Behavioral/Cognitive

Biased Neural Representation of Feature-Based Attention in the Human Frontoparietal Network

Mengyuan Gong[1,2] and Taosheng Liu[2]
[1]Department of Psychology and Behavioral Sciences, Zhejiang University, Hangzhou 310028, Zhejiang, China, and [2]Department of Psychology, Michigan State University, East Lansing, Michigan 48824

Selective attention is a core cognitive function for efficient processing of information. Although it is well known that attention can modulate neural responses in many brain areas, the computational principles underlying attentional modulation remain unclear. Contrary to the prevailing view of a high-dimensional, distributed neural representation, here we show a surprisingly simple, biased neural representation for feature-based attention in a large dataset including five human fMRI studies. We found that when human participants (both sexes) selected one feature from a compound stimulus, voxels in many cortical areas responded consistently higher to one attended feature over the other. This univariate bias was consistent across brain areas within individual subjects. Importantly, this univariate bias showed a progressively stronger magnitude along the cortical hierarchy. In frontoparietal areas, the bias was strongest and contributed largely to pattern-based decoding, whereas early visual areas lacked such a bias. These findings suggest a gradual transition from a more analog to a more abstract representation of attentional priority along the cortical hierarchy. Biased neural responses in high-level areas likely reflect a low-dimensional neural code that can facilitate a robust representation and simple readout of cognitive variables.

Key words: **feature-based attention; fMRI; frontoparietal network; neural representation**

Significance Statement

It is typically assumed that cognitive variables are represented by distributed population activities. Although this view is rooted in decades of work in the sensory system, it has not been rigorously tested at different levels of cortical hierarchy. Here we show a novel, low-dimensional coding scheme that dominated the representation of feature-based attention in frontoparietal areas. The simplicity of such a biased code may confer a robust representation of cognitive variables, such as attentional selection, working memory, and decision-making.

Introduction

How neural activities represent sensory and cognitive information is a fundamental question in system and cognitive neuroscience. A prevailing view is that the brain uses distributed neural activities to represent information, which is best illustrated by studies in the sensory cortex. For example, neurons in early visual areas have smooth tuning functions that span a range of feature values (Hubel and Wiesel, 1962; Maunsell and Van Essen, 1983). A single stimulus feature would evoke a profile of population response across a group of such neurons. Computational studies have demonstrated that such population responses can be used for encoding and decoding sensory information (Pouget et al., 2000). Consistent with these neuronal level studies, human functional magnetic resonance imaging (fMRI) studies have shown that patterns of BOLD responses can be used to decode and reconstruct visual stimulus (Kamitani and Tong, 2005; Kay et al., 2008).

Although there is a general consensus that stimulus properties are represented via distributed population activity in sensory areas, much less is known about how cognitive variables are represented. Cognitive functions related to task control and target selection have been associated with activity in parietal and prefrontal cortical areas, collectively known as the multiple-demand network (Duncan, 2010; Woolgar et al., 2016). Neurons in this network can flexibly adapt to task demands (Duncan, 2001; Jackson and Woolgar, 2018), and they have been shown to encode task rules (Freedman, et al., 2001; Stokes et al., 2013), attentional priority (Bisley and Goldberg, 2003), and learned category (Swaminathan and Freedman, 2012). Consistent with these electrophysiological studies in nonhuman primates, many human fMRI studies have decoded these cognitive variables using multivariate BOLD response patterns in this network (Li et

Received Mar. 25, 2020; revised Aug. 21, 2020; accepted Sep. 23, 2020.

Author contributions: M.G. and T.L. designed research; M.G. and T.L. performed research; M.G. and T.L. contributed unpublished reagents/analytic tools; M.G. and T.L. analyzed data; M.G. and T.L. wrote the paper.

The authors declare no competing financial interests.

This work was supported by a National Institutes of Health Grant R01-EY-022727. We thank Dr. David Zhu and Scarlett Doyle for assistance in collecting the neuroimaging data. We also thank Michael Jigo for assistance in data collection and analysis.

Correspondence should be addressed to Taosheng Liu at tsliu@msu.edu.

https://doi.org/10.1523/JNEUROSCI.0690-20.2020

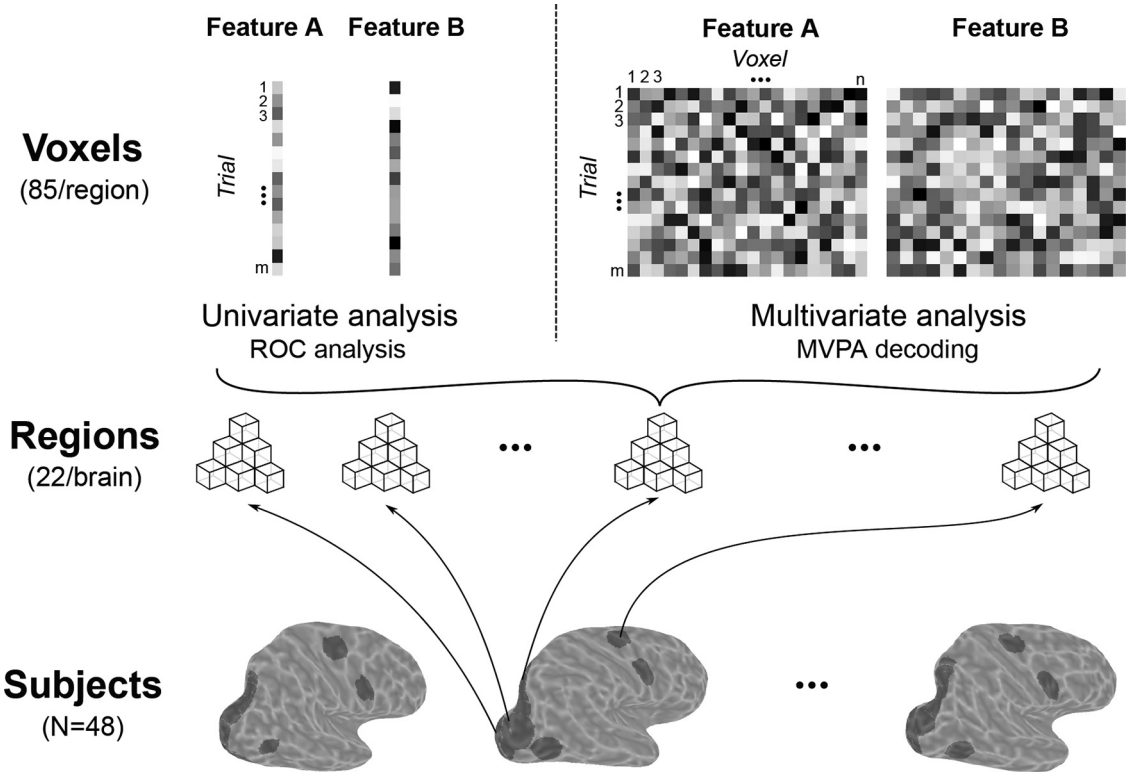

Figure 1. Overview of the data structure and roadmap for the analyses. Bottom row, fMRI data from 48 subjects performing attentional selection tasks were analyzed. Three left hemispheres from different subjects were shown, each overlaid with 11 predefined brain areas (dark shaded areas). Middle row, Each brain contained 22 areas (11 per hemisphere), each of which was composed of individual voxels. Top row, Data structure for the analyses. For multivariate analysis (right), we obtained two matrices of response patterns, each containing BOLD response amplitude from each voxel on each trial in each attention condition. This is an $m \times n$ matrix, where m is the number of trials and n is the number of voxels; m varied across experiments, and n was set to be 85 for the main analysis (for voxel selection, see Materials and Methods). For the univariate analysis (left), we obtained two response vectors by averaging the response patterns for each attention condition across voxels, resulting in an $m \times 1$ vector that contained the average BOLD response from each trial for each condition.

al., 2007; Liu et al., 2011; Liu and Hou, 2013; Erez and Duncan, 2015; Bettencourt and Xu, 2016). Thus, it appears that cognitive variables are also encoded by distributed population activities in high-level brain areas.

