Methods

MRI acquisition. Scanning was performed on a 3.0-Tesla Philips Intera Achieva scanner using a standard bird-cage 8-channel head coil at the Vanderbilt University Institute of Imaging Science. A high-resolution 3D anatomical T1-weighted scan was acquired from each participant (FOV 256 x 256, 1 x 1 x 1 mm resolution). To measure BOLD contrast, standard gradient-echo echoplanar T2*-weighted imaging was used to collect 28 slices perpendicular to the calcarine sulcus, which covered the entire occipital lobe as well as the posterior parietal and temporal cortex (TR, 2000 ms; TE = 35 ms; flip angle, 80º; FOV 192 x 192; slice thickness, 3 mm (no gap); in-plane resolution, 3 x 3 mm). Participants used a custom-made bite bar to stabilize head position and minimize motion.

Functional MRI data preprocessing. All fMRI data underwent three-dimensional (3D) motion correction using automated image registration software[31]. This was followed by slice scan-time correction to correct for the different times of slice acquisition, and linear trend removal to eliminate slow drifts in signal intensity. No spatial or temporal smoothing was directly applied to the data. (Note however, that a small degree of spatial blurring would be expected to result from our preprocessing steps of motion correction, spatial realignment and Talairach transformation, and data reinterpolation.) The fMRI data were aligned to retinotopic mapping data collected from a separate session, using Brain Voyager software (Brain Innovation). All automated alignment was subjected to careful visual inspection and manual fine-tuning to correct for any potential residual misalignment. Rigid-body transformations were performed to align fMRI data to the within-session 3D anatomical scan, and then to the retinotopy data. After across-session alignment, fMRI data underwent Talairach transformation and reinterpolation using 3 x 3 x 3 mm voxels. This procedure

allowed us to delineate individual visual areas on flattened cortical representations and to restrict the selection of voxels around the grey-white matter boundary.

Retinotopic mapping of visual areas. Retinotopic visual areas of each subject were delineated in a separate experimental session using well-established methods[32-35]. Subjects maintained fixation while viewing 'traveling wave' stimuli consisting of rotating wedges and expanding rings, which were used to construct phase-encoded retinotopic maps of polar angle and eccentricity, respectively. Boundaries between visual areas were delineated on flattened cortical representations using field-sign mapping, which identifies reversals in polar-angle phase encoding relative to topographic changes in eccentricity phase encoding[34].

Regions of interest. Voxels used for orientation decoding analysis were selected bilaterally from the cortical surface of areas V1, V2, V3, V3A, and V4. First, voxels near the grey-white matter boundary were identified within each visual area using retinotopic maps delineated on a flattened cortical surface representation. Next, voxels were sorted according to the reliability of their responses to the visual field localizer using a *t*-statistic. To facilitate comparison of decoding performance across visual areas, we wished to select an equal number of voxels from each area while ensuring that all selected voxels were highly responsive to the visual localizer stimulus. We used the 120 most active voxels from each of V1, V2, V3 (minimum *t* of 8.55 or greater for every subject), as well as V3A/V4 combined. We combined V3A and V4 because fewer voxels were available in these areas; individually, they showed similar levels of decoding performance. For the V1-V4 region of interest, we selected the 480 most active voxels across the entire region.

After completing the analyses described below, we confirmed that decoding performance was reliable with varying numbers of voxels from each region (**Supplementary Fig. 7**). In general, orientation decoding tended to improve with increasing voxel number and showed high levels of performance when 120 voxels from each visual area were selected.

fMRI data samples used for decoding. For the working memory experiment, fMRI data samples were produced by averaging the activity of individual voxels across time points 6-10s (i.e., TRs 4-6) after the start of each trial **(Fig. 1b**, grey region). We selected the start point of this time window to account for the hemodynamic lag of the BOLD response (4-6s); given that the cue appeared at 1200 ms, some BOLD activity associated with processing of the cue would be expected to emerge by a time of 6s. We adopted a conservative strategy in selecting the end point of 10s; this prevented the possible inclusion of any BOLD activity associated with the presentation of the test grating at time 13s (which, in principle, could begin to influence fMRI activity partway through the acquisition of TR 7 and beyond). Thus, an fMRI sample for a working memory trial consisted of a “spatial pattern” of time-averaged activity spanning all voxels within the region of interest.

For the unattended gratings experiment, fMRI data samples were created by averaging activity over each 16-s stimulus block, after accounting for a 4-s delay in the BOLD response. For classification analysis of individual fMRI time points, no temporal averaging was performed.

All fMRI data were transformed from MRI signal intensity to units of percent signal change, calculated relative to the average level of activity for each voxel across all samples within a given run. We also performed spatial normalization to

remove potential differences in overall amplitude within each visual area. Here, the amplitude of all selected voxels in the region of interest were transformed into z-scores, so that the mean activity across the voxels was set to 0 and the standard deviation was set to 1. (For reference, we found that decoding accuracy in the main working memory experiment and the unattended gratings experiment was identical for fMRI data with and without spatial normalization for every visual area and subject.)

All fMRI data samples for a given experiment were labelled according to the corresponding orientation, and served as input to the orientation classifier.

