Supporting Information

Stokes et al. 10.1073/pnas.0905306106

SI Text

Multivoxel Pattern Analysis. MVPA was performed using a linear classifier that is a variant of correlation-based nearest-neighbor classification (1–4). Unlike other approaches to neural classification, including distance metrics, linear discriminant analysis, and support vector machines, correlation-based classification is mean-invariant. Therefore, pattern analysis is not influenced by global differences in condition-specific activation levels. This is of particular advantage when it is the pattern of differential activity, not magnitude differences, that is used to infer differential population coding (5, 6).

All pattern analyses were performed on minimally preprocessed data to preserve the high spatial frequency information used to characterize differential population codes. Functional images for each participant were spatially aligned, unwarped, and slice-time corrected, but not spatially normalized or smoothed. A gray matter mask was produced for each subject using the segment function in SPM5. Only voxels classified as gray matter were included for MVPA. The time-series data from each voxel were high-pass filtered (cut-off = 128 s) and data from each session were scaled to have a grand mean value of 100 across all voxels and volumes. Using a searchlight procedure (4, 7–9), neural classifications were performed at the location of each cortical voxel, based on the pattern of activity observed within the surrounding cortical volume (radius = 10 mm sphere, containing ≈90 voxels). Classification accuracy for each searchlight was recorded at the central voxel, which repeated for all cortical voxels produced a 3-D accuracy map. Accuracy maps for each participant were then spatially normalized to the MNI template, and assessed via a random-effects group analysis. For more detailed analyses, we also examined MVPA results within predefined regions of interest (ROI) using the MarsBar toolbox (10). To avoid circular inferences (11), ROIs were based on independent data from a previous study that identified regions of visual cortex specifically involved in shape-processing (12). Spiridon et al. (12) identified separate coordinates for anterior and posterior LOC. We used the peak activation co-ordinates for each of these regions (averaged across hemispheres) to construct spherical ROIs (r = 10 mm) for anterior LOC (aLOC; central co-ordinate ± 37, −51, −14) and posterior LOC (pLOC; central co-ordinate ± 45, −82, −2). All pattern-analyses were implemented in Matlab using customized procedures, and SPM 5 (Wellcome Department of Cognitive Neurology, London, U.K.) was used for image handling and second-level analyses.

Localizer Analysis. Initially, we trained a set of linear classifiers to discriminate between the two experimental conditions in the perceptual task: viewing X and O. This was performed separately within each searchlight sphere using a leave-one-out cross-validation procedure. Firstly, beta estimates were obtained for each scanning run using the same model used for the univariate analysis (see *Materials and Methods*). For each iteration of the cross-validation, data from 5 scanning runs were then used to construct a "training" pattern, whilst data from the remaining scanning run were used to construct a "test" pattern. The training pattern was constructed by first averaging the beta estimates from the five scanning runs of the current training set. A difference pattern representing the relative activation bias associated with viewing "X" or "O" was then calculated by subtracting the mean beta estimates for the two conditions in each voxel. Next, this difference pattern was divided by a squared estimate of the voxel-wise error covariance matrix, effectively weighting each voxel according to the multivariate structure of the noise. This was derived by first calculating the residual variance (i.e., observed–fitted data) of the training data for each voxel within the current searchlight sphere. We then used a shrinkage estimator (13) to calculate the covariance of these residual values, as this provides a more stable estimate than the sample covariance for data sets with many variables and/or few observations (4, 8). The resulting pattern of voxel biases defined the training pattern for each neural classifier. After deriving the training pattern, a statistically independent test pattern was obtained using exactly the same procedure on the beta estimates from the remaining session. Finally, classification accuracy was tested using a correlation-based procedure (1–4). The correlation between training and test patterns was calculated, and coefficients above zero were coded as correct classifications, whereas coefficients of zero or below were recorded as incorrect classifications. The leave-one-out procedure was then repeated for each of the remaining five possible train-test permutations. The overall classification accuracy value for each searchlight sphere was then calculated by taking the percentage of correct classifications across all 6 iterations of the cross-validation procedure.