Contrary to this population-based view, however, a previous neurophysiological study reported evidence supporting a biased neural code for cognitive variables (Fitzgerald et al., 2013). These researchers performed detailed analysis of lateral intraparietal area (LIP) neuronal activity from several experiments in which monkeys performed a variety of tasks (categorization, associative learning, perceptual decision-making). These tasks all require the subject to respond to a visual stimulus in a discrete manner (e.g., Does this stimulus belong to category A or category B?). The task design further ensured that the measured neural responses reflected cognitive variables instead of signals related to stimulus processing or motor response. Surprisingly, they found that for a given monkey, the majority of recorded LIP neurons showed similar response profiles, consistently exhibiting higher response to one task condition than the other (e.g., higher response to category A than category B). This largely biased population response implies a low-dimensional, rather than high-dimensional, neural representation for cognitive variables.

The finding of a biased response is surprising (Chafee, 2013), and thus it is important to know whether the biased representation can be generalized to other cognitive variables, brain areas, and species. In previous human fMRI studies, cognitive variables can be decoded in high-level brain areas in the absence of any obvious biases in the univariate response. However, these results were obtained by averaging data across subjects, which could obscure a biased response if the direction of bias varied across subjects. Therefore, investigating the existence of biased neural representation requires analyses that take into account the direction of bias for each individual subject. Here, we conducted such analyses on a large fMRI dataset to examine biased neural representations in the human brain.

In these experiments, human subjects attended to one of the two features in a compound stimulus that contained both features, which allowed us to measure attentional signals throughout the brain. We examined potential biases at multiple levels of analysis, as follows: single brain area, multiple brain areas within a subject, and multiple subjects at the group level (Fig. 1). Selective attention is a core operation that enables complex task control (Duncan, 2013) and is highly associated with activity in the frontoparietal network (Corbetta and Shulman, 2002), which constitutes a subset of the multiple-demand network. Thus, attentional signals in the brain provide a good test case for possible biased representation of a cognitive variable. Our analysis revealed a significant bias, mostly in regions within the frontoparietal network, suggesting that low-dimensional neural representations are used in the higher-order areas of human brain to encode cognitive variables such as attentional priority.

Materials and Methods

Participants

In total, 48 subjects (21 females; mean age, 25.1 years) from Michigan State University were included across five experiments. We based our sample size on previous studies using similar attention tasks, details of

which can be found in previous publications. All had normal or corrected-to-normal vision. Participants were paid for their participation at $20/h. All participants gave informed consent according to the study protocol approved by the Institutional Review Board at Michigan State University (LEGACY08-211).

Experimental design and statistical analysis

Overview of the experimental procedures. We reanalyzed data from five previously published fMRI experiments (Liu et al., 2011; Liu, 2016; Jigo et al., 2018; Gong and Liu, 2020). The details of the methods can be found in previous publications, so only an abbreviated description is provided here. All experiments used a similar task design: two features were presented in the same location, and subjects were cued to attend to one of the features on a trial-by-trial basis. The stimulus features varied across experiments, and we will refer to them as feature A and feature B in this report. There are thus two experimental conditions: attend A and attend B. The exact features are as follows. In the first experiment, six subjects attended to a color in a superimposed red-green color display (Liu et al., 2011). In the second and third experiment, 6 and 12 subjects attended to a motion direction in a superimposed clockwise–counterclockwise rotating dot display (Liu et al., 2011; Jigo et al., 2018). In the fourth experiment, 12 subjects attended to a linear motion direction in a superimposed up-left/up-right moving dot field (Gong and Liu, 2020). In the fifth experiment (Liu, 2016), 12 subjects attended to a dynamic object in a superimposed display containing two Gabor patches (object 1 or object 2) that continuously changed their features in multiple dimensions (color, orientation, and spatial frequency). Data for each subject were collected in a single 1.5–2 h scanning session. In total, the dataset contained 48 subjects. The number of trials in each condition (attend A or attend B) varied from 29 to 136 across experiments. The order of conditions was fully randomized.

In all experiments, subjects performed a threshold-level change detection task on the attended feature (e.g., detecting a speedup event in the attended motion direction). Each subject was extensively trained on the task before the scanning session with their performance calibrated by a psychophysical staircase procedure. The task was sufficiently challenging to engage feature selection. In all experiments, we verified that performance did not differ between attend feature A and attend feature B conditions (for details, see previous publications).

Retinotopic mapping. For each subject in each experiment, we ran a separate scanning session of visual field mapping to define visual and parietal topographic areas. We used standard phase-encoded checkerboard stimuli to define retinotopic visual areas (Sereno et al., 1995; DeYoe et al., 1996; Engel et al., 1997) and a memory delay saccade task to map topographic areas in the parietal cortex (Sereno et al., 2001; Schluppeck et al., 2006; Konen and Kastner, 2008). All areas were defined and visualized on computationally flattened representations of the cortical surface, which were generated from high-resolution anatomical images using FreeSurfer (http://surfer.nmr.mgh.harvard.edu) and custom MATLAB code. Detailed descriptions of the mapping procedure can be found in our previous publications. The following regions of interest (ROIs) in each hemisphere were identified with this procedure: V1, V2, V3, V3A/B, V4, V7, MT+, and subregions of the intraparietal sulcus (IPS), IPS1 to IPS4.

Univariate analysis: deconvolution. We used the deconvolution approach by fitting the time series of each voxel with a general linear model whose regressors modeled the two attention conditions with finite impulse responses. The design matrix was pseudoinversed and multiplied by the time series to obtain an estimate of the hemodynamic response evoked by each condition. For each voxel, we computed a goodness-of-fit measure (r^2 value), corresponding to the amount of variance explained by the deconvolution model (Gardner et al., 2005). The r^2 value represents the degree to which the response of the voxel over time is correlated with the attention task. Thus, when using the r^2 value to select voxels (see below), we essentially selected voxels based on their overall modulation in BOLD response during the task, regardless of any differential activity among conditions.

We also performed a permutation test to assess the statistical significance of the r^2 values to aid our voxel selection. For each subject, we repeated the deconvolution analysis 1000 times, each time with a random reshuffling of the trial labels. For each of the 1000 analyses, we took the maximum r^2 value across all voxels in the brain to obtain a distribution of 1000 maximum r^2 values. This null distribution thus contained the maximum possible r^2 value expected by chance for all voxels and can be used to assess the statistical significance of the observed r^2 values while controlling for familywise type I error (Nichols and Holmes, 2001). The p value of each voxel was calculated as the percentile of voxels from the null distribution that exceeded the observed r^2 value. We also used the r^2 value in conjunction with the anatomical constraints to define the following two frontal areas as clusters of active voxels during the attention task: frontal eye field (FEF) in the vicinity of the precentral sulcus and superior frontal sulcus, and inferior frontal junction (IFJ) at the intersection between inferior frontal sulcus and inferior portion of precentral sulcus.

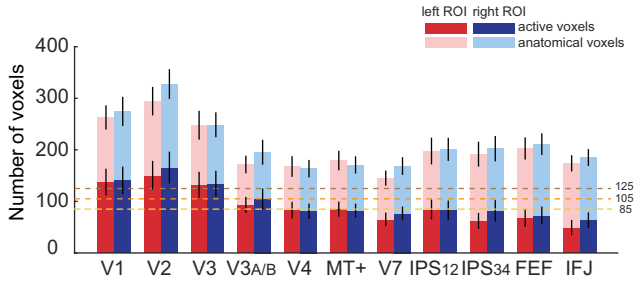

Figure 2. Number of anatomical voxels and task-related active voxels. We assessed the responsiveness of each voxel to the task using a permutation test, which allowed us to derive a p value for the observed r^2 value in each voxel (see Materials and Methods). We labeled voxels with $p < 0.05$ as active, and plotted the average number of active voxels (dark bars) and the total number of anatomical voxels (light bars) for each predefined brain area. Three different levels of voxel inclusion criteria (85, 105, and 125) are indicated by horizontal dashed lines. Black vertical lines denote 95% confidence intervals across subjects.