Decoding analysis. A variety of methods have been proposed for pattern classification analysis of fMRI data[36-42]. Here, we used methods previously developed in our lab to decode orientation-selective responses in the human visual cortex[43]. fMRI activity patterns were analyzed using a linear classifier to predict the orientation that was seen (unattended gratings experiment) or held in memory (working memory experiment). Linear support vector machines (SVM)[44] were used to obtain a linear discriminant function that could distinguish between the two orientations. Mathematically, this function can be expressed by:

$g(x) = \mathbf{w}_i\mathbf{x}_i + w_o$

where $\mathbf{x}_i$ is a vector specifying the fMRI amplitude of the voxel *i*, $\mathbf{w}_i$ is a vector specifying the weight of each voxel *i*, and w_o is the overall bias. For a training data set, linear SVM computes the optimal weights and bias for the discriminant function, such that this discriminant function, $g(x)$ satisfies:

$g(x) > 0$ when fMRI activity is induced by one orientation

$g(x) < 0$ when fMRI activity is induced by the other orientation

To evaluate orientation-classification performance, we performed an *N*-fold cross-validation procedure using independent samples for training and testing. This involved dividing the data set into *N* pairs of 25º and 115º trials (or stimulus blocks for the unattended gratings experiment), training the classifier using data from *N*-1 pairs, and then testing the decoder on the remaining pair. We performed this validation procedure repeatedly until all pairs were tested, to obtain a measure of classification accuracy for each orientation case and subject.

For the time-resolved decoding analysis, we used a leave-one-run-out procedure for cross validation, since temporally adjacent fMRI time points are not fully independent of one another. Additionally, for this analysis, the linear classifier was trained using the data from time points corresponding with TRs 3–8, but then tested on all eight TRs. We avoided training on TRs 1–2 since relevant BOLD activity would have yet to evolve because of hemodynamic lag.

Supplementary Analyses

Post-experimental participant reports. After each experimental fMRI session, participants were asked to discuss the strategies they had used to perform each task. For the main working memory experiment as well as the random-orientation variant of this task, all participants reported relying on an effortful strategy of maintaining a representation of the cued grating throughout the delay period, and then comparing this representation to the orientation of the test grating.

For the experiment consisting of mixed trials of working memory and immediate report, participants reported that immediate report trials led to brief cognitive engagement. After seeing the two sample gratings, if the cue indicated

immediate report, participants reported briefly recalling the appearance of the cued orientation to judge its orientation relative to the vertical axis. After completing this response, however, they made no further effort to sustain this representation of the cued grating.

In the visual expectancy experiment, most participants reported that they interpreted the cue, which could momentarily call to mind a representation of the upcoming orientation, but they did not feel compelled to maintain this representation throughout the delay period. Instead, participants reported they were able to make a discrimination judgment immediately after the test grating was presented, by relying on long-term memories of the central tendencies of the two sets of orientations.

Orientation decoding accuracy on correct versus incorrect trials. For our decoding analysis of the main working memory experiment, we included all trials, regardless of whether participants could discriminate small differences in orientation correctly or incorrectly. This was based on the assumption that participants always maintained the cued orientation in working memory on every trial, even if they could not do so with the necessary precision to respond correctly in this near-threshold discrimination task. Ideally, if fMRI decoding were sufficiently sensitive to discriminate between orientations just a few degrees apart, then we might expect to be sensitive to error trials in a near-threshold task. Participants had the opportunity to report at the end of every working memory run whether they had failed to perform the task on any trials, or had accidentally remembered the uncued grating. For most subjects, neither of these situations ever arose.

For the sake of comparison, below we report mean decoding accuracy on correct vs. incorrect trials, respectively: 81.5% vs. 83.8% in V1-V4 pooled, 73.2% vs. 74.2% in V1, 78.7% vs. 71.6% in V2, 78.2% vs. 72.1% in V3, and 70.6% vs. 71.6% in V3A/V4. A repeated-measures ANOVA across these areas failed to reveal a significant difference in decoding accuracy between correct and incorrect trials ($F(1,5) = 0.42$, $P = 0.5$).

Eye-tracking control experiment. Participants completed an additional session outside of the fMRI scanner in which eye movements were monitored using an Applied Science Laboratories XC-HR50 eye-tracking system. They completed ten working memory runs, identical to those performed in the scanner.

In general, participants maintained good fixation. Eyes were well centered during the working memory task, with mean x and y positions of 0.09° and -0.02°, respectively. More important, there was negligible evidence of eye movements in our observers, as the average standard deviation in eye position across subjects was just fractions of a degree (SD for horizontal and vertical axes, 0.19° and 0.30°, respectively).

We also determined whether small systematic eye movements could account for our fMRI decoding of remembered orientations, by submitting these eye-tracking data as input to the orientation classifier. Input to the classifier consisted of the mean x and y positions of the eye, their product, as well as the standard deviation and covariance of these values for the delay period of each working memory trial. We averaged the eye-position information over the time period immediately after the cue was presented up until the time that the test grating appeared. Although eye position measures were very stable and we obtained an equal number of samples of eye movement data as fMRI data,

decoding of the remembered orientation based on eye position was at chance levels (50.2%, $P = 0.94$). Thus, eye position was not a reliable predictor of the orientation held in working memory.

References

[31] Woods, R.P. *et al.* Automated image registration: I. General methods and intrasubject, intramodality validation. *J. Comput. Assist. Tomogr.* **22**, 139-152 (1998).

