Biased neural coding of feature-based attention in human brain

Mengyuan Gong and Taosheng Liu*

Department of Psychology, Michigan State University, East Lansing, MI, 48824, USA

Summary

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. Here, we show a biased neural coding of feature-based attention in a large fMRI dataset. We found that when participants 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 robust at the level of single brain areas and 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 gradient of coding mechanism implements the transition from a more analog to a more abstract representation of attentional priority. Biased neural responses in high-level areas likely reflect a low-dimensional neural code (possibly one-dimensional scalar code) that facilitates robust representation and simple read-out of cognitive variables.

Keywords

Biased neural coding, feature-based attention, fMRI, frontoparietal network

Introduction

The human brain has the capacity to represent an enormous variety of information. How neural activities represent sensory and cognitive information remains a fundamental question in system and cognitive neuroscience. Answers to this question will inform us how neural computations support adaptive behavior, thus providing important clues regarding the computational principles of the brain.

The prevailing view of neural coding is that the brain uses distributed neural activities to represent information. This view is best illustrated by work that examined neural activities in sensory areas of the brain. Classical neurophysiological studies showed that neurons in early visual areas have smooth tuning functions that span a range of feature values (e.g., Hubel and Wiesel, 1962; Blasdel, 1992; Maunsell, and Van Essen, 1983). Hence, a single stimulus feature would evoke different responses across a population of such neurons, resulting in a specific profile of population activity. Computational studies have demonstrated that such population responses can be used for encoding and decoding sensory information (Pouget, 2000). In parallel to these neuronal level findings, human functional magnetic resonance imaging (fMRI) studies have shown that patterns of BOLD responses can be used to decode and reconstruct visual stimulus (e.g., Kamitani and Tong, 2005; Brouwer and Heeger, 2009; Kay et al., 2008).

Although there is a general consensus that stimulus properties are represented via distributed population codes in sensory areas, much less is known about how cognitive variables are represented in the brain. Cognitive functions related to task control and target selection have been widely associated with activity in parietal and

prefrontal cortical areas, collectively known as the Multiple-Demand (MD) network (Duncan, 2010). A key feature of neurons in this network, in contrast to sensory neurons, is that they can flexibly adapt their tuning profile to accommodate task demands (Duncan, 2001; Fedorenko et al., 2013). For instance, neurophysiological studies suggest that population-level neural patterns in prefrontal cortex (PFC) reflect the coding of task rule (Stokes et al., 2013; Freedman, et al., 2001), and the pattern of activity in lateral intraparietal areas (LIP) can encode attentional priority (Bisley and Goldberg, 2003) and learned category (Swaminathan and Freedman, 2012). Contrary to this population-based view, however, a recent neurophysiological study reported evidence supporting a scalar neural coding for cognitive variable in LIP (Fitzgerald et al., 2013). These researchers performed detailed analysis of LIP neuronal activity from several experiments in which monkeys performed a variety of cognitive tasks (categorization, associative learning, perceptual decision making). Surprisingly, the majority of recorded neurons from a monkey showed similar response profile, exhibiting largely biased response to one task condition (e.g., higher response to learned category A than category B). This biased population coding implies a low-dimensional, rather than high-dimensional, neural coding scheme for cognitive variables.

The finding of a biased response is surprising (Chaffee, 2013) and thus it is important to know whether the biased coding scheme is restricted to non-human primates or also applies to humans. Many human fMRI studies have decoded cognitive variables from the measured multivariate BOLD response patterns in the MD network, without observing any obvious biases to a particular condition in the univariate response (e.g., Li et al., 2007; Liu et al., 2011; Liu and Hou, 2013; Erez and Duncan, 2015;

Bettencourt and Xu, 2016). Although such findings are consistent with a distributed population coding mechanism, the univariate results were obtained by averaging data across subjects, which could obscure a biased coding if the direction of bias varied across subjects. Therefore, investigating the existence of biased neural codes requires us to examine neural responses at the single-subject level. Here, we conducted such analyses on a large fMRI dataset from five attention experiments to examine whether biased neural codes are present in the human brain. Selective attention is a key cognitive function and it is thought to be a core operation that enables complex task control (Duncan 2013). Attentional control is also highly associated with activity in the MD network (Corbetta and Shulman, 2002; Scolari et al., 2015). Thus, attentional signals in the brain provide a good test case for possible biased representation of a cognitive variable.

In these experiments, subjects were instructed to attend to one of the two alternative features in a compound stimulus that contained both features, which allowed us to measure the attentional signals throughout the human cortex. We examined potential biased coding in attentional signals at multiple levels of analysis: single brain area, multiple brain areas within a subject, and multiple subjects at the group level. Our analysis revealed a gradient of increasing bias from sensory to high-level areas, suggesting that low-dimensional neural codes are indeed utilized in the human brain to represent cognitive variables such as attentional priority.

Results

Overview of experiments

All experiments utilized a similar paradigm, where subjects were cued to attend to one of the two superimposed stimuli at the same spatial location (Fig. 1A & B). To facilitate the presentation of the results, we arbitrarily named the two stimuli as feature A and B (see Fig. 1A for details). Subjects' task was to report brief changes (e.g. luminance or moving speed) contingent on the attended feature (see *Method Details*). 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. 1B). Figure 1C shows the overall brain activation during the task and Figure 1D shows the group-averaged mean time courses of 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. 1E). In contrast, the attended feature can be reliably decoded from distributed activity patterns in both visual and frontoparietal areas (Fig. 1F). 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, 2019). In total, the dataset contained BOLD responses from 48 subjects, each containing 22 brain areas (11 areas per hemisphere).