Voxel selection and response calculation. For each ROI, we first eliminated noisy voxels defined as any voxel that showed >10% signal change. We then sorted the voxels by their r^2 value in a descending order and selected the top 85 voxels for each ROI and subject for further analysis. This number of voxels was found in >95% of all ROIs (1056 in total). For ROIs that had <85 voxels using this criterion, we used all voxels that satisfied the criterion (average, 70 ± 12 voxels). This number of voxels was also approximately the number of active voxels in frontoparietal areas, as assessed by the permutation test described above (Fig. 2). Because of the anatomical difference, the proportion of selected voxels was smaller in visual areas than that in frontoparietal areas, we thus repeated the main analyses using r^2 sorted voxels at 105 and 125 voxels, which included ~90% and ~85%, respectively, of all ROIs that met this criterion.

For each voxel and each ROI, we obtained single-trial fMRI response amplitude, resulting in an $m \times n$ instance matrix for each ROI and attention condition, where m was the number of trials and n was the number of voxels. Taking into account the difference in study design (block vs event-related) across experiments, we obtained the instance matrix by averaging a time window of 10 s (5 time points) for block design in experiments 1 and 2 and 6.6 s (three time points) for event-related design in experiments 3–5. For each ROI and each subject, we adjusted the start time of the averaging window to account for variable response profiles across ROIs and subjects. Specifically, we aggregated data into time bins with variable start times, ranging from the second to fifth time point after the trial onset. The time bin with the largest overall response was then used for extracting the single-trial BOLD response. For multivariate analysis, we averaged the time points from the selected bins to obtain an $m \times n$ instance matrix for each ROI and attention condition. For univariate analysis, we further averaged responses across voxels to obtain $m \times 1$ trialwise vectors for each ROI and each attention condition.

Receiver operating characteristic analysis. For each subject and each ROI, we performed a receiver operating characteristic (ROC) analysis on

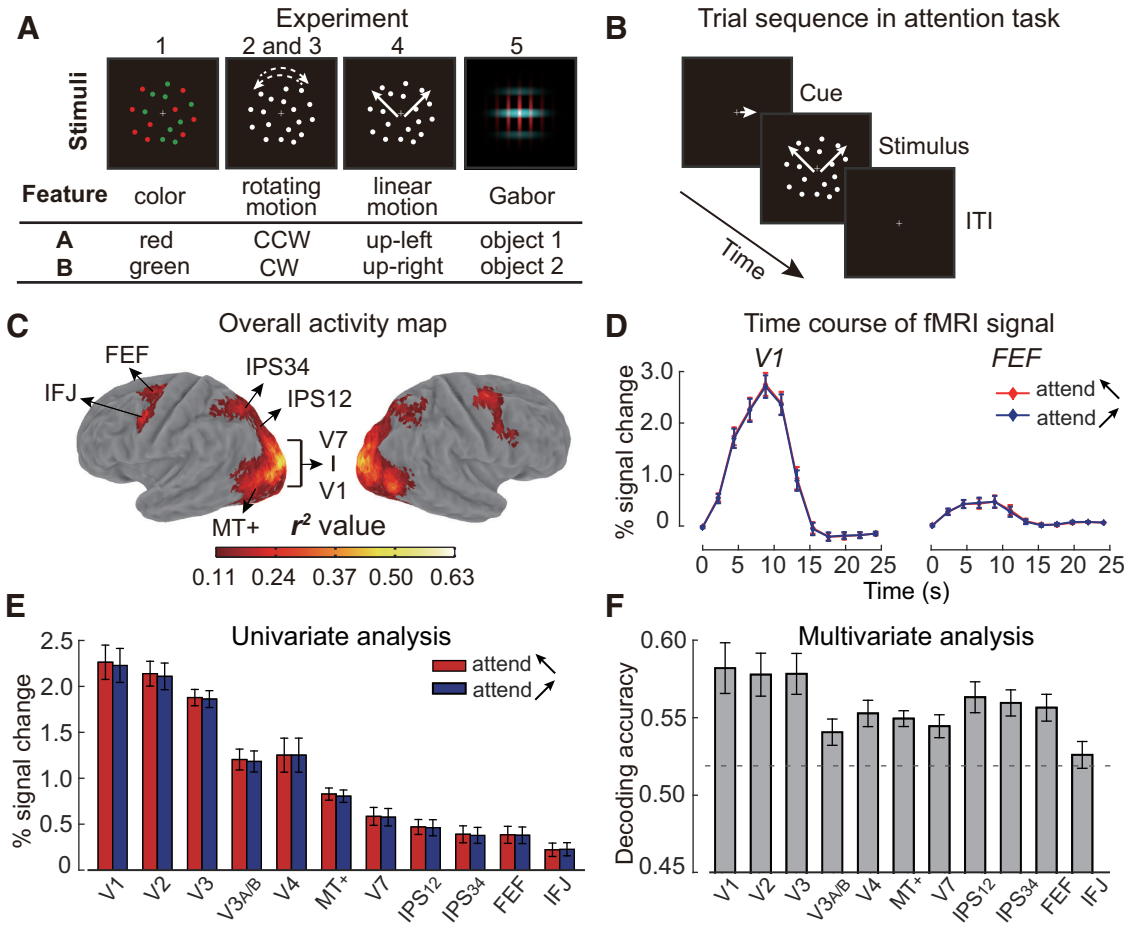

Figure 3. Schematic of stimuli and example results from one experiment. ***A***, Schematic of stimuli across experiments. Data from five experiments were reanalyzed, with a total of 48 subjects. All experiments used a compound stimulus containing two features, labeled as A and B, with their meaning explained in the table. ***B–F***, Example results from experiment 4 using linear motion ($N = 12$). ***B***, Trial sequence of the attention task. ***C***, Overall activity (r^2) map visualized on an inflated atlas surface. The approximate locations of the key brain areas in occipital and frontoparietal cortices are indicated. ***D***, Mean fMRI time course for two attention conditions in V1 and FEF. ***E***, Univariate analysis: mean fMRI responses in visual and frontoparietal areas for each attention condition. ***F***, Multivariate analysis: decoding accuracy in visual and frontoparietal areas. Gray dashed line indicates the maximal significance threshold (corresponding to $p < 0.05$) across brain areas obtained from a permutation test. Error bar denotes the standard error of the mean (SEM) across 12 subjects.

the trialwise response (see Voxel selection and response calculation) to quantify the univariate discriminant information between the two attention conditions. The ROC analysis is essentially a classification analysis based on one-dimensional, univariate response. The area under the ROC curve (AUC) indicates the reliability with which an ideal observer could distinguish the attended feature given the response distributions for both conditions (Britten et al., 1992). In general, we rectified the raw AUC values at ~0.5 (e.g., an AUC of 0.45 is rectified to 0.55) to quantify the discriminant information regardless of the direction of bias. For ease of explanation, we refer to the rectified AUC values simply as AUC, and the original, unrectified AUC as raw AUC. Because the AUC values were always >0.5 (theoretical chance level), we assessed the statistical significance of AUC using a permutation test. We shuffled the labels of trials to obtain an AUC of the shuffled data and repeated this procedure 1000 times to obtain a null distribution of 1000 AUCs for each subject and each ROI. To compute the group-level significance, we concatenated the null distribution across 48 subjects to obtain a 48 × 1000 matrix and then averaged the values over subjects to obtain a group-level null distribution of 1000 AUC values for each brain area. The 95th percentile of this group-level distribution was thus defined as the statistical significance level (corresponding to $p = 0.05$) to determine whether an observed AUC value significantly exceeded the chance level. For the correlation analysis between behavioral difference and AUCs, we used the raw AUCs such that the direction of neural bias (A > B vs A < B) was represented by values >0.5 or <0.5, respectively. We removed outliers based on the median absolute deviation method (Williams, 2011) separately for each brain area, which resulted in an exclusion of <5% data points. Finally, we used the median AUC across ROIs to label each subject's direction of bias (values >0.5 indicated bias for A, and vice versa). This subject level assignment was then used to assess the distribution of bias at the group level.