[32] DeYoe, E. A. *et al.* Mapping striate and extrastriate visual areas in human cerebral cortex. *Proc Natl Acad Sci U S A* **93**, 2382-2386. (1996).

[33] Engel, S. A., Glover, G. H., & Wandell, B. A. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cereb. Cortex* **7**, 181-192. (1997).

[34] Sereno, M.I. *et al.* Borders of multiple visual area in humans revealed by functional magnetic resonance imaging. *Science* **268**, 889-893 (1995).

[35] Wandell, B. A., Dumoulin, S. O., & Brewer, A. A. Visual field maps in human cortex. *Neuron* **56**, 366-383 (2007).

[36] Haxby, J. V. *et al.* Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425-2430 (2001).

[37] Cox, D. D. & Savoy, R. L. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* **19**, 261-270 (2003).

[38] Mitchell, T. M. *et al.* Classifying instantaneous cognitive states from FMRI data. *AMIA Annu Symp Proc*, 465-469 (2003).

[39] Haynes, J. D. & Rees, G. Decoding mental states from brain activity in humans. *Nat Rev Neurosci* **7**, 523-534 (2006).

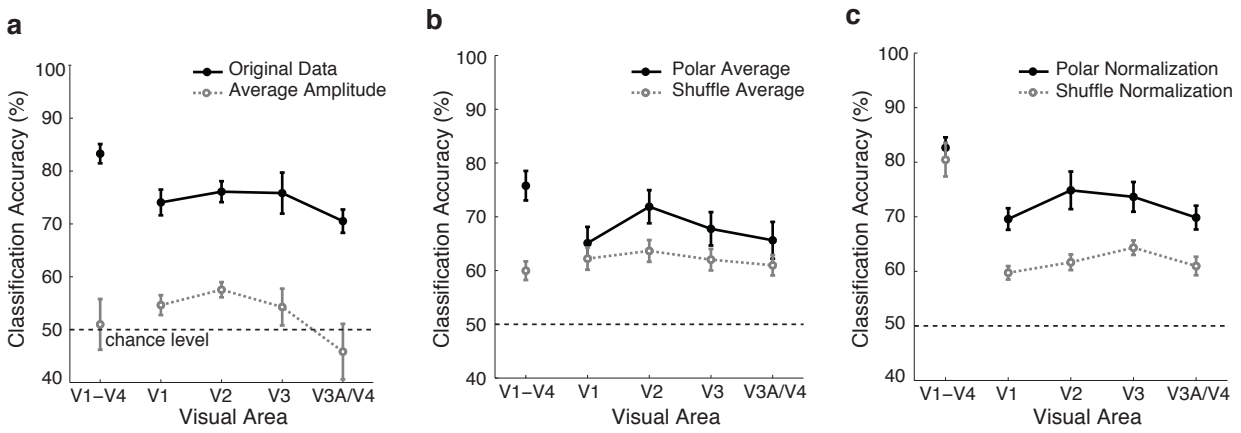
[40] Kriegeskorte, N., Goebel, R., & Bandettini, P. Information-based functional brain mapping. *Proc Natl Acad Sci U S A* **103**, 3863-3868 (2006).

[41] Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* **10**, 424-430 (2006).

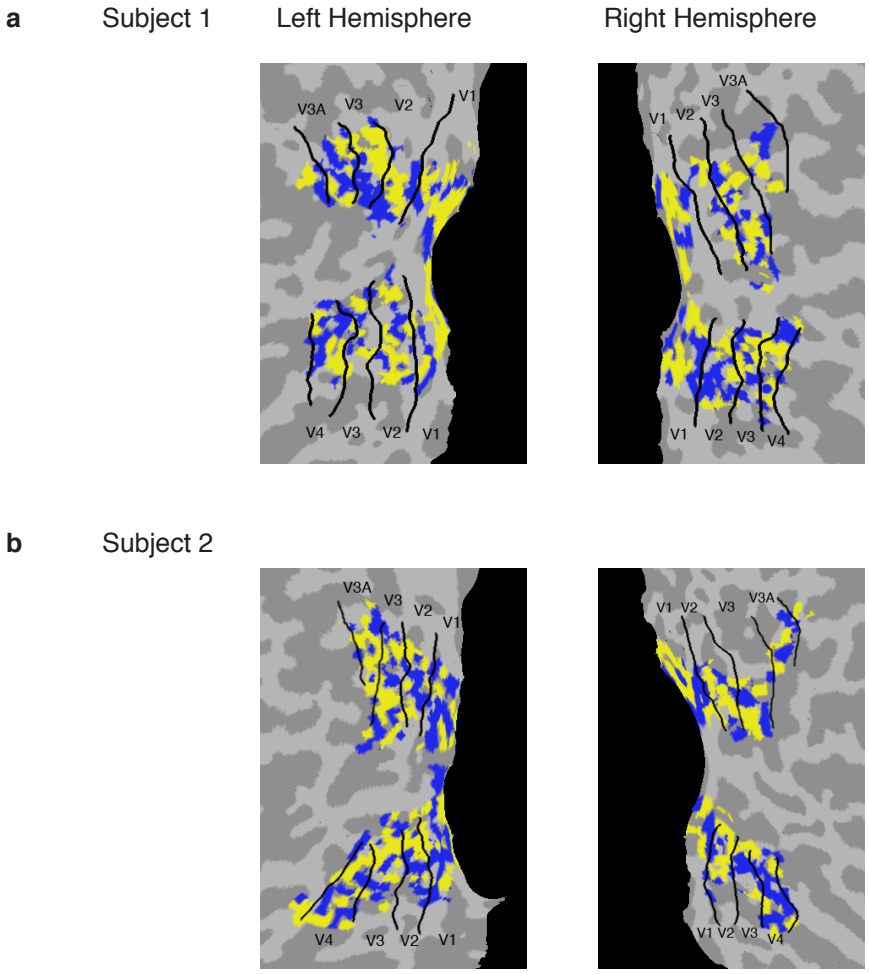
[42] Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. Identifying natural images from human brain activity. *Nature* **452**, 352-355 (2008).