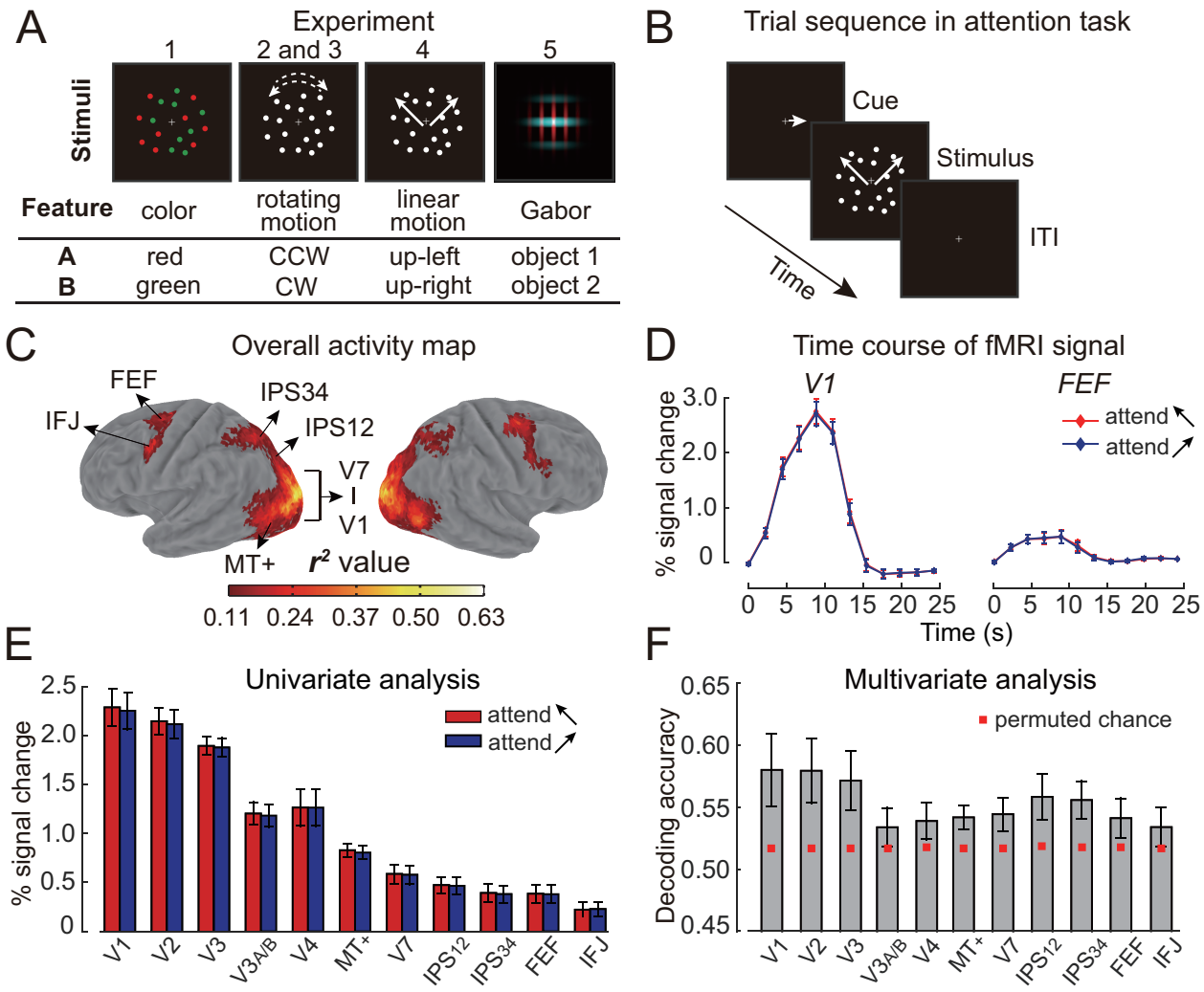

Figure 1. (A) Schematic of stimuli across experiments. Data from five experiments were re-analyzed, with a total of 48 subjects. All experiments employed a compound stimulus containing two features, labeled as 'A' and 'B', with their meaning explained in the table. (B-F) Example results from Experiment 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 cortex are indicated. IPS: intraparietal sulcus, FEF: frontal eye field, IFJ: inferior frontal junction. (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. Error bar denotes standard error of the mean across 12 subjects.

Biased neural coding of attention across brain areas and subjects

We first examined whether neural activity for the two attention conditions showed mean difference *within each brain area across voxels*, using paired t-tests (see *Method*

Details). Figure 2 summarizes all the t-values for all brain areas and subjects, with the color indicating the direction of the bias and the shade indicating its strength. The t-tests were significant in 816 out of 1056 brain areas (77%) after false discovery rate (FDR) correction for multiple comparisons (corrected $p < 0.05$, Benjamini and Hochberg, 1995). This is true for both left hemisphere (N=400/528) and right hemisphere (N=416/528), with no obvious difference in the proportion of significant areas between hemispheres (chi-square test for equal proportion: $\chi^2=1.38$, $p=0.24$). Thus, in the majority of the brain areas, voxel responses were on average higher for one attended feature than the other, exhibiting a univariate difference, i.e., biased coding.

Although the direction of the bias varied across subjects, it appeared consistent *within an individual subject across brain areas*. To test the degree of such within-subject consistency, we categorized the direction of bias according to the sign of the t-statistic in each area for each subject (i.e., positive t-values indicated bias towards feature A and negative t-values indicated biased toward feature B). If the direction of such bias is random across brain areas, we should expect approximately half of the brain areas showing a positive bias and the other half showing a negative bias, across the 22 brain areas (11 in each hemisphere). We used a binomial test to evaluate the deviation of observed bias pattern against this null hypothesis and found most of the subjects significantly deviated from the null (N=41/48, maximum $p=0.0307$, *FDR*-corrected). Across subjects, the mean proportion of brain areas that showed the same sign of bias was 81% (i.e., about 18 out of 22 brain areas showed the same direction of bias).

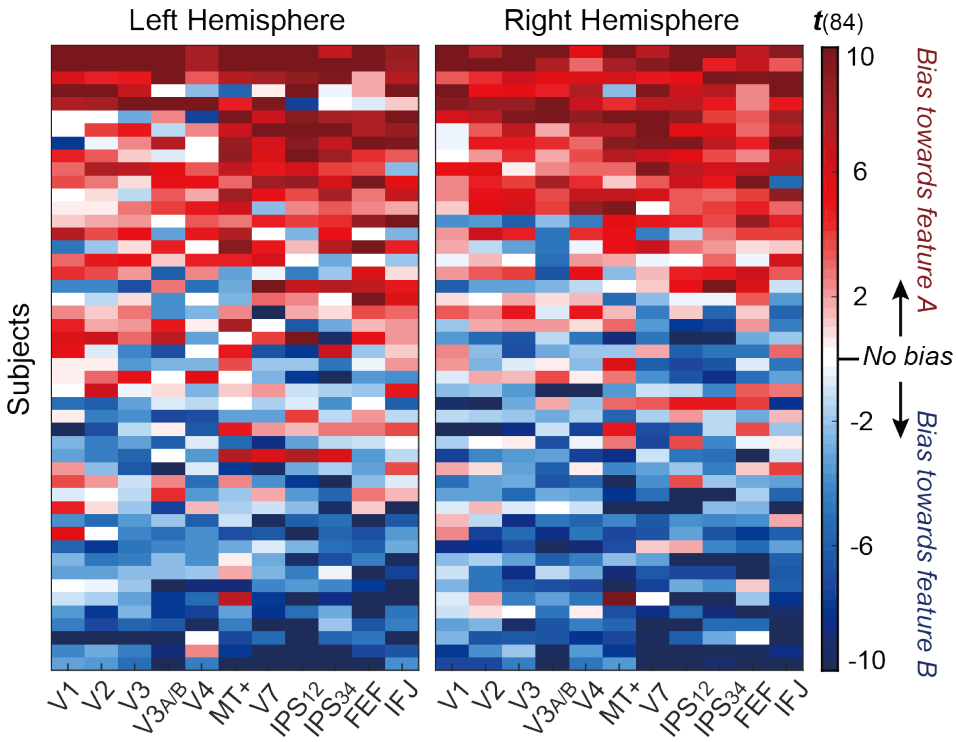

Figure 2. Biased neural coding of feature-based attentional modulation in individual brain areas and subjects. Data from left and right hemispheres are shown in separate graphs. Each row represents data of an individual subject, and each column represents a single brain area. Each cell is color-coded to indicate the *t*-value from the comparison of neural response between two attention conditions (85 voxels were used for this analysis). Red and blue colors indicate the direction of this difference, with their shade indicating the strength of the difference. We arranged this map by sorting the amount of bias per subject (indexed by the mean *t*-values across brain areas for each subject), such that from top to bottom, bias progressed from feature A to feature B.