Multivariate pattern analysis. For each voxel and each ROI, we obtained a single-trial fMRI response amplitude (see Voxel selection and response calculation). We then performed multivariate pattern analysis (MVPA) to discriminate between the two conditions using Fisher linear discriminant analysis. As we often have fewer trials than voxels, which made the estimated covariance matrix noninvertible, we added a ridge coefficient to the diagonal elements of the covariance matrix (Warton, 2008). We performed leave-one-run-out cross-validation to evaluate the classification accuracy by dividing the dataset into test data (one run) and training data (remaining runs). This procedure was repeated until each run was tested once. Classification accuracy was averaged across folds for each ROI. To assess the contribution of biased coding to MVPA, we subtracted the grand mean of each instance matrix from the instance matrix itself (i.e., mean removal separately for each attention condition) before applying the same MVPAs as before. Similar to the permutation test used for ROC analysis, we assessed the significance of decoding accuracy by shuffling the trial labels in the training data and calculated the decoding accuracy on the test data. We repeated this procedure 1000 times to compute a null distribution for each subject and each ROI. We then averaged the null distributions over all subjects to obtain a group-level distribution of decoding accuracies for each ROI.

The 95th percentile of this group-level distribution was thus defined as the statistical significance level (corresponding to $p = 0.05$) to determine whether an observed MVPA decoding accuracy significantly exceeded the chance level. For the correlation analysis between AUCs and decoding accuracies, we used rectified AUCs such that both indices reflected the amount of discriminant information between feature A and feature B. We removed outliers based on the median absolute deviation method (Williams, 2011) separately for each brain area, which resulted in an exclusion of <10% data points (range, 2–8.3%). In all analyses where multiple statistical tests were conducted, we corrected the p values using the false discovery rate (FDR) method (Benjamini and Hochberg, 1995).

Results

Overview of experiments

All experiments used a similar paradigm, where subjects were cued to attend to one of the two superimposed stimuli at the same spatial location. To facilitate the presentation of the results, we arbitrarily named the two stimuli feature A and feature B (Fig. 3*A*, for details). The task of the subjects was to report brief changes (e.g., luminance or moving speed) contingent on the attended feature (see Materials and Methods). This task design kept the physical stimuli constant but varied attentional instruction. Thus, differential neural responses in the two experimental conditions reflect attentional modulation, instead of stimulus-related changes. We show representative results from one of the experiments, where subjects were cued to attend to dots moving in either the upper left or upper right direction (Fig. 3*B*). Figure 3*C* shows the overall brain activation during the task, and Figure 3*D* shows the group-averaged mean time courses of the fMRI BOLD response in two representative regions (V1 and FEF). There was no univariate difference between the two attention conditions in average BOLD responses across subjects (Fig. 3*E*). In contrast, the attended feature can be reliably decoded from distributed activity patterns in both visual and frontoparietal areas (Fig. 3*F*). We obtained similar results from the other experiments; details can be found in previous publications (Liu et al., 2011; Liu, 2016; Jigo et al., 2018; Gong and Liu, 2020). In total, the dataset contained BOLD data from 48 subjects, each containing 22 brain areas (11 areas/hemisphere).

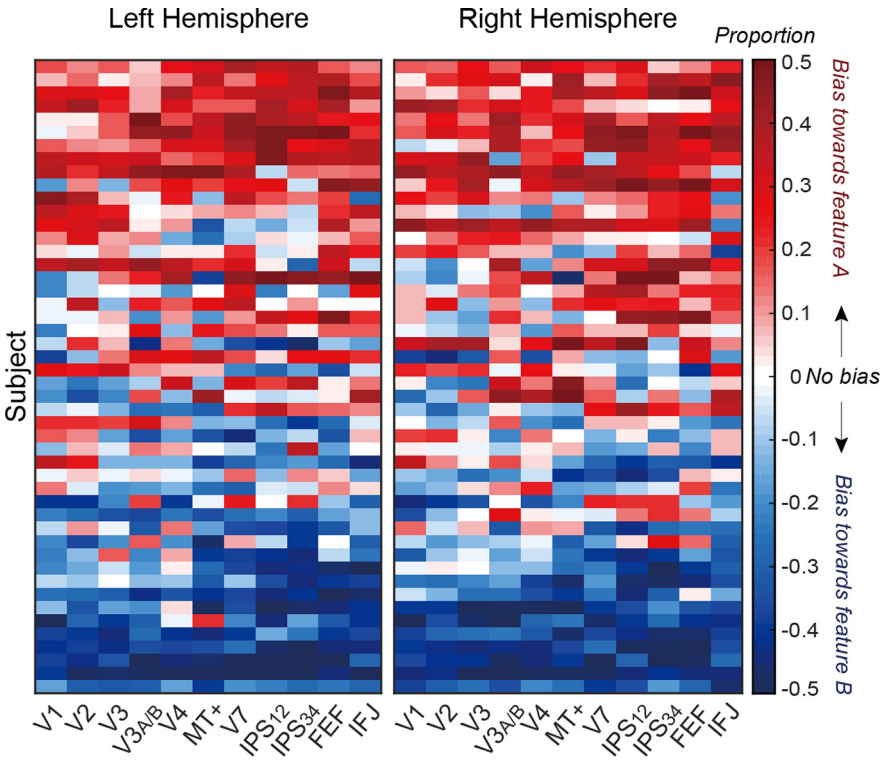

Figure 4. Quantifying the neural bias in individual brain areas and subjects. Data from left and right hemispheres are shown in separate maps. Each row represents data from an individual subject, and each column represents a single brain area. Each cell is color coded by the proportional bias, where positive and negative values indicate bias for feature A and feature B, respectively. We arranged this map by sorting the degree of the dominant bias per subject (indexed by the number of brain areas with the same direction of bias), such that, from top to bottom, bias progressed from feature A to feature B.

Biased neural representation of attention across brain areas and subjects

To explore whether neural activity showed a biased response, we used a method similar to the one used by Fitzgerald et al. (2013). In their work, they rank-ordered the response of each neuron for different categories to index the direction of bias and found that the majority of recorded neurons showed a biased response in the same direction. Similarly, we assessed the bias in fMRI BOLD signal by measuring the proportion of voxels showing stronger response to feature A than to feature B. A proportion value of 0.5 indicates no overall bias; we calculated the proportional bias by subtracting 0.5 from the proportion values such that a bias toward feature A or B would be indicated by positive or negative values, respectively. Figure 4 summarizes the proportion bias for all brain areas and subjects, with the color indicating the direction of the bias. The color map showed many areas exhibiting a biased response pattern across the two attention conditions. Furthermore, within a given subject, multiple brain areas tend to have the same direction of bias. We evaluated the consistency of the direction of bias across brain areas with a binomial test, for which the null hypothesis is random distribution of bias direction across brain areas (i.e., approximately half of the brain areas showing a bias toward feature A and the other half showing a bias toward feature B). We found that 37 of 48 subjects (maximum $p = 0.034$, FDR corrected) exhibited a consistent direction of bias across 22 brain areas (11 per hemisphere). The mean proportion of brain areas that showed the same direction of bias was ~77% across subjects (i.e., ~17 of 22 brain areas showed the same direction of bias). These results suggest that the bias is reliably consistent across regions for the majority of subjects.

