array}
\begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
1 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 1 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 1 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 1 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
-1 & 0 & 0 & 0 & 1 & 0 & \cdots \\
0 & -1 & 0 & 0 & 0 & 1 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix},
$$

Here, the rows correspond to the 144 regressors per run, with FIR regressors for all 12 time-points ('T') of each orientation condition ('O') grouped together. For the two contrasts per model we estimated the amount of multivariate variance explained by the encoded effect relative to the multivariate error variance using cross-validated MANOVA[10].

cvMANOVA estimates the variance of the multivariate fMRI time series that can be explained by a contrast between conditions (here: differences in multivariate responses to different orientations). The 'pattern distinctness'

$$\mathrm{D} = \mathrm{trace}\left(\frac{1}{\mathrm{n}}\mathrm{B}'\mathrm{C}\mathrm{C}^{-}\mathrm{X}'\mathrm{X}\mathrm{C}\mathrm{C}^{-}\mathrm{B} \cdot \Sigma^{-1}\right),$$

where C is the contrast matrix, B is the parameter matrix, n is the number of scans, X is the design matrix, and Σ is the error covariance matrix. D is the amount of multivariate variance explained by the effect encoded in the contrast in units of the multivariate error variance. In order to obtain

an unbiased estimate of the explained variance cross-validation is used (see Ref.[10] for details). Because of the in some cases large number of voxels in a region, explained variance was computed relative to an estimate of the multivariate error variance Σ which was regularized towards the diagonal, using an optimized regularization parameter[44].

In the present application, if different orientations would elicit the same multivariate response, D would on average be 0, while different responses to different orientations would lead to an average D larger than 0. D for attended and unattended orientations was averaged across orientations shown in the left and right visual field.

This procedure was performed separately for all ROIs using varying voxel counts (between 20 and 500 voxels, see above). In order to select the optimal number of voxels for each area within each subject and analysis while avoiding double-dipping, we used a nested cross-validation approach. For every subject, we averaged D across all other subjects for all 25 possible ROI sizes and selected the ROI size with maximal D. D for the left-out subject with this ROI size was kept and this procedure was repeated for every subject.

While the main analysis used one contrast for attended and unattended contents, each across time points, we also performed a post-hoc exploratory analysis that used separate contrasts for the twelve time points we estimated in the (FIR-based) multivariate general linear model to generate time courses (Fig. S4). Above-zero D for each time point indicates that the multivariate responses to different orientations are different at that time point, but not necessarily that this multivariate response difference stays the same over time. We thus finally wanted to assess whether the informative brain patterns are stable across time (akin to generalization in a cross-classification analysis), in particular between the time point directly prior to the onset of the cue and the two

time-points at 6–10 seconds after the cue onset. Where standard cvMANOVA quantifies the pattern information in the form of variance explained by a contrast (here between different orientations), for this purpose we used a variant that estimates the amount of similarly encoded pattern information in the form of explained variance shared between two contrasts (orientation-specific responses at two different time points), the 'pattern stability'

$$\mathrm{xD} = \mathrm{trace}\left(\frac{1}{\mathrm{n}}\mathrm{B}'\mathrm{C}_1\mathrm{C}_1^{-}\mathrm{X}'\mathrm{X}\mathrm{C}_1\mathrm{C}_2^{-}\mathrm{B}\cdot\Sigma^{-1}\right),$$

where C_1 and C_2 are the two different contrast matrices, B is the parameter matrix, n is the number of scans, X is the design matrix, and Σ is the error covariance matrix. Again, cross-validation is used to obtain an unbiased estimate. xD is on average 0 if there is no shared variance between the two contrasts, i.e. if the respective informative patterns are orthogonal to each other.

Statistical testing

Group-level statistics (n = 87 human subjects) were performed using one-sample and paired t-tests. Please note that one-sample t-tests do not provide population inference[45]. The data was tested for deviation from normality using Kolmogorov–Smirnov tests. We applied Bonferroni correction to all resulting p-values to account for the number of areas tested in a given analysis.

To test whether the reduction of representational strength in UMI vs AMI is different between regions, one would normally test for an interaction using an ANOVA with the factors attention and region. Standard interaction analyses test the difference (e.g. between areas) of a difference (e.g. between the conditions, D_{AMI} - D_{UMI}) for significance. However, information measures are generally not comparable between brain areas as a result of their unique neural topology, vascular structure, and levels of physiological noise[46-48]. In a post-hoc exploratory analysis, we therefore adopted the strategy to first quantify the effect of attention within region by

an attentional modulation factor D_{AMI} / D_{UMI}, i.e. the ratio between the average D under AMI and UMI, which should be comparable between regions. We then computed the three pairwise differences of attenuation factors between the three regions, and assessed significant difference from 0 by means of bootstrap confidence intervals on these differences[49] based on 100,000 resamples of the 87 subjects, at 95 % confidence corrected for multiple comparisons.