Thus, although mean BOLD response *across subjects* never exhibited reliable differences between the two attention conditions (e.g., Fig. 1E), voxel responses *within a single brain area* exhibited reliable univariate difference in the majority of brain areas examined. The direction of such bias also remained fairly consistent across brain areas within each subject.

Biased representation increases from sensory to frontoparietal areas

Having observed significant univariate bias, we next quantified such bias to examine how it varies among cortical areas, using a ROC analysis. As an example, Figure 3A shows response distributions from the two attention conditions in a single brain area of a subject. Here the two distributions appear separable with the “attend feature B” condition having larger responses on average than “attend feature A”. This univariate bias can be indexed by the area under the ROC curve (abbreviated as AUC, Figure 3B), which provides a standardized, non-parametric measure of the separation between the two distributions, such that a value of 0.5 indicates no separation and a value of 1 indicates perfect separation. We calculated AUCs for all 1056 brain areas and averaged across subjects. Consistent with the t-test results above, we observed significant AUCs in many brain areas, compared to the significance threshold obtained from permutation tests (see *Method Details*), except for early visual areas including V1 and V2 in both hemispheres, and left V3 (Fig. 3B). Furthermore, there is a general trend of increasing AUC values along the cortical hierarchy. This was confirmed by a two-way repeated-measures ANOVA (11 brain areas × 2 hemispheres), showing a strong effect of brain areas ($F_{(10,470)}$=11.05, p<0.001, η^2=.190), without a significant difference between hemispheres (p=0.151), or interaction between hemisphere and brain areas (p=0.578).

To further characterize the main effect of brain areas, we grouped the 11 areas into four region groups based on anatomical considerations: V1, extrastriate visual areas (ExS), consisting of V2, V3, V3A/B, V4 and MT+, parietal areas (IPS), consisting of V7/IPS0, IPS12 and IPS34, and prefrontal network (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 analysis on AUCs to assess whether the biased representation varied with region group and stimulus domain.

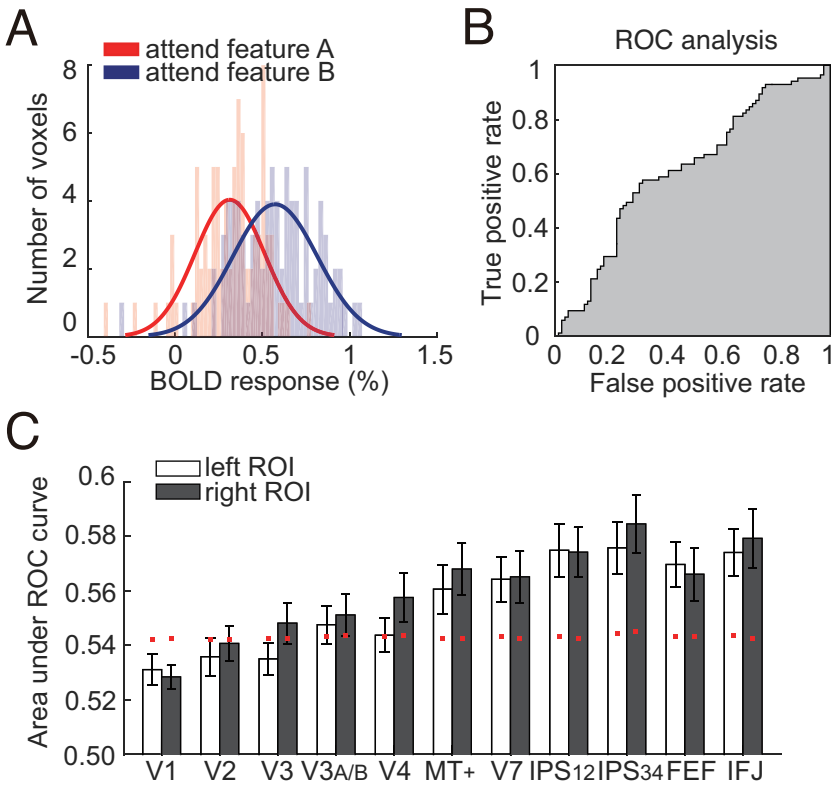

Figure 3. (A) Illustration of the ROC analysis in a single brain area. Histograms show the response distributions (across 85 voxels) in a brain area for the two attention conditions. Solid curves are normal density functions fitted to each distribution. These density curves are for visualization purpose only. (B) The AUC value is calculated from the area under the constructed ROC curve (gray shaded area), a non-parametric measure of the separation between two distributions. (C) Average AUC values across all subjects in all areas of interest. Red square indicates statistical significance threshold ($p<0.05$) obtained from a permutation test for each ROI.

Comparison across region groups. When grouped into four main anatomical groups, we observed a clear increase in AUC from V1 to extrastriate visual areas and further into parietal and frontal areas (Fig. 4A, red plot). A one-way repeated-measures ANOVA on AUCs revealed a significant effect of region group ($F_{(3,141)}=22.55$, $p<0.001$, $\eta^2=.324$). Pairwise comparisons further showed a stronger bias in frontoparietal areas

than that in V1 and ExS (*ps*<0.01), without a significant difference between IPS and PFC (*p*=1.0). To confirm that this result was not due to our specific voxel selection criterion, we repeated the same analysis using different number of voxels (n=100 and 150) and found similar results (Fig. 4A, cyan and black plots). A two-way repeated-measures ANOVA (4 region group × 3 voxel number) revealed a main effect of region group ($F_{(3,282)}$=21.56, *p*<0.001, η^2=.314) and voxel number ($F_{(2,282)}$=33.16, *p*<0.001, η^2=.414), as well as a two-factor interaction ($F_{(6,282)}$=5.13, *p*<0.001, η^2=.098), showing that greater bias was found with fewer voxels in ExS, IPS and PFC (*ps*<0.001), but not in V1 (p=0.49). Because we selected voxels based on the rank-ordered degree of activity during the task, this finding suggests that bias is stronger in more active voxels. Importantly, these results demonstrate the robustness of biased representation of attention in higher-order areas.