Biased representation in higher-order frontoparietal areas

Comparison across individual brain areas and hemispheres

To quantify the amount of bias in individual brain areas, we used a ROC analysis using trial-level responses from two conditions (see Materials and Methods). Note that this analysis was conducted on trialwise responses instead of voxelwise responses, as fMRI voxels within a brain area are likely nonindependent, which could complicate statistical inference. The univariate bias can be indexed by the AUC, which provides a standardized, nonparametric measure of the separation between the two

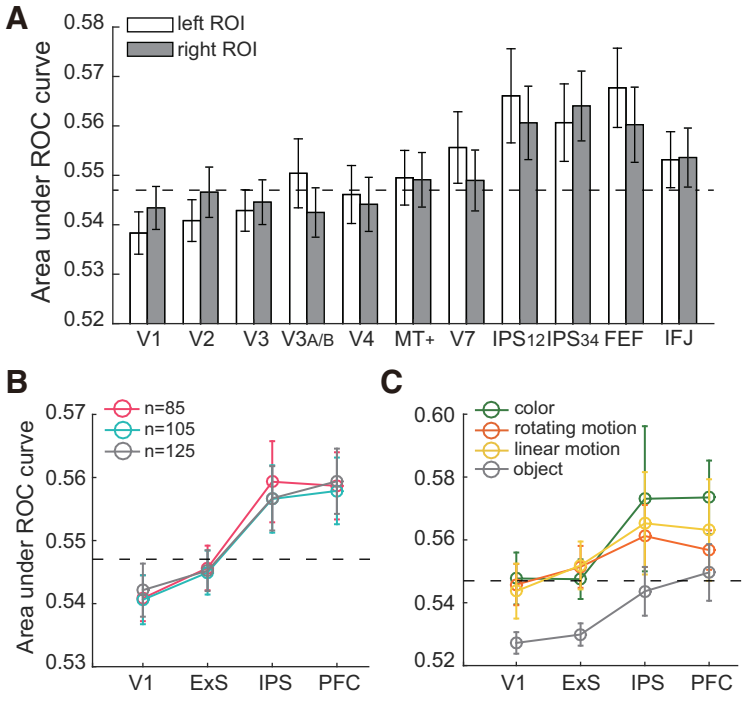

Figure 5. Results from the ROC analysis. ***A***, Average AUC values across all subjects in all areas of interest. ***B***, Average AUC values for different region groups, obtained with different numbers of voxels ($n = 85$, 105, and 125) in the ROC analysis. ***C***, Average AUC values for each stimulus domain, obtained with data from different experiments. The error bar denotes SEM. In all of the plots, the horizontal dashed lines indicate the maximal significance threshold (corresponding to $p < 0.05$) across brain areas obtained from a permutation test.

distributions, such that a value of 0.5 indicates no separation and a value of 0 or 1 indicates perfect separation. To assess the bias regardless of its direction, we further rectified the AUC values around 0.5 for individual brain areas (see Materials and Methods). We then averaged AUCs across subjects and performed permutation tests to determine the significance threshold (see Materials and Methods). We found a general trend of increasing AUC values along the cortical hierarchy, which were statistically reliable in the frontoparietal areas (Fig. 5*A*; maximal threshold, 0.547), but not in most of the visual areas. This regional difference was confirmed by a two-way repeated-measures ANOVA (11 brain areas × 2 hemispheres), showing a significant effect of brain areas ($F_{(10,470)} = 4.53$, $p < 0.001$, $\eta^2 = 0.088$) without a significant difference between hemispheres ($p = 0.593$) or interaction between hemisphere and brain areas ($p = 0.596$).

To further characterize the pattern of biased neural response, we grouped the 11 areas into the following four region groups based on anatomical considerations: V1, extrastriate visual areas (ExS), consisting of V2, V3, V3A/B, V4, and MT+; areas in the intraparietal sulcus (IPS), consisting of V7/IPS0, IPS12, and IPS34; and prefrontal cortex (PFC), consisting of FEF and IFJ. Within a region group, we averaged AUCs across constituting brain areas. Because there was no laterality effect, we further collapsed data across two hemispheres. Then, we conducted separate analyses on AUCs to assess whether the biased representation varied with region group and stimulus domain.

Comparison across region groups and stimulus domains

When grouped into four main anatomical region groups, we observed a clear increase in AUC from V1 to extrastriate visual areas, and further into parietal and frontal areas (Fig. 5*B*, red plot). A one-way repeated-measures ANOVA on AUCs revealed a significant effect of region group ($F_{(3,141)} = 7.18$, $p < 0.001$, $\eta^2 = 0.133$). Pairwise comparisons further showed a stronger bias in frontoparietal areas than that in V1 and ExS (p values < 0.01), without a significant difference between IPS and PFC ($p = 0.865$). To confirm that this result was not because of our specific voxel selection criterion, we repeated the same analysis using different numbers of voxels ($n = 105$ and 125) and found similar results (Fig. 5*B*, cyan and black plots). A two-way repeated-measures ANOVA (4 region groups × 3 voxel numbers) revealed a significant main effect of region group ($F_{(3,282)} = 6.51$, $p < 0.001$, $\eta^2 = 0.122$) without a main effect of number of voxels ($F_{(2,282)} = 0.59$, $p = 0.555$) or the interaction effect ($F_{(6,282)} = 0.89$, $p = 0.501$). These results demonstrate that the attentional modulation lacks bias in visual areas but shows a significant and stronger bias in higher-order areas.

Because we used different stimuli across experiments, it is possible that the observed bias was primarily driven by a particular stimulus. We thus separated AUCs according to the stimulus domain and performed a mixed-effect ANOVA (4 region groups × 4 stimuli), which showed only a main effect of region group ($F_{(3,132)} = 5.23$, $p = 0.002$, $\eta^2 = 0.108$). There was no main effect of stimuli ($F_{(3,44)} = 1.124$, $p = 0.358$) or interaction ($F_{(9,132)} = 0.732$, $p = 0.718$). These results indicated that the biased representation of attentional signal was not significantly modulated by stimuli. We do note, however, that, numerically, AUCs were the largest for colors, intermediate for motion directions, and smallest for the dynamic objects (Fig. 5*C*). When comparing with the significance threshold obtained from the permutation tests, AUCs for dynamic objects even dropped to chance level in IPS. This pattern of results hints at a possible decrease of bias along with increasing complexity of the attended information (e.g., from simple features to complex objects).

Bias removal produces dissociable effects in sensory and frontoparietal areas

The AUC analysis demonstrates a univariate bias between two attention conditions in many brain areas. Previously, we have shown significant above-chance multivariate decoding using pattern classification techniques in all those areas (Fig. 3*F*). Given that both methods index the neural discriminability between conditions, this raises the question of how much the univariate bias contributes to the multivariate decoding. We used the grand mean of the BOLD signal from each attention condition (across voxels and trials) as a proxy measure of this bias. We then subtracted this grand mean from each attention condition and performed both the ROC analysis and MVPA decoding separately for each brain area (see Materials and Methods). As expected, AUCs in all region groups fell below the significance threshold (i.e., not different from chance) because mean removal essentially eliminated univariate difference (Fig. 6*A*). If multivariate decoding relies mostly on univariate differences, we expect that removing the mean would diminish the decoding accuracy. Alternatively, if multivariate decoding relies on multidimensional pattern variability, we expect little impact of mean