For behavioral analyses, proportions of correct responses were calculated treating missed responses as errors. Reaction time analyses were conducted excluding missed responses. Differences in proportions of correct responses were tested using two-tailed Wilcoxon signed rank test (n = 87). Discrimination thresholds were averaged across all trials. Correlations between discrimination thresholds and measures of pattern distinctness for attended and unattended memory items were tested using Pearson's linear correlation coefficients.

Data availability

The MRI and behavioral data that were used in this study are available to researchers from the corresponding author upon request.

Code availability

MATLAB source code for cvMANOVA is available online (https://github.com/allefeld/cvmanova/releases). For the analyses in this paper we used v2 (2015–1–12).

Additional References

21. Christophel, T.B., Hebart, M.N. & Haynes, J.-D. J. Neurosci. 32, 12983–12989 (2012).
22. Christophel, T.B. & Haynes, J.-D. Neuroimage 91, 43–51 (2014).
23. Christophel, T.B., Cichy, R.M., Hebart, M.N. & Haynes, J.-D. Neuroimage 106, 198–206 (2015).

24. Christophel, T.B., Allefeld, C., Endisch, C. & Haynes, J.-D. Cereb Cortex 1–16 (2017).doi:10.1093/cercor/bhx119
25. Schneegans, S. & Bays, P.M. Journal of Cognitive Neuroscience 1–18 (2017).doi:10.1162/jocn_a_01180
26. Brainard, D.H. Spat. Vis. 10, 433–436 (1997).
27. Pratte, M.S. & Tong, F. J. Vis. 14, 22, 1–12 (2014).
28. Görgen, K., Hebart, M.N., Allefeld, C. & Haynes, J.-D. arXiv:1703.06670 [q-bio, stat] (2017).at <http://arxiv.org/abs/1703.06670>
29. Friston, K.J. et al. Hum. Brain Map. 2, 189–210 (1994).
30. Wang, L., Mruczek, R.E.B., Arcaro, M.J. & Kastner, S. Cereb Cortex 25, 3911–3931 (2015).
31. Ashburner, J. & Friston, K.J. Neuroimage 26, 839–851 (2005).
32. Timm, N.H. (Springer: New York (NY), 2002).
33. Guggenmos, M., Wilbertz, G., Hebart, M.N. & Sterzer, P. eLife 5, e13388 (2016).
34. Haynes, J.D. & Rees, G. Nat. Neurosci. 8, 686–691 (2005).
35. Haynes, J.-D. & Rees, G. Curr. Biol. 15, 1301–1307 (2005).
36. Haynes, J.-D. & Rees, G. Nat. Rev. Neurosci. 7, 523–534 (2006).
37. Kamitani, Y. & Tong, F. Nat. Neurosci. 8, 679–685 (2005).
38. Soon, C.S., Brass, M., Heinze, H.J. & Haynes, J.D. Nat. Neurosci. 11, 543–545 (2008).
39. Sterzer, P., Haynes, J.-D. & Rees, G. J. Vis. 8, 10.1-12 (2008).
40. Cichy, R.M., Chen, Y. & Haynes, J.D. Neuroimage 54, 2297–307 (2011).
41. Riggall, A.C. & Postle, B.R. J. Neurosci. 32, 12990–12998 (2012).
42. Lee, S.-H., Kravitz, D.J. & Baker, C.I. Nat. Neurosci. 16, 997–999 (2013).
43. Bettencourt, K.C. & Xu, Y. Nat. Neurosci. 19, 150–157 (2016).
44. Schäfer, J. & Strimmer, K. Stat. Appl. Genet. Mo. B. 4, 1–30 (2005).
45. Allefeld, C., Görgen, K. & Haynes, J.-D. Neuroimage 141, 378–392 (2016).
46. Dubois, J., Berker, A.O. de & Tsao, D.Y. J. Neurosci. 35, 2791–2802 (2015).
47. Haynes, J.-D. Neuron 87, 257–270 (2015).
48. Hebart, M.N. & Baker, C.I. NeuroImage (2017).doi:10.1016/j.neuroimage.2017.08.005
49. Efron, B. & Tibshirani, R.J. (CRC Press: 1994).

Competing financial interests

The authors have no competing interests, or other interests that might be perceived to influence the results or discussion reported in this paper.