Comparison across stimulus domains. Because we used different stimuli (i.e., colors, rotating motion, linear motion directions and dynamic objects) across experiments, the amount of bias may differ across stimulus domains. When we plotted the AUCs separately for each stimulus type (Fig. 4B), we observed largest AUCs for colors, intermediate AUCs for motion directions, and smallest AUCs for the dynamic objects. This observation was confirmed by a mixed-effect ANOVA (4 region group × 4 stimuli), which showed a main effect of region group ($F_{(3,132)}$=23.93, *p*<0.001, η^2=.352) and stimuli ($F_{(3,44)}$=7.45, *p*<0.001, η^2=.337), as well as a significant two-factor interaction ($F_{(9,132)=}$2.29, *p*=0.020, η^2=.135). These results indicated that the regional differences on AUCs were more pronounced for attending to colors and rotating/linear motion directions (one-way ANOVA *ps*<0.01), but it became much weaker, and indeed,

statistically unreliable, for attending to dynamic objects (p=0.64). This finding suggests a potential decrease of bias along with increasing complexity of the attended information (e.g., from simple features to complex objects).

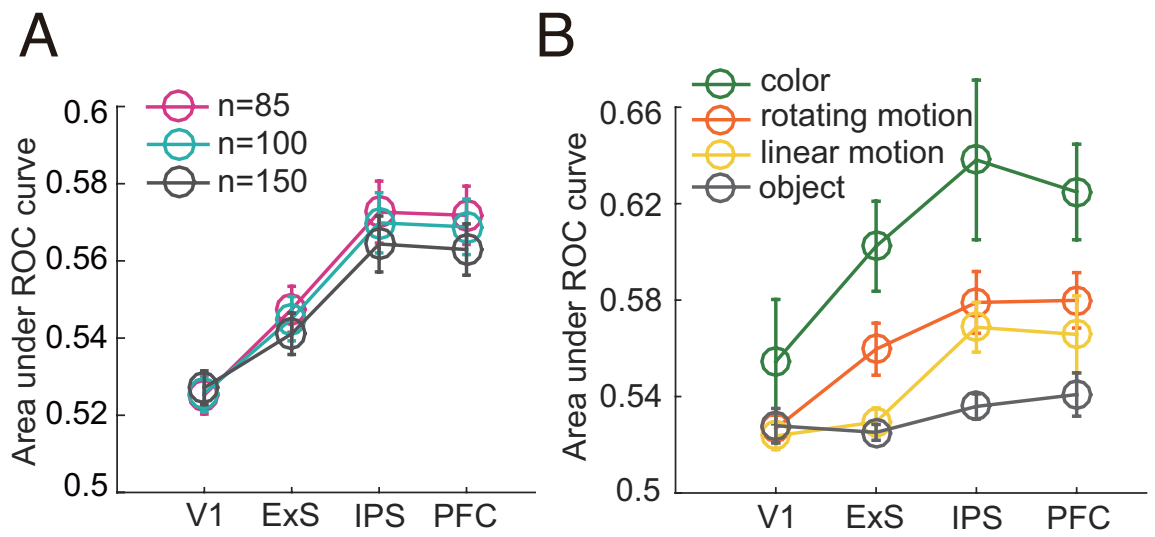

Figure 4. (A) Average AUC values for different region groups, obtained with different numbers of voxels (n=85, 100 and 150) in the ROC analysis. (B). Average AUC values for each stimulus type, obtained with data from different experiments. Error bar denotes standard error of the mean.

Bias removal produces dissociable effects in sensory and frontoparietal areas.

The above-chance AUCs demonstrate a univariate bias between two attention conditions in the majority of brain areas (except for early visual areas). Previously we have shown significant above-chance multivariate decoding using pattern classification techniques in all those areas (e.g., Fig. 1F). Given that both methods measure the neural discriminability between conditions, this raises the question of how much the univariate bias contributes to the multivariate decoding of attention. We used the grand-mean of BOLD signal from each attention condition as a proxy measure of this bias (equivalent to the means of each distribution in Fig. 3A). 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 *Method details*). 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. 5A). If multivariate decoding relies mostly on univariate differences, we would expect that removing the mean diminishes (or eliminates) attention decoding. Alternatively, if multivariate decoding relies on multi-dimensional pattern variability, we would expect little impact of mean removal on attention decoding. We found that mean removal had progressively stronger impact on decoding accuracy along the cortical hierarchy (Fig. 5B). This was confirmed by a two-way repeated-measures ANOVA (region group × mean removal), showing a main effect of mean removal ($F_{(1,141)}$=25.14, p<0.001, η^2=.349), and importantly, a significant interaction between region group and mean removal ($F_{(3,141)}$=14.07, p<0.001, η^2=.230). Follow-up t-tests showed that mean removal produced significant drop of MVPA-based decoding in ExS, IPS, and PFC (*ps*<0.001), but not in V1 (p=0.83). A significant interaction was also obtained if we excluded V1 data from the analysis ($F_{(2,94)}$=11.41, p<0.001, η^2=.195), indicating a relatively larger drop in decoding accuracy due to mean removal in IPS and PFC compared to ExS. Indeed, after mean removal, decoding accuracy in PFC dropped to the chance level (Figure 5B, red squares). 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, the MVPA decoding appears to rely on the univariate bias exclusively.

To further examine the contribution of univariate bias to MVPA-based decoding, we performed a correlation analysis between AUCs and MVPA decoding accuracies across subjects. In particular, for regions where the univariate bias contributes to MVPA-based decoding, we should expect a dependency between these two measures. Indeed, we found significant correlations between these AUC and MVPA decoding in frontoparietal areas (IPS: r=0.33, p=0.024; PFC: r=0.30, p=0.036), but no such correlations in sensory areas (V1: r=0.06, p=0.697; ExS: r=0.17, p=0.249). This result further supports the possibility that attention decoding in the frontoparietal network were mainly driven by the univariate bias.