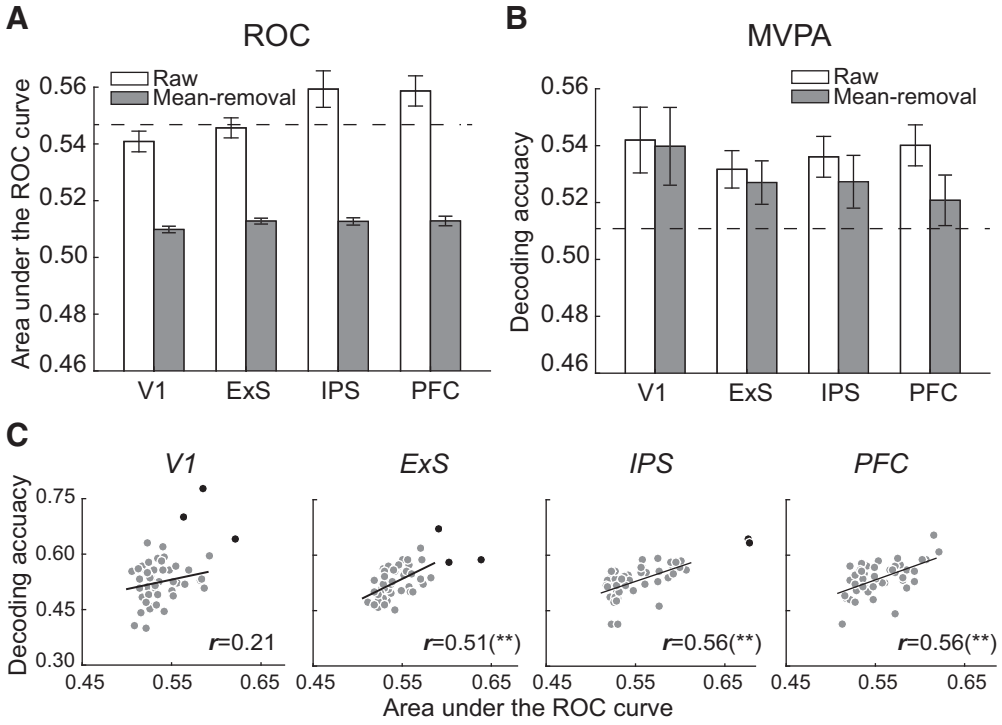

Figure 6. The relationship between ROC and MVPA. ***A***, ROC results using the original data (Raw) and after removing the grand mean from each attention condition (Mean removal). ***B***, MVPA results before and after removing the grand mean from each attention condition. In both plots, horizontal dashed lines indicate maximal significance thresholds obtained from permutation tests. The error bar denotes the SEM. ***C***, Intersubject correlation between AUC and MVPA decoding accuracy for each region group (**statistical significance at $p < 0.01$). Black dots indicate outliers detected by the median absolute deviation method (Williams, 2011).

removal on decoding accuracy. We found that mean removal had a progressively stronger impact on decoding accuracy along the cortical hierarchy (Fig. 6*B*). This was confirmed by a two-way repeated-measures ANOVA (region group × mean removal), showing a main effect of mean removal ($F_{(1,141)} = 11.46$, $p = 0.001$, $\eta^2 = 0.196$), and, importantly, a significant interaction between region group and mean removal ($F_{(3,141)} = 4.25$, $p = 0.006$, $\eta^2 = 0.083$). Follow-up t tests showed that mean removal produced a significant drop of MVPA-based decoding in ExS ($p = 0.024$), IPS ($p = 0.043$), and PFC ($p < 0.001$), but not in V1 ($p = 0.679$). A significant interaction was also obtained if we excluded V1 data from the analysis ($F_{(2,94)} = 7.28$, $p = 0.001$, $\eta^2 = 0.134$), indicating a relatively larger drop in decoding accuracy because of mean removal in PFC compared with IPS and ExS. These results suggest that the progressively stronger univariate bias along the cortical hierarchy, as shown by the ROC analysis, contributes significantly to MVPA-based decoding. Indeed, in the PFC, MVPA decoding appears to rely mostly on the univariate bias. We obtained a similar pattern of results when using 105 or 125 voxels.

To further examine the contribution of univariate bias to MVPA-based decoding, we calculated Pearson correlations between AUCs and MVPA decoding accuracies across subjects for each brain area and averaged them into four region groups (Fig. 6*C*). For regions where the univariate bias contributes to MVPA-based decoding, we should expect a dependence between these two measures. Indeed, we found a progressively stronger relationship from sensory to frontoparietal areas. The AUC and MVPA decoding were positively correlated in frontoparietal areas (IPS, $r = 0.562$; PFC, $r = 0.562$; p values < 0.001, FDR corrected), and such correlation was also observed in extrastriate areas ($r = 0.512$, p values < 0.001, FDR corrected), but not for V1 ($r = 0.21$, p values $= 0.170$, FDR corrected). These results further support the possibility that attention decoding was driven in part by the univariate bias.

Biased coding does not correlate with behavioral selection

We examined whether the observed neural bias was a consequence of preferential behavioral selection. For example, a neural bias in favor of a particular feature may result from a stronger top-down attention to that feature, contributing to higher accuracy and faster reaction time, although subjects were always instructed to attend equally to individual features. Such a preferential selection should lead to better behavioral performance in the attention tasks. We thus tested this possibility by correlating the difference in behavioral performance (both accuracy and reaction time) between the two attention conditions (i.e., feature A – feature B) and AUCs. For this analysis, we used raw AUCs without rectification, which captured the direction of the neural bias (see Materials and Methods). This analysis revealed no significant correlations in any of the region groups between the magnitude of the neural bias and the behavioral preference (Fig. 7; accuracy, p values > 0.75; reaction time, p values > 0.90; FDR corrected), making it unlikely that the observed neural bias is because of behavioral preferences.

Equivalent distribution of biased coding at the group level

Last, we examined the group-level distribution of the direction (or sign) of the biased representation by assigning each subject a preferred feature (see Materials and Methods). We grouped individuals according to the stimuli they viewed during the task (i.e., rotating motion, linear motion, and dynamic object). We excluded data for the attend-color experiment because of the small sample size in that experiment ($N = 6$). Figure 8 shows approximately equal distribution of biased direction across

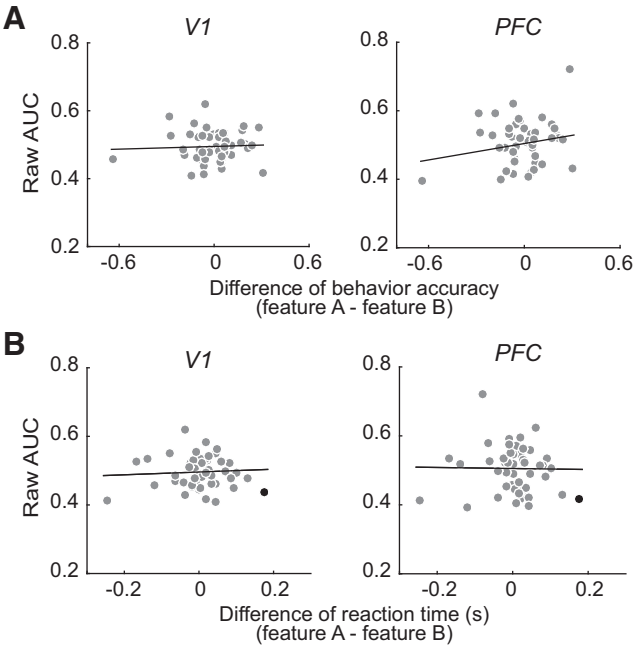

Figure 7. The relationship between raw AUC and behavioral difference. ***A***, ***B***, Correlations between raw AUC values and behavioral difference, as indexed by accuracy (***A***) and reaction time (***B***), between two attention conditions in two representative areas (V1 and PFC). Black dots indicate outliers detected by the median absolute deviation method (Williams, 2011).