Attention Task. The principal aim of this study was to explore attentional activation of perceptual representations. More specifically, we sought to test whether the same neural populations that are selectively activated during perception of X or O are also selectively activated while participants are attending for X or O. Therefore, the critical test of our hypothesis is the extent to which discriminative patterns observed during attention match the discriminative patterns observed during the pattern localizer. More conventional leave-one-out cross-validation can only demonstrate that there is *something* reliable that differentiates the two conditions. Because *any* differences in activity could lead to accurate classifications as long as they are reproducible across observations, it is necessary to constrain the range of possible inferences to test the current hypothesis. Therefore, we applied a variant of the cross-validation procedure described above that explicitly tests the prediction that neural populations for seeing X will be activated during preparatory attention for X, whereas populations for seeing O will be activated during attention for O. The cross-comparison procedure was the same as the MVPA procedure for the pattern localizer, except here the classifier was trained using data from all sessions of the localizer task and tested against data from the attention task. For each searchlight sphere, data from the localizer task were used to construct a training pattern representing the relative activation patterns associated with viewing X or O. This was done with data from all six scanning runs using a procedure identical to that described above. Next, beta parameters were estimated for the attentional task using the FIR design matrix described for the univariate analyses (see *Materials and Methods*). Importantly, all time points following the onset of any letter stimulus were discarded to ensure that our results were not contaminated by stimulus-driven visual information. Test patterns representing the relative activation associated with attending for X or O (averaged across all six scanning runs) were then calculated for each time bin by subtracting the pattern of beta values for the attend O condition from the pattern of beta values associated with the attend X condition. These difference patterns were then divided by the square of the residual covariance matrix, and the weighted patterns were used to test the cross-comparison between per-

ception and attention at each time point. As before, a positive correlation between the discriminative pattern for perception and the equivalent pattern observed during attention was coded as a correct cross-comparison classification, whereas zero or negative coefficients were coded as incorrect classifications. This procedure generated an estimate of the pattern-match between stimulus-driven perception and top-down preparatory attention for each 2-s time bin of the attention trial. A positive match between the perceptual and attentional task directly implies the presence of shared features across the population response. The results of the full searchlight analysis were then reconstructed to form perception-to-attention cross-comparison accuracy maps at each time point and for each participant. These were then normalized to the standard MNI template. Time-course analyses were examined using the aLOC and pLOC ROIs described above. An overall summary classification score was also calculated for each subject by averaging the accuracy data across the eight time bins (16 s) beginning 4 s after the cue onset (4- to 20-s post-cue). This summary value was used to assess the results of the full searchlight analysis (Fig. 2*C*). The time-averaged classification score was also used to verify above-chance classification within each predefined ROI (Fig. 3*A*), and to examine the relationship between behavior and attentional bias in aLOC and pLOC (Fig. 3*D*). Finally, to examine the time-course of the brain-behavior relationship (Fig. 3*E*), a more stable down-sampled estimate of the time-course of classification accuracy was calculated by averaging together classification scores from consecutive time points to create a series of new, 4-s time bins.

As noted above, data from all time-points following with the presentations of letter stimuli were discarded from analyses of preparatory attention. However, for completeness, we also performed a separate analysis to examine the time-course of the pattern-specific response elicited by target stimuli. We used the same cross-comparison approach described above with the exception that target-specific patterns defined using the localizer were compared to activation patterns time-locked to the onset of the semitransparent target stimuli presented during the attention task. Second-level analyses were performed on pattern-classification scores for aLOC and pLOC at each 2-s time point spanning 0–14 s from the onset of the target stimulus.

1. Haxby JV, et al. (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293:2425–2430.
2. Williams MA, et al. (2008) Feedback of visual object information to foveal retinotopic cortex. *Nat Neurosci* 11:1439–1445.
3. Williams MA, Dang S, Kanwisher NG (2007) Only some spatial patterns of fMRI response are read out in task performance. *Nat Neurosci* 10:685–686.
4. Stokes M, Thompson R, Cusack R, Duncan J (2009) Top-down activation of shape-specific population codes in visual cortex during mental imagery. *J Neurosci* 29:1565–1572.
5. Haynes JD, Rees G (2006) Decoding mental states from brain activity in humans. *Nat Rev Neurosci* 7:523–534.
6. Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* 10:424–430.
7. Haynes JD, et al. (2007) Reading hidden intentions in the human brain. *Curr Biol* 17:323–328.
8. Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. *Proc Natl Acad Sci USA* 103:3863–3868.
9. Soon CS, Brass M, Heinze HJ, Haynes JD (2008) Unconscious determinants of free decisions in the human brain. *Nat Neurosci* 11:543–545.
10. Brett M, Anton J-L, Valabregue R, Poline J-B (2002) Region of interest analysis using an SPM toolbox [abstract]. *8th International Conference on Functional Mapping of the Human Brain.*
11. Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI (2009) Circular analysis in systems neuroscience: The dangers of double dipping. *Nat Neurosci* 12:335–340.
12. Spiridon M, Fischl B, Kanwisher N (2006) Location and spatial profile of category-specific regions in human extrastriate cortex. *Hum Brain Mapp* 27:77–89.
13. Ledoit O, Wolf M (2003) Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J Empir Finance* 10:603–621.