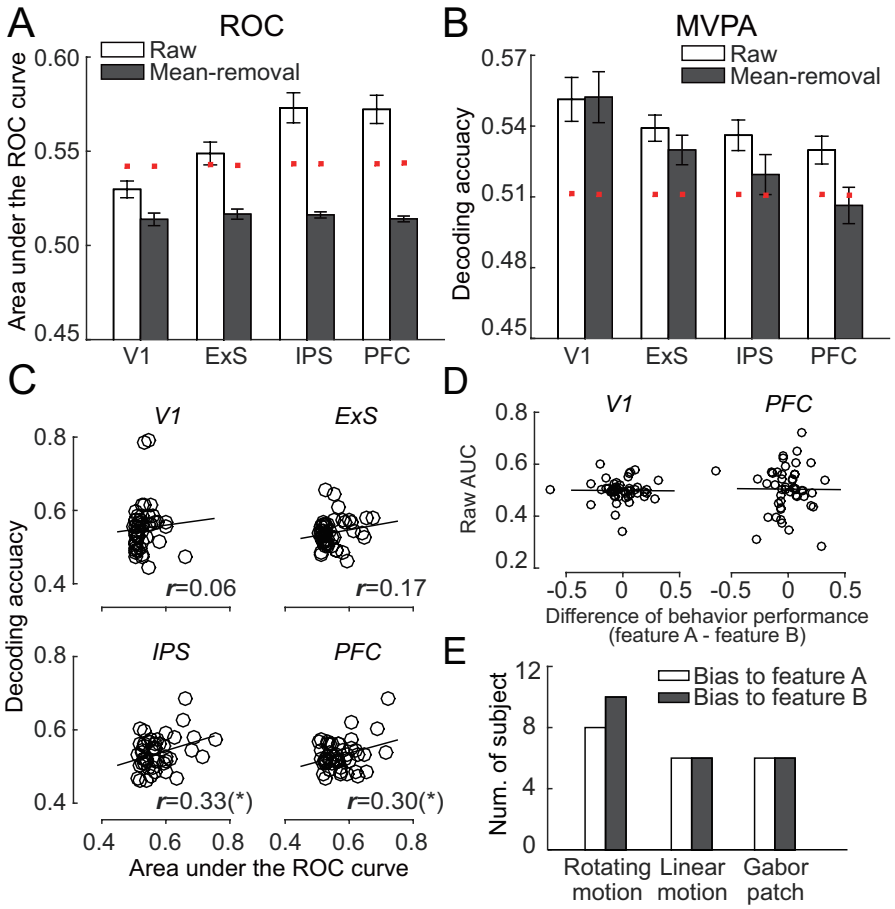

Figure 5. The comparison between (A) ROC analysis and (B) multivariate pattern analysis before and after removing the grand mean from each attention condition. Red squares indicate

statistical significance thresholds ($p<0.05$) obtained from permutation tests. Error bar denotes SEMs. (C) Inter-subject correlation between AUC and MVPA decoding accuracy for each region group. (D) Correlations between raw AUC values and behavioral difference between two attention conditions in two representative areas (V1 and PFC). (E) Group-level bias distribution across subjects.

Biased coding does not correlate with behavioral selection

We examined if 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, even though subjects were always instructed to attend equally to individual features. Such a preferential selection should lead to better behavioral performance in our threshold tasks. We thus tested this possibility by examining neural and behavior relationship, by correlating the difference in behavioral performance 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 (Fig. 5D, see *Method details).* This analysis revealed no such correlations in any of the region groups between the magnitude of the neural bias and behavioral preference (*ps*>0.36), ruling out the account that the biased neural coding of attention is due to behavioral preferences.

Equivalent distribution of biased coding at the group level

Lastly, we examined the group-level distribution of the biased coding (see *Method details*). 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 5E shows approximately equal distribution of biased direction across subjects (Chi-square test against equal proportion: *ps*=1.0). 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 cancelled each other out.

Discussion

We observed a biased population neural coding that represents the attentional priority in a large fMRI dataset containing 48 subjects across multiple experiments. This biased coding was robust, as demonstrated by a convergence of findings from single brain area and consistency across multiple brain areas within a subject. At the level of brain areas, we found that voxels in the majority of 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, which also explains the lack of group-average univariate difference in the standard analyses. Importantly, when we quantified the amount of univariate bias using the ROC analysis, we found a progressively stronger biased response from sensory to frontoparietal areas. We ruled out the possibility that the results were due to specific voxel selection criterion, stimulus domain or behavioral preference. Collectively, our findings provide the first evidence for a biased neural coding scheme of cognitive variables in the human brain.

The initial observation of biased coding (Fitzgerald et al., 2013) in non-human primates was indeed rather surprising (Chafee, 2013). However, it is unknown whether such bias is only present in non-human primates who typically undergo extensive training in cognitive tasks before neural recording experiments, or the specific brain area examined, namely LIP. Our results thus suggest that a similar bias is also present

in the human brain under different tasks, which requires much less training and engages a multitude of brain areas. Here we further leveraged the large sample size and whole-brain coverage afforded by fMRI and show that the direction of bias is consistent among brain areas within individual subject, but it varies across subjects. We note that the magnitude of the observed bias here is weaker than the neuronal level data, as Fitzgerald et al., observed nearly all recorded neurons had similar direction of bias. There are likely multiple reasons for this difference in magnitude, such as differences in species, stimulus/task, measured signals (neural spiking vs. BOLD response), and selection criteria (feature selective neurons vs. stimulus-responsive voxels regardless of feature selectivity). Notwithstanding these methodological differences, the findings from both studies converge in demonstrating a biased neural coding for cognitive variables.

Notably, we found a dissociable pattern of bias between early visual areas and frontoparietal areas. First, the magnitude of bias grows progressively larger from early visual areas to frontoparietal areas, with an absence of bias in early visual areas. Second, the MVPA analysis on mean removed data suggests that the univariate bias made progressively larger contribution to pattern-based decoding as the cortical hierarchy ascends. In particular, the chance-level decoding in PFC stands in sharp contrast to the unchanged decoding accuracy in V1 after removing the univariate bias (Fig. 5B). The contribution of the univariate bias to pattern-based decoding was further supported by the finding of significant correlation between AUC values and decoding accuracies in the frontoparietal areas. These contrasting results thus support a

distinction between two coding mechanisms: distributed population coding in sensory areas and biased population coding in high-level areas.

From a computational point of view, our results may reflect different dimensionality of the neural signals at different levels of cortical hierarchy. Specifically, the differential impact of removing the grand mean on multivariate decoding suggests that sensory areas contain high-dimensional neural signals whereas frontoparietal areas contain low-dimensional, possibly one-dimensional, neural signals (see Supplementary Materials for more quantitative explanation of this rationale). Such a coding scheme is also supported by 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 among cortical areas could be one manifestation of this transition.

Analog coding is naturally implemented with sensory neurons that are continuously tuned to stimulus features (Hubel and Wiesel, 1962; Blasdel, 1992; Maunsell, and Van Essen, 1983). Consistent with this idea, we observed a distributed population code 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 (Serences et al., 2009; Harrison and Tong, 2009; Riggall and Postle, 2012) from activity patterns in early visual areas. Together,

these findings support the distributed population coding of cognitive variables in early visual areas.

While analog coding 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 (Bisley and Goldberg, 2010; Kastner and Ungerleider, 2000) and also part of the Multiple-Demand (MD) network (Duncan, 2013; Fedorenko et al., 2013). 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 system (see Fitzgerald et al., 2013). Furthermore, if a cortical area provides a scalar output, it will also simplify the readout of downstream area and 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 (i.e., scalar coding). The investigators suggested that by reducing local neural signals to one-dimensional activity, the brain can achieve robust temporal control of behavior such as the timing in shifting attention and crossing a decision threshold during evidence accumulation. Although these ideas of the 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, especially if recurrent activity is weak in individual areas (Ganguli, et al., 2008). A limitation of previous single-unit work is that all 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 wide-spread MD 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 considering the generality of biased neural representation. Given our human fMRI data and previous monkey single-unit data were obtained from a variety of behavioral paradigms, biased representation appears to be a general principle of neural coding. However, we should note that a commonality shared among these behavioral paradigms is that all tasks entail a few discrete task conditions. It is possible that biased representation is particularly useful in this type of regimes. Theoretical studies suggest that neural dimensionality could scale with task complexity (Fusi, Miller, & Rigotti, 2016; Gao and Ganguli, 2015). This idea has found some support in studies where the number of task conditions appears to drive estimates of dimensionality in

monkey PFC (Brincat, et al., 2018; Rigotti et al., 2013). There is also some hint in our data supporting this notion, as the univariate bias was weaker for dynamic multi-feature objects than single features (Figure 4B). Future studies are necessary to systematically evaluate the influence of task and stimulus complexity on the dimensionality of neural signals in the association cortex.