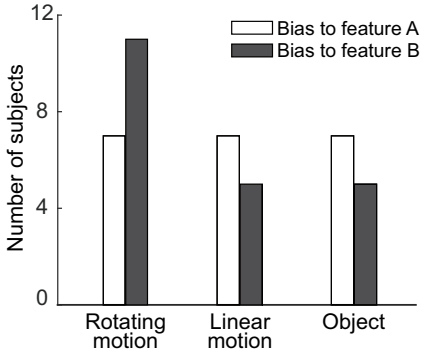

Figure 8. Group-level distribution of biased representation across subjects. The distribution of biased representation across subjects was shown separately for different stimuli (i.e., rotating motion, linear motion, and dynamic object).

subjects (χ^2 test against equal proportion, p values > 0.18). Thus, about half of the subjects showed a biased response to feature A, and the other half showed the opposite bias. This observation explains the lack of systematic univariate bias at the group level because averaging the two opposite biases likely canceled each other out.

Discussion

We observed a biased neural representation of the attended feature in the frontoparietal network, using a large fMRI dataset containing 48 subjects across multiple experiments. At the level of brain areas, we found that many brain areas showed differential univariate activity between attention conditions. Within individual subjects, the direction of bias remained consistent among brain areas. However, at the group level, the direction of bias varied across subjects with no predominant direction, thus explaining the lack of a group average univariate difference in standard analyses. When we quantified the amount of univariate bias, we found a progressively stronger biased response from sensory to frontoparietal areas, with reliable above-chance bias in the latter areas. Importantly, biased responses had a major contribution to multivariate decoding in frontoparietal areas. We ruled out the possibility that the results were because of a specific voxel selection criterion, stimulus domain, or behavioral preference. Collectively, our findings provide novel evidence for a biased neural representation of cognitive variables in the frontoparietal network of the human brain.

The initial observation of biased representation (Fitzgerald et al., 2013) in nonhuman primates was indeed rather surprising (Chafee, 2013). However, it is unknown whether such bias is present only in nonhuman primates, who typically undergo extensive training in behavioral tasks before neural recording experiments, in the specific cognitive tasks investigated, or in the specific brain area examined, namely LIP. Although our attention task and the tasks performed by the monkeys are different, they shared some formal similarities. First, all tasks were designed to isolate neural signals for cognitive, associative representations without contributions from stimulus- and motor-related processes. Second, these cognitive representations establish an association between visual input and a categorical structure as specified by the task. Our results thus extend biased neural representation to a new class of tasks and to humans with much less training. Here we further leveraged the large sample size and whole-brain coverage afforded by fMRI and show that bias occurs in a multitude of brain areas with consistent direction among brain areas within an individual subject, although the direction of bias varies across subjects.

Notably, with the whole-brain coverage of fMRI data, we found a dissociable pattern of bias between sensory areas and frontoparietal areas. First, the magnitude of bias grows progressively larger from early visual areas to frontoparietal areas, with an absence of statistically reliable bias in most of the visual areas. Second, the MVPA on mean removed data suggests that the univariate bias made progressively larger contribution to pattern-based decoding along the cortical hierarchy (Fig. 6*B*). The contribution of the univariate bias to pattern-based decoding was further supported by the finding of a significant correlation between AUC values and decoding accuracies, which were highest in frontoparietal areas. These contrasting results thus support a distinction between the following two mechanisms: distributed population representation in sensory areas and biased representation in high-level areas.

From a computational point of view, our results likely reflect different dimensionality of the neural signals at different levels of cortical hierarchy. Specifically, the differential impacts of removing the grand mean on multivariate decoding suggests that sensory areas contain high-dimensional neural signals whereas frontoparietal areas contain low-dimensional neural signals. Such a coding scheme is also consistent with general theories of visual information processing, which typically assume that sensory input is processed along hierarchical stages that start with analog representations and gradually transition to task-related, abstract representations (Riesenhuber and Poggio, 1999; Hochstein and Ahissar, 2002; Deco and Rolls, 2004). This transition likely involves changes in the coding properties in different brain areas, and our observation of different amount of bias (and signal dimensionality) among cortical areas could be one manifestation of this transition.

Analog representations are naturally implemented with sensory neurons that are continuously tuned to stimulus features (Hubel and Wiesel, 1962; Maunsell, and Van Essen, 1983; Blasdel, 1992). Consistent with this idea, we observed a distributed population response in sensory areas without appreciable univariate bias. Our results are also consistent with a large body of fMRI studies that have reliably decoded the attended stimulus features (Kamitani and Tong, 2005, 2006; Liu et al., 2011; Liu and Hou, 2013) and memorized stimulus features (Harrison and Tong, 2009; Serences et al., 2009; Riggall and Postle, 2012) from activity patterns in early visual areas.

While analog representation via distributed population response in early sensory areas is well established, how information transitions to an abstract representation in high-level brain areas is much less known. In our data, we found evidence for a biased, or low-dimensional, representation for attended feature in frontal and parietal areas, which are the likely sources of attentional control (Kastner and Ungerleider, 2000; Bisley and Goldberg, 2003) and also part of the multiple-demand network (Duncan, 2013; Fedorenko et al., 2013; Woolgar et al., 2016). It may seem counterintuitive that with the myriad of neurons and their complex connections, the brain uses a low-dimensional, or possibly scalar, code to represent an abstract cognitive variable. Computational analyses, however, have pointed out some benefits of using such a simple code. Because individual neurons are always noisy, a biased response pattern would allow a simple operation, such as averaging, to achieve a reliable representation that is robust to noisy fluctuations in neural systems (Fitzgerald et al., 2013). Furthermore, such a simple neural code can also simplify the readout of downstream areas and the control of behavior. A related idea was proposed in a modeling study (Ganguli et al., 2008), in which LIP neuronal data from categorization and decision-making tasks were found to obey one-dimensional dynamics, such that slowly evolving activity patterns are proportional to spontaneous activity. The investigators suggested that by reducing local neural signals to one-dimensional activity, the brain achieves robust temporal control of behavior such as the timing in shifting attention and crossing a decision threshold during evidence accumulation. Although these proposed benefits of low-dimensional representations reflect different aspects of information coding, they are similar in that unreliable and heterogeneous neural activities from individual neurons can be pooled to achieve more robust representation of cognitive variables.

A natural question concerns how low-dimensional neural activity is generated in the brain. While simulations with simple network models show that local, sparse, recurrent excitatory connections can generate low-dimensional neural activity, it is also possible that coupling among cortical areas plays a role (Ganguli et al., 2008). A limitation of previous single-unit work is that all of the data come from a single brain area, namely LIP. Thus, it is unknown whether low-dimensional neural activity is restricted to one, or a few, brain areas, or is instead a network phenomenon. Our data showed a biased response pattern in the widespread frontoparietal network, and, critically, the direction of such bias was consistent across nodes in this network. Our results thus suggest that network-level interaction could contribute to the generation and maintenance of low-dimensional neural activity.

It is worthwhile to consider the generality of biased neural representation. Given that our human fMRI data and previous monkey single-unit data were obtained from a variety of behavioral paradigms, biased representation of cognitive variables appears to be a general phenomenon. However, we should note that a commonality shared among these behavioral paradigms is that all tasks entail a few (often two) discrete task conditions. It is possible that biased representation is particularly useful in this type of regime but would be less useful with more complex task contexts (e.g., an attention task with increased number of features). Theoretical studies suggest that neural dimensionality could scale with task complexity (Gao and Ganguli, 2015; Fusi et al., 2016). This idea has found support in studies where the number of task conditions appears to drive estimates of dimensionality in monkey PFC (Rigotti et al., 2013; Brincat, et al., 2018). There is also some hint in our data supporting this notion, as the univariate bias was numerically weaker for dynamic multifeature objects than single features (Fig. 5*C*). Future studies are necessary to systematically evaluate the influence of task and stimulus complexity on the dimensionality of neural signals.

We also do not know how biased representation arises in the first place. Each subject in our dataset only performed a task once in a single scanning session, which does not allow us to test the stability of the bias. It is possible that the direction of such bias is determined by each individual's past experience, which, hence, is more or less fixed for that individual, or, alternatively, that such bias arises stochastically when performing a particular task. It would be interesting to examine the consistency and origin of the neural bias in future studies.