[43] Kamitani, Y. & Tong, F. Decoding the visual and subjective contents of the human brain. *Nat Neurosci* **8**, 679-685 (2005).

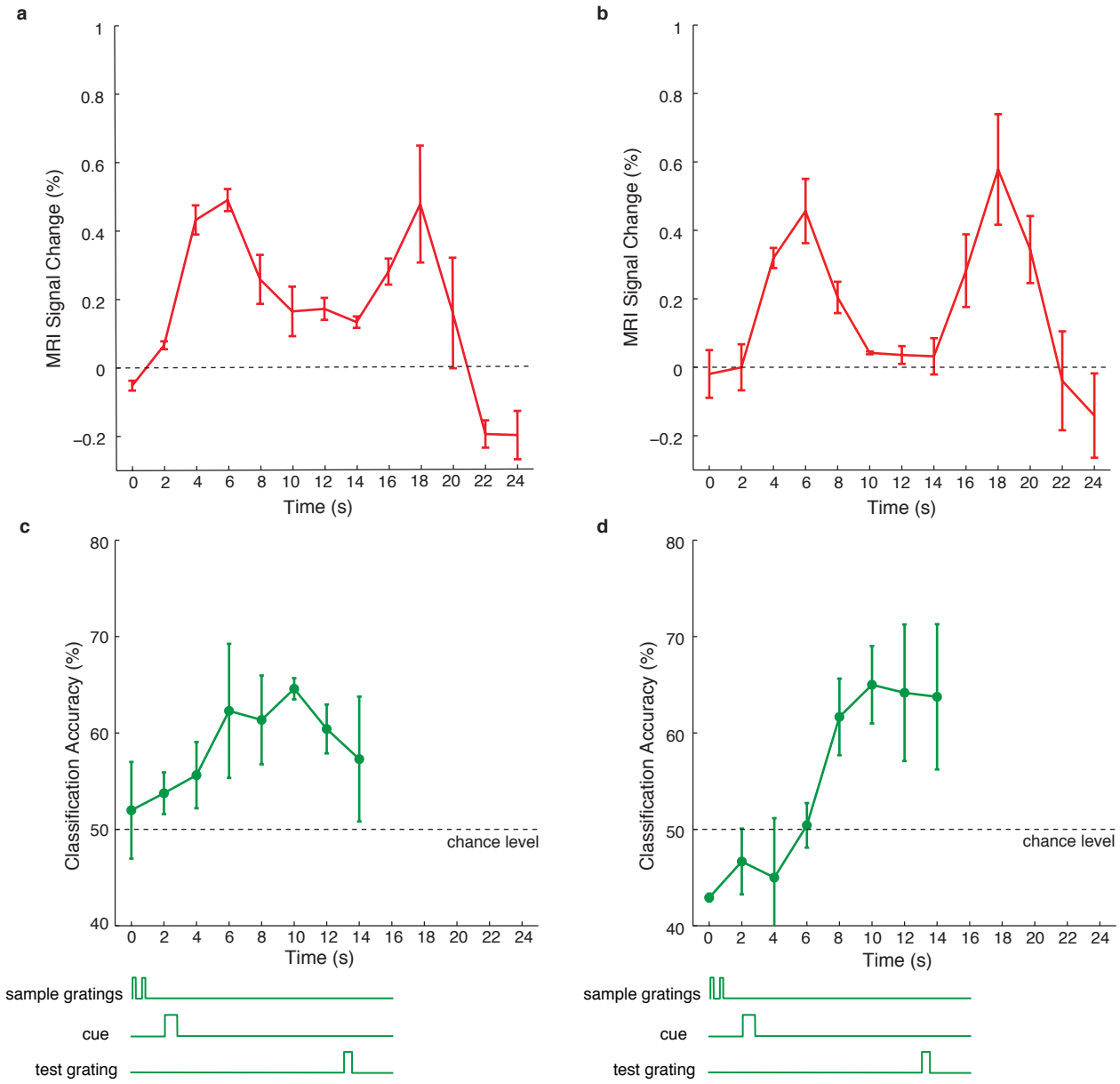
[44] Vapnik, Vladimir Naumovich *Statistical learning theory*. (Wiley, New York, 1998).