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 if the direction of the bias persists over a longer period of time. It is possible that the direction of such bias is determined by each individual's past experience, hence more or less fixed for that individual, or alternatively, such bias arises stochastically when performing a particular task. It would be interesting to examine the consistency and origin of the biased neural coding in future studies.

In conclusion, we highlight biased neural coding as a potential mechanism for representing cognitive variables in the brain. Although the simplicity of this coding scheme seems counterintuitive, it can facilitate a robust representation and simple read-out of information critical for stimulus selection and cognitive control. Together with the findings of distributed population coding in early sensory areas, our results suggest a gradual transition from high- to low-dimensional representation along the cortical hierarchy. Such a gradient of cognitive coding could play complementary roles at different stages of processing, enabling representations at multiple levels of abstraction to support adaptive behavior.

STAR*METHODS

Detailed methods are provided including the following:

- KEY RESOURCES TABLE
- CONTRACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
    - Overview of the experimental procedures
    - Retinotopic mapping
- QUANTIFICATION AND STATISTICAL ANALYSIS
    - Univariate Analysis: Deconvolution
    - Voxel Selection and Response Calculation
    - Receiver Operating Characteristic (ROC) Analysis
    - Multivariate Pattern Analysis (MVPA)

Acknowledgements

We thank Dr. David Zhu and Ms. Scarlett Doyle for their assistance in collecting the neuroimaging data. We also thank Michael Jigo and Youyang Hou for collecting some of the datasets.

Declaration of Interests

The authors declare no competing financial interests.

Supplementary Information

Supplementary information includes two figures and texts of the simulation method and results.

Reference

Bettencourt, K. C., & Xu, Y. (2016). Decoding the *content of visual short-term memory under distraction in occipital and parietal areas.* Nat. Neurosci, *19*, 150-7.

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

Bisley, J. W., & Goldberg, M. E. (2003). Neuronal activity in the lateral intraparietal area and spatial attention. Science, *299*, 81-86.

Blasdel, G. G. (1992). Orientation selectivity, preference, and continuity in monkey striate cortex. J. Neurosci., *12*, 3139-3161.

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. J. Neurosci., *29*, 13992-14003.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: A comparison of neuronal and psychophysical performance. J. Neurosci., *12,* 4745–4765.

Brincat, S. L., Siegel, M., von Nicolai, C., & Miller, E. K. (2018). Gradual progression from sensory to task-related processing in cerebral cortex. Proc. Natl. Acad Sci., *115*, E7202-E7211.

Chafee, M. V. (2013). A scalar neural code for categories in parietal cortex: representing cognitive variables as “more” or “less.” Neuron, *77*, 7–9.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. Nat. Rev. Neurosci., *3*, 201.

Deco, G., & Rolls, E. T. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. Vis. 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., *93:2382–2386.*

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

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., *110*, 16616–16621.

Fitzgerald, J. K., Freedman, D. J., Fanini, A., Bennur, S., Gold, J. I., & Assad, J. A. (2013). Biased associative representations in parietal cortex. Neuron, *77*, 180-191.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. Science, *291*, 312-316.

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. Curr. Opin. Neurobiol., *37*, 66-74.

Jigo, M., Gong, M., & Liu, T. (2018). Neural Determinants of Task Performance during Feature-Based Attention in Human Cortex. *ENeuro*, *5*.

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.

Gao, P., & Ganguli, S. (2015). On simplicity and complexity in the brave new world of large-scale neuroscience. Curr. Opin Neurobiol, *32*, 148-155.

Gong, M., & Liu, T. (2019). Continuous and discrete representations of feature-based attentional priority in human frontoparietal network. Cogn. Neurosci.

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. Nature, *458*, 632.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. Neuron*, 36*, 791-804.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. J. Physiol., *160*, 106-154.

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

Kamitani, Y., & Tong, F. (2006). Decoding seen and attended motion directions from activity in the human visual cortex. Curr. Biol., *16*, 1096-1102.

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.

Ganguli, S., Bisley, J. W., Roitman, J. D., Shadlen, M. N., Goldberg, M. E., & Miller, K. D. (2008). One-dimensional dynamics of attention and decision making in LIP. Neuron, *58*, 15-25.

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

Kastner, S., & Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. Ann. Rev. Neurosci., *23,* 315-341.

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., Hospadaruk, L., Zhu, D. C., & Gardner, J. L. (2011). Feature-specific attentional priority signals in human cortex. J. Neurosci., *31*, 4484-4495.

Liu, T., & Hou, Y. (2013). A hierarchy of attentional priority signals in human frontoparietal cortex. J. Neurosci., *33*, 16606-16616.

Liu, T. (2016). Neural representation of object-specific attentional priority. Neuroimage, *129*, 15-24.

Maunsell, J. H., & Van Essen, D. C. (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.

Scolari, M., Seidl-Rathkopf, K. N., & Kastner, S. (2015). Functions of the human frontoparietal attention network: Evidence from neuroimaging. Curr Opin Behav. Sci., 1, 32-39.

Serences, J. T., Ester, E. F., Vogel, E. K., & Awh, E. (2009). Stimulus-specific delay activity in human primary visual cortex. Psy. 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.

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.

Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. Neuron, *78*, 364-375.

Swaminathan, S. K., & Freedman, D. J. (2012). Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. Nat. Neurosci., *15*, 315.

Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. Nat. Rev. Neurosci., *1*, 125.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nat. Neurosci., *2*, 1019.

Riggall, A. C., & Postle, B. R. (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, M. R., Wang, X. J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. Nature, *497*, 585.

STAR*METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
MATLAB	MathWorks	https://www.mathworks.com
MGL	Stanford University	http://gru.stanford.edu/doku.php/mgl/overview
mrTools	Stanford University	http://gru.stanford.edu/doku.php/mrTools/overview
Palamedes	Prins and Kingdom, 2009	http://www.palamedestoolbox.org
FreeSurfer	Harvard University	http://surfer.nmr.mgh.harvard.edu

CONTACT FOR REAGENT AND RESOURCE SHARING

As Lead Contact, Taosheng Liu is responsible for all reagent and resource requests. Please contract Taosheng Liu at tsliu@msu.edu with requests and inquiries.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

In total, forty-eight subjects from Michigan State University were included across five experiments. All had normal or corrected-to-normal vision. Eleven were right-handed and one was left-handed. Participants were paid for their participation at $20/hr and gave informed consent according to the study protocol approved by the Institutional Review Board at Michigan State University.