In conclusion, we highlight biased neural representation as a potential mechanism for coding cognitive variables in higher-order frontoparietal areas. Although the simplicity of this coding scheme seems counterintuitive, it may facilitate an abstract representation and simple readout of information critical for stimulus selection and cognitive control. Together with the findings of distributed population representation in sensory areas, our results suggest a gradual transition from high- to low-dimensional representation along the cortical hierarchy. Such a gradient of neural representation could enable information processing at multiple levels of abstraction to support adaptive behavior.

References

Bettencourt KC, Xu Y (2016) Decoding the content of visual short-term memory under distraction in occipital and parietal areas. Nat Neurosci 19:150–157.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57:289–300.

Bisley JW, Goldberg ME (2003) Neuronal activity in the lateral intraparietal area and spatial attention. Science 299:81–86.

Blasdel GG (1992) Orientation selectivity, preference, and continuity in monkey striate cortex. J Neurosci 12:3139–3161.

Britten KH, Shadlen MN, Newsome WT, Movshon JA (1992) The analysis of visual motion: a comparison of neuronal and psychophysical performance. J Neurosci 12:4745–4765.

Brincat SL, Siegel M, von Nicolai C, Miller EK (2018) Gradual progression from sensory to task-related processing in cerebral cortex. Proc Natl Acad Sci U S A 115:E7202–E7211.

Chafee MV (2013) A scalar neural code for categories in parietal cortex: representing cognitive variables as "more" or "less. Neuron 77:7–9.

Corbetta M, Shulman GL (2002) Control of goal-directed and stimulus-driven attention in the brain. Nat Rev Neurosci 3:201–215.

Deco G, Rolls ET (2004) A neurodynamical cortical model of visual attention and invariant object recognition. Vision Res 44:621–642.

DeYoe E, Carman G, Bandettini P, Glickman S, Wieser J, Cox R, Miller D, Neitz J (1996) Mapping striate and extrastriate visual areas in human cerebral cortex. Proc Natl Acad Sci U S A 93:2382–2386.

Duncan J (2001) An adaptive coding model of neural function in prefrontal cortex. Nat Rev Neurosci 2:820–829.

Duncan J (2010) The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends Cogn Sci 14:172–179.
Duncan J (2013) The structure of cognition: attentional episodes in mind and brain. Neuron 80:35–50.
Engel SA, Glover GH, Wandell BA (1997) Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cereb Cortex 7:181–192.
Erez Y, Duncan J (2015) Discrimination of visual categories based on behavioral relevance in widespread regions of frontoparietal cortex. J Neurosci 35:12383–12393.
Fedorenko E, Duncan J, Kanwisher N (2013) Broad domain generality in focal regions of frontal and parietal cortex. Proc Natl Acad Sci U S A 110:16616–16621.
Fitzgerald JK, Freedman DJ, Fanini A, Bennur S, Gold JI, Assad JA (2013) Biased associative representations in parietal cortex. Neuron 77:180–191.
Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2001) Categorical representation of visual stimuli in the primate prefrontal cortex. Science 291:312–316.
Fusi S, Miller EK, Rigotti M (2016) Why neurons mix: high dimensionality for higher cognition. Curr Opin Neurobiol 37:66–74.
Ganguli S, Bisley JW, Roitman JD, Shadlen MN, Goldberg ME, Miller KD (2008) One-dimensional dynamics of attention and decision making in LIP. Neuron 58:15–25.
Gao P, Ganguli S (2015) On simplicity and complexity in the brave new world of large-scale neuroscience. Curr Opin Neurobiol 32:148–155.
Gardner JL, Sun P, Waggoner RA, Ueno K, Tanaka K, Cheng K (2005) Contrast adaptation and representation in human early visual cortex. Neuron 47:607–620.
Gong M, Liu T (2020) Continuous and discrete representations of feature-based attentional priority in human frontoparietal network. Cogn Neurosci 11:47–59.
Harrison SA, Tong F (2009) Decoding reveals the contents of visual working memory in early visual areas. Nature 458:632–635.
Hochstein S, Ahissar M (2002) View from the top: hierarchies and reverse hierarchies in the visual system. Neuron 36:791–804.
Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J Physiol 160:106–154.
Jackson JB, Woolgar A (2018) Adaptive coding in the human brain: distinct object features are encoded by overlapping voxels in frontoparietal cortex. Cortex 108:25–34.
Jigo M, Gong M, Liu T (2018) Neural determinants of task performance during feature-based attention in human cortex. eNeuro 5:ENEURO.0375-17.2018.
Kamitani Y, Tong F (2005) Decoding the visual and subjective contents of the human brain. Nat Neurosci 8:679–685.
Kamitani Y, Tong F (2006) Decoding seen and attended motion directions from activity in the human visual cortex. Curr Biol 16:1096–1102.
Kastner S, Ungerleider LG (2000) Mechanisms of visual attention in the human cortex. Annu Rev Neurosci 23:315–341.
Kay KN, Naselaris T, Prenger RJ, Gallant JL (2008) Identifying natural images from human brain activity. Nature 452:352–355.
Konen CS, Kastner S (2008) Representation of eye movements and stimulus motion in topographically organized areas of human posterior parietal cortex. J Neurosci 28:8361–8375.
Li S, Ostwald D, Giese M, Kourtzi Z (2007) Flexible coding for categorical decisions in the human brain. J Neurosci 27:12321–12330.
Liu T (2016) Neural representation of object-specific attentional priority. Neuroimage 129:15–24.
Liu T, Hou Y (2013) A hierarchy of attentional priority signals in human frontoparietal cortex. J Neurosci 33:16606–16616.
Liu T, Hospadaruk L, Zhu DC, Gardner JL (2011) Feature-specific attentional priority signals in human cortex. J Neurosci 31:4484–4495.
Maunsell JH, Van Essen DC (1983) Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. J Neurophysiol 49:1127–1147.
Nichols T, Holmes A (2001) Nonparametric permutation tests for functional neuroimaging. Hum Brain Mapp 15:1–25.
Pouget A, Dayan P, Zemel R (2000) Information processing with population codes. Nat Rev Neurosci 1:125–132.
Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. Nat Neurosci 2:1019–1025.
Riggall AC, Postle BR (2012) The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. J Neurosci 32:12990–12998.
Rigotti M, Barak O, Warden MR, Wang XJ, Daw ND, Miller EK, Fusi S (2013) The importance of mixed selectivity in complex cognitive tasks. Nature 497:585–590.
Schluppeck D, Curtis CE, Glimcher PW, Heeger DJ (2006) Sustained activity in topographic areas of human posterior parietal cortex during memory-guided saccades. J. Neurosci 26:5098–5108.
Serences JT, Ester EF, Vogel EK, Awh E (2009) Stimulus-specific delay activity in human primary visual cortex. Psychol Sci 20:207–214.
Sereno MI, Dale AM, Reppas JB, Kwong KK, Belliveau JW, Brady TJ, Rosen BR, Tootell RB (1995) Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. Science 268:889–893.
Sereno MI, Pitzalis S, Martinez A (2001) Mapping of contralateral space in retinotopic coordinates by a parietal cortical area in humans. Science 294:1350–1354.
Stokes MG, Kusunoki M, Sigala N, Nili H, Gaffan D, Duncan J (2013) Dynamic coding for cognitive control in prefrontal cortex. Neuron 78:364–375.
Swaminathan SK, Freedman DJ (2012) Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. Nat Neurosci 15:315–320.
Warton DI (2008) Penalized normal likelihood and ridge regularization of correlation and covariance matrices. J Am Stat Assoc 103:340–349.
Williams DC (2011) Finite sample correction factors for several simple robust estimators of normal standard deviation. J Stat Comput Simul 81:1697–1702.
Woolgar A, Jackson J, Duncan J (2016) Coding of visual, auditory, rule and response information in the brain: 10 years of multi-voxel pattern analysis. J Cogn Neurosci 28:1433–1454.