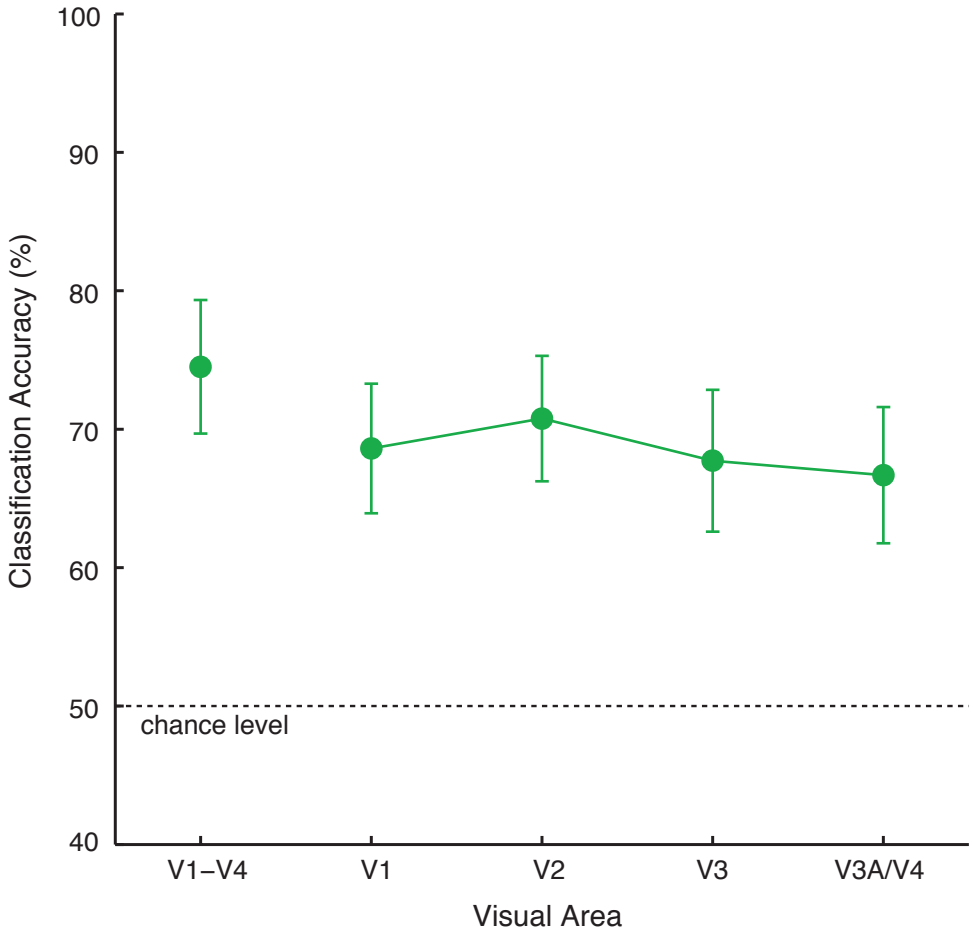
Supplementary Figure 1. Contribution of global and local fMRI signals to orientation decoding during working memory. a, Comparison of decoding applied to original data and the averaged response of each visual area. Classification accuracy was significantly worse after the response of all originally selected voxels was averaged to obtain the mean response amplitude of each region of interest ($F = 43.72$, $P < 0.005$). Performance fell to chance level for most areas, with only V2 exhibiting above-chance decoding ($T = 5.23$, $P < 0.005$). **b,** Decoding of global radial bias signals. Neighboring voxels corresponding to different radial segments of the visual field were sorted according to their polar-angle preference, as determined by a separate retinotopic mapping session. Using 16 polar angles to divide the visual field, we calculated the average response of all isopolar voxels and submitted the 16 averaged responses to the orientation decoder. Decoding of this global spatial activity ("polar average" condition), though above chance, was significantly worse than for the original data ($F = 33.22$, $P < 0.005$), indicating that global radial bias does not capture the full amount of orientation information that is available in these activity patterns. For comparison, we randomly shuffled the polar-angle assignment of the 120 voxels in each visual area, and performed the same analysis on their averaged responses ("shuffle average"). These data also led to above-chance orientation decoding performance, indicating that even after averaging, arbitrary sets of voxels with unrelated orientation biases still contain some bias in their combined response. The modest advantage in decoding performance for the polar average condition, as compared to the shuffle average condition, reflects the amount of orientation information available in the global radial component. **c,** Decoding of local spatial patterns. We examined classification of local orientation-selective responses within a polar angle, independent of the global activity pattern. The average response amplitude of a given polar wedge was subtracted from each voxel exhibiting that polar angle preference, to minimize the potential contribution of global radial bias. Decoding of these local signals was highly effective ("polar normalization" condition), comparable to that seen for the original data ($F = 1.16$, $P = 0.33$) and significantly better than decoding of the polar average data ($F = 9.96$, $P < 0.05$). By comparison, much poorer performance was obtained when we subtracted out the mean signal of polar-shuffled voxels (i.e., "shuffle average" data) from the original data, likely because some genuine orientation signals were removed by this procedure. Note that no spatial normalization was applied in the control analyses described above (with the exception of the original data). For the shuffle analyses, we performed 100 iterations of the randomization procedure to determine classification accuracy for each region of interest and subject. Error bars indicate ±1 S.E.M.