METHOD DETAILS

Overview of the Experimental Procedures

We re-analyzed data from five previously published fMRI experiments (Liu et al., 2011; Liu, 2016; Jigo et al., 2018; Gong and Liu, 2019). 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 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, six and twelve 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, twelve subjects attended to a linear motion direction in a superimposed up-left/up-right moving dot field (Gong and Liu, 2019). In the fifth experiment (Liu, 2016), twelve 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 hr 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.

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+, IPS1 to IPS4.

QUANTIFICATION AND STATISTICAL ANALYSIS

Univariate Analysis: Deconvolution

We used the deconvolution approach by fitting each voxel's time series with a general linear model whose regressors modeled the two attention conditions with finite impulse responses. The design matrix was pseudo-inversed and multiplied by the time series to obtain an estimate of the hemodynamic response (HRF) 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 voxel's response over time is correlated with the attention task. The statistical significance of the r^2 value was evaluated by a permutation test that repeated the deconvolution analysis with shuffled trial labels

(Gardner et al. 2005; Liu et al. 2011). The p-value of each voxel was calculated as the percentile of voxels from the null (permutated) distribution that exceeded the observed r^2 value. We used r^2 value in conjunction with the anatomical constraints to define 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.

Voxel Selection and Response Calculation

We performed our analyses on voxels that exhibited statistically significant activity in our tasks ($p<0.05$, based on r^2 value as determined above), and eliminated noisy voxels with more than 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 >90% of all ROIs (1056 in total). For ROIs that resulted in fewer voxels using this criterion, we used all voxels that satisfied the criterion (average: 64 voxels). We also repeated the analyses using different number of sorted voxels at 100 and 150 voxels).

To determine the time window that takes into account the types of analysis (univariate vs. multivariate) and the difference in study design (block- vs. event-related) across experiments, we calculated the voxel response by averaging different time windows. For univariate analysis, we averaged time points that showed elevated response on the basis of deconvolved response (2.5 – 20 s after trial onset for block-design in Exp. 1 and 2; 4.4 - 11 s after trial onset for event-related design in Exp. 3 to 5). For multivariate analysis, we extracted time points from each trial in the raw time series,

while taking into account of the hemodynamic delay and minimizing signal contributions from adjacent trials (5 – 22.5 s after trial onset for Exp. 1 and 2; 4.4 – 11 s after trial onset for event-related design in Exp. 3 to 5).

Receiver Operating Characteristic (ROC) Analysis

For each subject and each ROI, we performed a receiver operating characteristic (ROC) analysis on the voxel-wise deconvolved response to quantify the univariate discriminant information between the two attention conditions. The ROC analysis is a desirable tool because of its invariance to the scale of data and decision criteria. The area under the ROC curve (abbreviated as AUC) indicates the reliability with which an ideal observer could distinguish the attended feature given the response distributions for both conditions (e.g., Britten et al., 1992). In general, we rectified the raw AUC values around 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. Because the rectified AUC values were always greater than 0.5 (theoretical chance level), we assessed the statistical significance of AUC using a permutation test. We shuffled the labels of voxel response to obtain an AUC of the shuffled data and repeated this procedure 1000 times to obtain a null distribution of rectified AUCs for each subject and each ROI. To compute the group-level significance, we averaged the null distributions over subjects to obtain a group-level distribution of AUC values for each brain area. The 95 percentile of this group-level distribution was thus defined as the statistical significance level (corresponding to $p=0.05$) to determine if an observed AUC value significantly exceeded the chance level. For the correlation analysis between behavioral difference and AUCs (Fig. 5D), we did

not rectify the AUCs such that the direction of neural bias (A>B vs. A<B) was represented by AUCs greater or less than 0.5, respectively.

Multivariate Pattern Analysis (MVPA)

For each voxel and each ROI, we obtained single-trial fMRI response amplitude (see *Voxel selection and response calculation*). We thus obtained a m×n instance matrix for each ROI and attention condition, where m was number of trials and n was number of voxels. We then performed MVPA to discriminate between the two conditions using Fisher linear discriminant analysis. We performed leave-one-run-out cross-validation to evaluate the classification accuracy, by dividing the data set into testing (one run) and training data (remaining runs). This procedure was repeated until each run was tested once. Classification accuracy was averages across folds for each ROI. To assess the contribution of biased coding to MVPA, we subtracted the grand mean of each m×n instance matrix from the instance matrix itself (i.e., mean removal separately for each attention condition) before applying the same MVPA analyses 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 for 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 95 percentile of this group-level distribution was thus defined as the statistical significance level (corresponding to p=0.05) to determine if an observed MVPA decoding accuracy significantly exceeded the chance level. In all analyses where

multiple statistical tests were conducted, we corrected the p-values using the false discovery rate method (Benjamini and Hochberg, 1995).

Supplementary Materials

In the main analysis, we evaluated the effect of removing the grand mean of the BOLD response within each attention condition on multivariate pattern decoding. We found that cortical areas exhibited differential sensitivity to such an operation and inferred the difference of underlying dimensionality of the neural signals. Specifically, in visual areas, the neural signal is likely high-dimensional because removing the mean difference hardly had an impact on the pattern-based decoding accuracy, whereas in frontoparietal areas, the neural signal seems low-dimensional because removing the mean significantly reduced the decoding accuracy (see Fig. 5B). This inference relies on the idea that the overall mean is a single scalar and hence by definition a one-dimensional signal. Therefore, the approach of mean removal allows us to evaluate the contribution of one-dimensional signal to pattern-based decoding.

Although the rationale above seems straightforward and reasonable, we nevertheless conducted simulations to confirm the validity of this inference. This rationale predicts that mean removal will have different impacts on data with varying intrinsic dimensionality. To generate multivariate data of known dimensionality, we adopted the pattern component model (PCM), a framework that has been used to model and recover the dimensionality of neural population responses (Diedrichsen, et al., 2011; Diedrichsen & Kriegeskorte, 2017). The PCM is a standard linear mixed model that assumes that observed neural data are caused by a set of underlying pattern components (U), which is linearly combined with a set of latent features (F, Fig. S1). The pattern components represent all the possible distinct neural patterns in a given brain area and the latent features are essentially weights on these pattern components.

The pattern component and latent features thus define a neural feature space, with the number of such components (columns in U, and also the number of latent features, rows in F) as the dimensionality of the neural representation. This is analogous to how we conceptualize a three-dimensional Cartesian space, where each point's coordinate (x, y, z) can be considered as features (or weights) on three components (three cardinal vectors $\vec{\imath}, \vec{\jmath}, \vec{k}$). More precisely, PCM assumes that the neural data can be generated by the following process:

$$\boldsymbol{y_{k,n}} = \sum_{d=1}^{D} \boldsymbol{u_d} f_k^{(d)} + \varepsilon_{k,n} \qquad \text{Eq. 1}$$

where $y_{k,n}$ is the multi-voxel pattern on a single trial (Px1, P is the number of voxels/neurons), u_d is a pattern component, and f_k is a feature vector associated with each condition. Note k can assume any value in [1, K] (K is the total number of conditions); n can assume any value in [1, N] (N is the number of trials per condition). In this formula, $y_{k,n}$ is a linear combination of each u_d weighted by the dth element of the latent feature vector for condition k ($f_k^{(d)}$), plus random noise (ε).