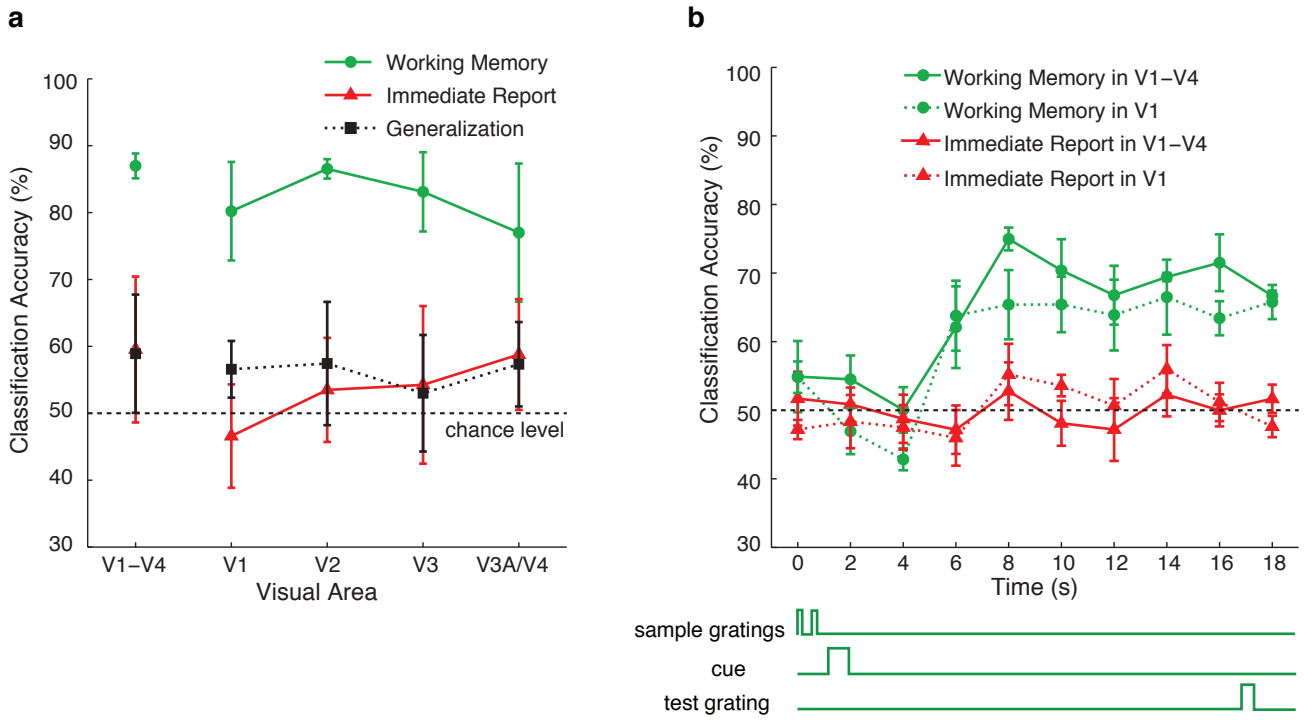
Supplementary Figure 2. Orientation preference of individual voxels plotted on the flattened cortical surface. Decoding was performed separately on areas V1, V2, V3, and V3A/V4, using the 120 most-active voxels from each region. Weights resulting from the trained classifier were used to indicate the preferred orientation of each voxel (yellow 25º; blue 115º). Maps of the left and right visual cortices are shown for two representative subjects. Visual inspection of the spatial arrangement of orientation preference suggests the presence of considerable local variability.



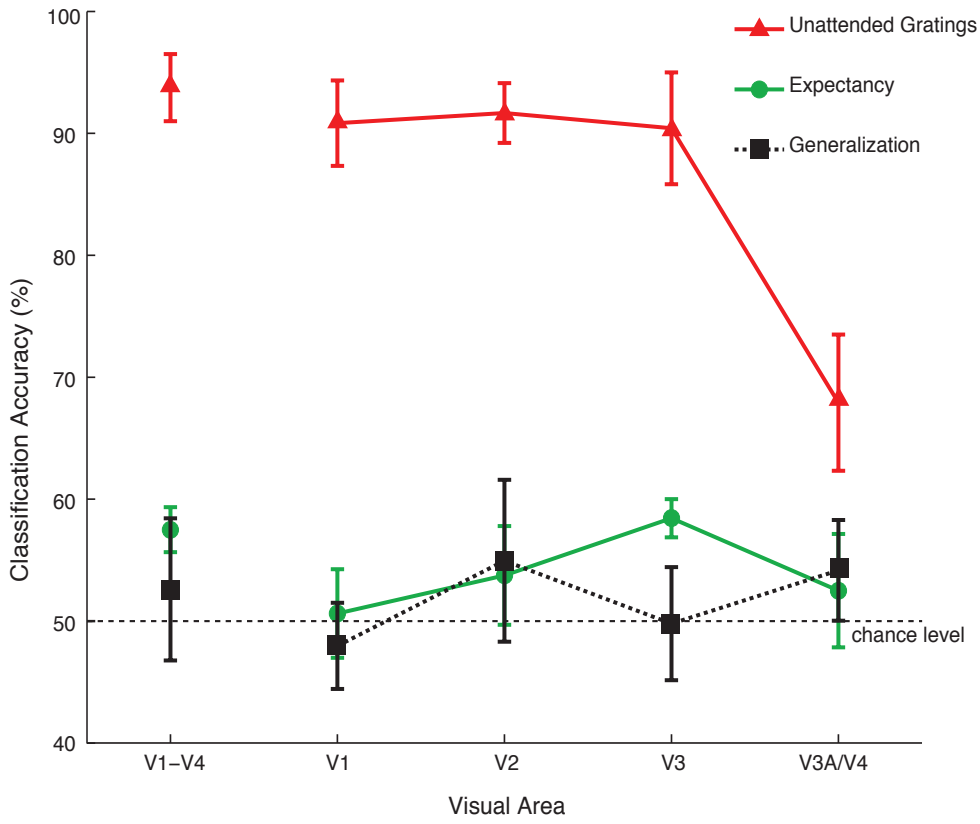
Supplementary Figure 3. Comparison of BOLD amplitudes and decoding accuracy for delay-period activity in V1. Time course of mean BOLD activity in V1 during the working memory task for subjects who showed sustained activity (**a**) or negligible activity (**b**) towards the end of the delay period. Subjects were grouped according to whether their individual V1 activity fell to baseline levels between time 10-14s. The three subjects in (**a**) exhibited activity that was significantly above baseline levels at each of time points 10, 12, and 14s (min $T = 2.89$, $P < 0.005$ in all cases). Conversely, by 12s all three subjects in (**b**) exhibited activity that was not significantly greater than fixation baseline (max $T = 1.10$, $P > 0.25$ in all cases). Orientation decoding of individual fMRI time points for the subjects with reliable delay-period activity (**c**) and negligible delay-period activity (**d**). Error bars indicate ±1 S.E.M.



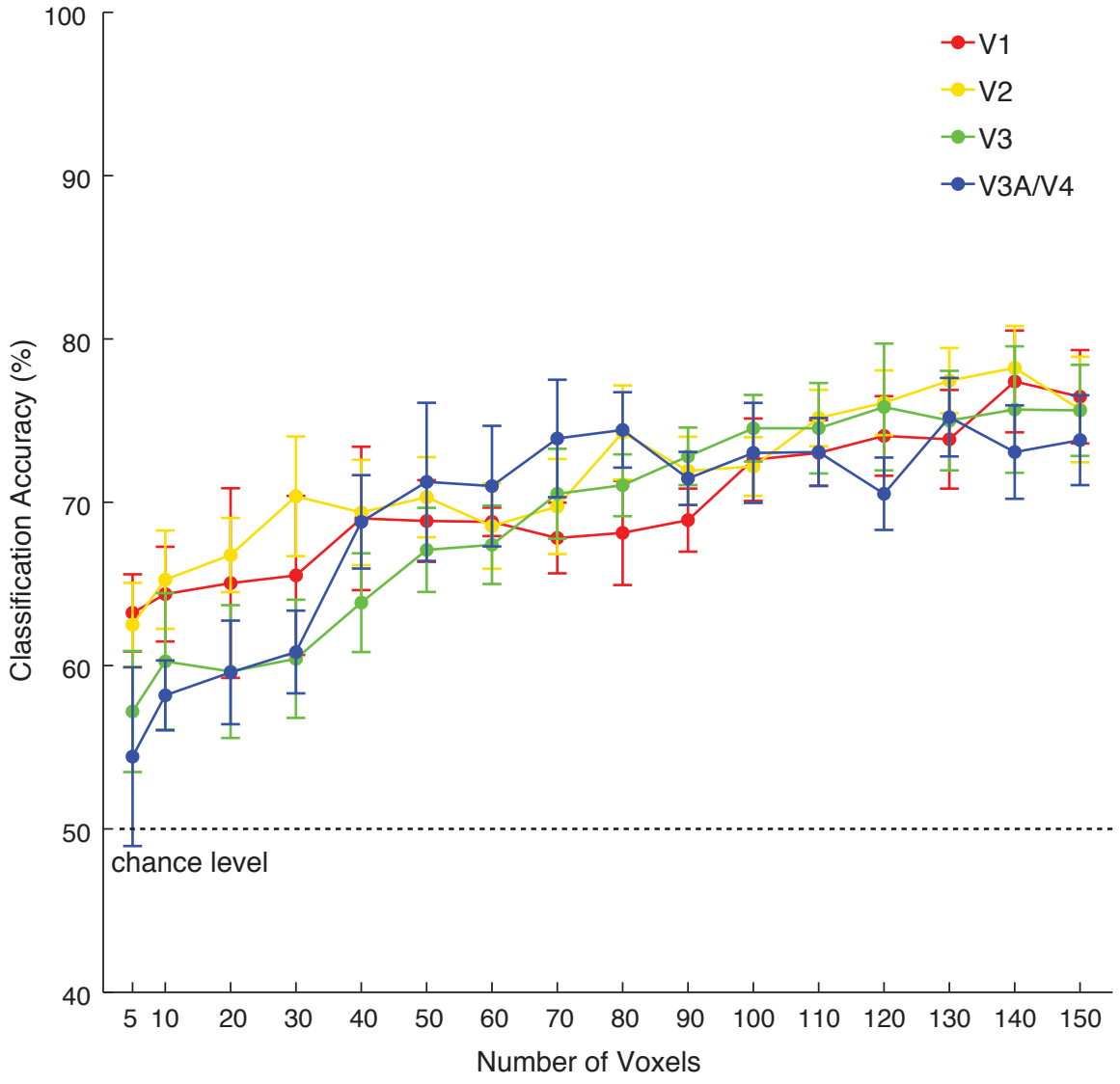
Supplementary Figure 4. Decoding accuracy for random orientations held in working memory. Four subjects participated in an additional working memory experiment that required remembering a different random orientation on every trial. The experimental design and timing of stimulus events were identical to the original experiment (**Fig. 1a**), except that randomly selected pairs of orientations (80-100º apart) were presented on every trial and the test grating was rotated ±5º or ±10º relative to the cued orientation. Participants performed 16 runs (128 trials) over the entire fMRI session, with mean behavioral performance of 76.2% correct. Orientation decoding accuracy was assessed using a leave-one-run-out cross-validation procedure, based on the specific orientations of the cued and uncued gratings for each test sample. Specifically, labels were assigned to each training sample based on whether the cued orientation on that trial fell within ±35º of the cued orientation of the test sample (label 1) or within ±35º of the uncued orientation (label 2). Any orientations that fell outside of this range were excluded from decoding analysis. The figure shows average orientation decoding performance plotted by visual area; error bars indicate ±1 S.E.M. Averaging over a time window of 6–10s, decoding of the orientation held in working memory was well above chance for every visual area at the group level (min $T = 3.39$, $P < 0.05$). Additionally, all individual subjects exhibited reliable decoding for areas V1– V4 pooled, V1, and V2 (performance exceeding 58.5%, $P < 0.05$, two-tailed binomial test). In areas V3 and V3A/V4, three out of four subjects surpassed this criterion. Decoding performance for random orientations held in working memory was slightly lower than that found in the original working memory experiment, but not significantly different ($F = 3.68$, $P = 0.15$).