Here, we focus on the consequence of mean removal on multivariate pattern classification. We generated simulated data with 1, 2, 3, 4 dimensions (D) using Eq. 1 with the following parameters: K=2, P=100, N=40. This yields a PxN data matrix for each of the two conditions at each level of assumed dimensionality. We then performed the receiver operating characteristics (ROC) analysis and multivariate pattern analysis (MVPA) on these simulated data using the same methods as we analyzed the real data. To create a reliable estimate of performance, we repeated this process for 100 times and examined the mean of AUC values and MVPA decoding accuracies. We adjusted the parameters of the multivariate normal distributions that controlled U, as well as the

additive noise (ε), so that MVPA decoding was at an intermediate level for all simulated datasets. We note that the qualitative patterns of the results are robust with respect to variations in the model parameters.

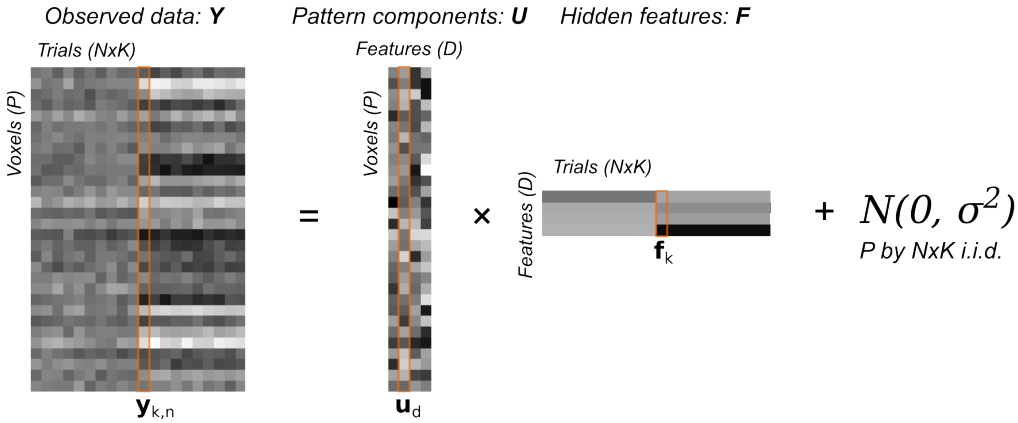

Figure S1. Illustration of the pattern component model. Neural data (***Y***) can be decomposed as a multiplication of pattern components (***U***) and hidden features (***F***), plus Gaussian noise. Columns in both ***U*** and ***F*** were sampled from multivariate normal distributions. *P*: number of voxels, *N*: number of trials per condition, *K*: number of stimulus conditions, *D*: number of dimensions. In this illustration, *P*=30, *N*=10, *K*=2, *D*=4. Note feature vector $\mathbf{f_k}$ is the same for trials within a condition. The PCM is a general framework that can accommodate any combination of P, N, K, and D.

The simulation results were shown in Figure S2. For ROC analysis, removing the mean from each condition led to a drop of AUC to essentially the chance level of 0.5, regardless of the intrinsic dimensionality of data. This is expected because AUC reflected univariate difference and can be eliminated with mean removal. For the MVPA result, mean removal progressively reduced decoding accuracy along with decreased data dimensionality, supporting our rationale that mean removal eliminated the contribution of one-dimensional signal to pattern-based decoding such that decoding of low dimensional data will be impaired more than decoding of high dimensional data.

The simulations assumed that pattern components are multivariate normal distributions, which generated data matrix (PxN) of equal covariance between the two

conditions. This could be a necessary condition for the above results. We thus assessed whether the empirical fMRI data had comparable structure to the simulated data, by testing the equality of the covariance matrix between two attention conditions. Because the data tend to have fewer trials than voxels (i.e., N < P), we used a testing procedure developed for high-dimensional and sparse setting (Cai et al., 2018). Of the 1056 tests, we found the null hypothesis of equal covariance was rejected only 4 times at alpha of 0.05. Thus there is no statistical evidence that the empirical fMRI data had different covariance structure between the two attention conditions, supporting our use of the simulation to make inferences about the structure of real data.

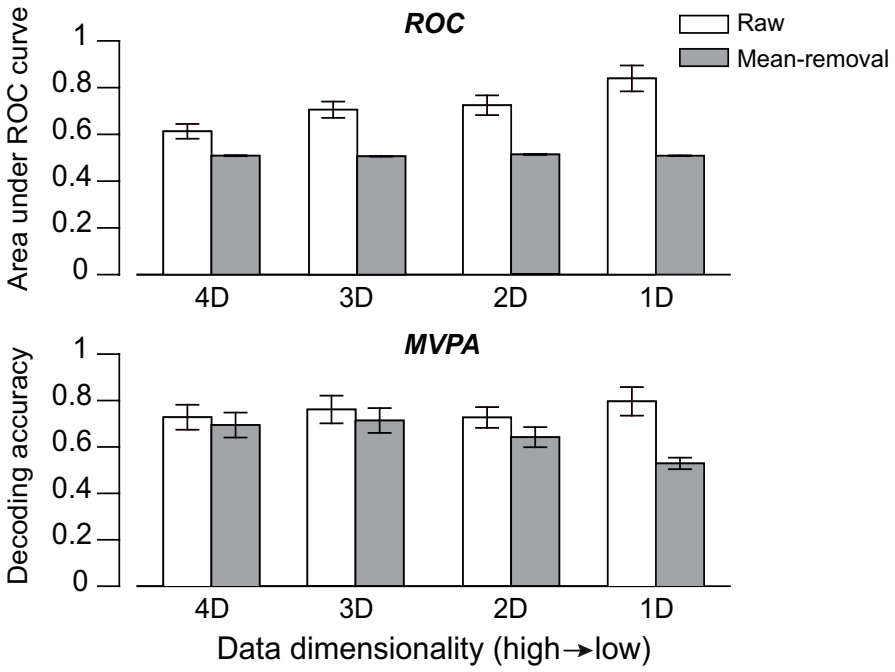

Figure S2. Simulation results for datasets with different dimensions (D=4, 3, 2, 1). Results are shown for the ROC analysis (top) and the MVPA decoding (bottom), before and after removing the mean from each condition. Error bar denotes SEMs.

References

Cai, T., Liu, W., & Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. J. Am. Stat. Assoc., *108*, 265–277.

Diedrichsen, J., Wiestler, T., & Ejaz, N. (2013). A multivariate method to determine the dimensionality of neural representation from population activity. NeuroImage, *76*, 225–235.

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. PLoS Comput. Biol., *13*, e1005508.