Supplementary Figure 5. Sustained orientation-selective activity found during working memory but not immediate report. Four subjects participated in a separate control experiment in which working memory trials were intermixed with "immediate report" trials. On each trial, participants viewed two sample gratings (~25º and ~115º) followed by a numerical cue. If the cue was green, participants performed the standard working memory task. However, if the cue was red, then participants were instructed to make a speeded response regarding whether the cued orientation was rotated clockwise or counterclockwise relative to the vertical axis. The timing of stimulus events was identical for both trial types, although participants were instructed not to make a response to the test grating on immediate report trials. Participants performed 16 runs in the fMRI scanner, with mean behavioral performance of 76.0% on the working memory task and 93.5% on the immediate report task. Decoding was performed on averaged fMRI activity from 6–14s to accommodate the lengthier delay period of 15s used in this experiment. **a,** Orientation decoding accuracy for both experiments, as well as generalization performance between the two experiments. Decoding performance for the working memory task (green curve) was well above chance in areas V1-V4 pooled, V1, V2, and V3 at the group level (min $T = 4.74$, $P < 0.05$), and significant in every visual area for individual subjects (performance exceeding 62.5%, $P < 0.05$, two-tailed binomial test). By comparison, immediate report trials (red curve) failed to elicit reliable orientation-selective activity (max $T = 1.22$, $P > 0.3$). Overall, orientation decoding was significantly better on working memory trials than immediate report trials ($F = 19.7$, $P < 0.05$). **b,** Orientation decoding of individual fMRI time points for areas V1-V4 pooled and V1. In the working memory task, both regions of interest showed a significant increase in orientation decoding performance over time (min $F = 7.34$, $P < 0.001$); performance reached asymptotic levels within the first 8 seconds and then remained high throughout the rest of the delay period. Decoding in the immediate report condition remained at chance-level performance throughout the trial. These results indicate that merely selecting the cued grating from memory is not sufficient to obtain sustained orientation-selective responses; instead, active maintenance in working memory appears to be critical. Error bars indicate ±1 S.E.M.



Supplementary Figure 6. Orientation decoding results for expectancy control experiment. Four subjects participated in an additional imaging session, which consisted of 10 runs of a control experiment resembling that of the working memory task, and 5 runs of the unattended gratings experiment. The purpose of the control experiment was to determine whether orientation-selective activity during the delay might reflect the participants' anticipation of the test grating as opposed to the maintenance of the task-relevant sample grating. The timing of stimulus events was identical to the working memory experiment (**Fig. 1a**), except that no sample gratings were shown. Instead, participants were shown a colored numerical cue indicating which of the two approximate angles (~25° or ~115°) would be presented at the end of the trial. After presentation of the test grating, participants were required to judge whether this stimulus was likely an example of a clockwise or counterclockwise rotation (behavioural performance, 77.5% correct). To successfully complete this task, participants had to rely on long-term memory representations of the central tendency of the two base orientations. Classification accuracy for the unattended gratings experiment was once again highly significant in all visual areas, corroborating our results from the main experiment (V1–V4 pooled: 92.75%, $P < 0.01$). By contrast, orientation decoding performance was very poor for the visual expectancy experiment, and failed to reach significance in individual areas V1 and V2 (max $T = 0.93, P > 0.4$). Performance was above chance for V3 and areas V1-V4 pooled, but these regions failed to show reliable generalization across the visual expectancy and unattended gratings experiments. These results suggest that the orientation signals decoded in our main experiment were reflective of the contents of working memory. Error bars indicate ±1 S.E.M.



Supplementary Figure 7. Effect of voxel number on decoding accuracy for orientations held in working memory. Decoding accuracy gradually improved as a function of voxel number for each visual area, reaching near-asymptotic performance at about 100-150 voxels. We used 120 voxels per visual area for all subsequent decoding analyses, since all voxels in V1-V3 showed strong responses to the visual-field localizer at this cutoff (with *t*-values of 8 or greater) and decoding performance was at near-asymptotic levels. Error bars indicate ±1 S.E.M.