nature communications

Article

https://doi.org/10.1038/s41467-024-50717-y

Prefrontal neuronal dynamics in the absence of task execution

Received: 16 September 2022

Accepted: 18 July 2024

Published online: 06 August 2024

Check for updates

Shusen Pu[1,2,6], Wenhao Dang[1,6], Xue-Lian Qi[3] & Christos Constantinidis[1,4,5]

Prefrontal cortical activity represents stimuli in working memory tasks in a low-dimensional manifold that transforms over the course of a trial. Such transformations reflect specific cognitive operations, so that, for example, the rotation of stimulus representations is thought to reduce interference by distractor stimuli. Here we show that rotations occur in the low-dimensional activity space of prefrontal neurons in naïve male monkeys (*Macaca mulatta*), while passively viewing familiar stimuli. Moreover, some aspects of these rotations remain remarkably unchanged after training to perform working memory tasks. Significant training effects are still present in population dynamics, which further distinguish correct and error trials during task execution. Our results reveal automatic functions of prefrontal neural circuits allow transformations that may aid cognitive flexibility.

Neurons in the prefrontal cortex (PFC) represent sensory stimuli in a dynamic, task-specific manner[1–3]. Analysis of population activity with methods of dimensionality reduction reveals that stimuli are typically represented in a low dimensional space, or manifold, with the majority of the firing rate variance captured by a few dimensions[4,5]. Furthermore, the representation of stimuli in the reduced space changes dynamically during the course of a trial, as the subjects perform a cognitive operation according to task demands; the stimulus manifold may therefore be projected to a different subspace, rotated, or otherwise geometrically transformed[6–8]. Orthogonal rotation of a stimulus representation has been proposed as a possible mechanism for reducing interference between sensory and memory representations, protecting the memory of an initial stimulus from the interference of a subsequent stimulus presentation[9]. The same stimuli are represented in different subspaces when used in the context of different tasks[10] and errors are characterized by changes in stimulus representation geometry[11].

While it is clear that geometric transformation of stimulus information in neuronal populations occurs during cognitive operations[12], less is known on how the acquisition of a cognitive task may alter stimulus representation geometry. We thus addressed this question by analyzing prefrontal populations, both before and after subjects were trained to perform working memory tasks involving identical stimuli. We examined multiple aspects of the activity space, including subspace alignment, geometrical similarity, and dynamics, to determine the relative contribution of training and regional specificity in the formation of the observed low dimensional geometry.

Results

Neurophysiological recordings were collected from the lateral PFC of six monkeys in total; pre-training data were acquired from all animals, and post-training data were recorded from three of these six subjects (Table S1)[13,14]. Once fully trained, the monkeys viewed two stimuli appearing in sequence with intervening delay periods between them and reported whether the second stimulus (sample) was the same as the first stimulus (cue) and constituted a match, or was different and constituted a nonmatch (Fig. 1A, B). The stimulus sets used in these experiments varied in terms of spatial location or shape (Fig. 1A, B). Recordings were also obtained from these animals, viewing the same stimuli presented with the same timing prior to any training in the task. A total of 1164 neurons from six monkeys in five prefrontal subdivisions (Fig. 1C) were recorded during the passive, pre-training viewing of spatial stimuli; 847 neurons were recorded during the passive, pre-training viewing of feature stimuli. Additionally, 1031 neurons from three monkeys were obtained after training, when the animals performed actively the spatial working memory task; 796 neurons were

[1]Department of Biomedical Engineering, Vanderbilt University, Nashville, TN 37235, USA. [2]Department of Mathematics and Statistics, University of West Florida, Pensacola, FL 32514, USA. [3]Department of Neurobiology & Anatomy, Wake Forest School of Medicine, Winston-Salem, NC 27157, USA. [4]Neuroscience Program, Vanderbilt University, Nashville, TN 37235, USA. [5]Department of Ophthalmology and Visual Sciences, Vanderbilt University Medical Center, Nashville, TN 37232, USA. [6]These authors contributed equally: Shusen Pu, Wenhao Dang. e-mail: Christos.Constantinidis.1@vanderbilt.edu

https://doi.org/10.1038/s41467-024-50717-y

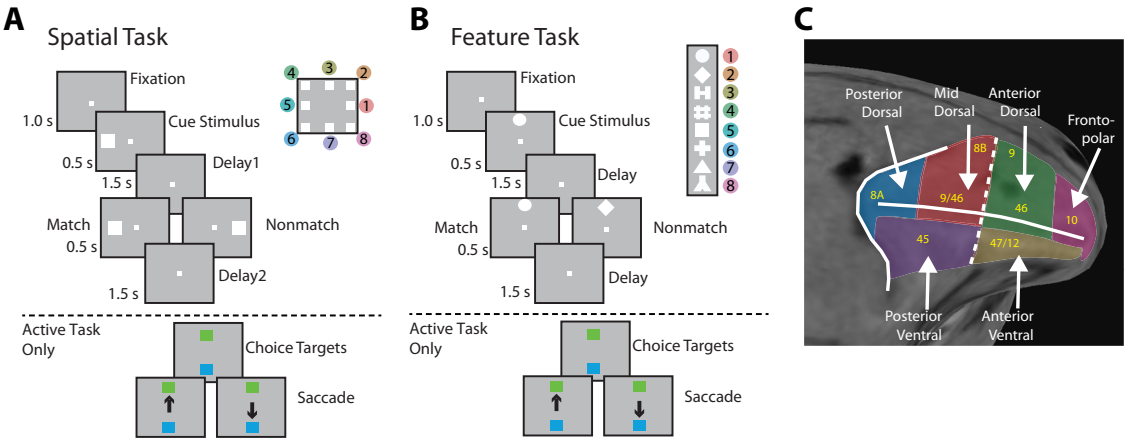


Fig. 1 | Task structure and recording areas. The animals were required to maintain center fixation throughout both active and passive task trials. At the end of active tasks trials, monkeys were required to make a saccade to a green target if the stimuli matched or to a blue target if the stimuli did not match. **A** Spatial match-to-sample task; the eight cue locations analyzed are shown in the inset. **B** Shape feature match-to-sample task; eight possible shapes in a session shown are shown in the inset. **C** Neurons were recorded from 5 subdivisions of lateral PFC, which we refer to as Anterior Dorsal (AD), Anterior Ventral (AV), Mid-Dorsal (MD), Posterior Dorsal (PD), and Posterior Ventral (PV).

obtained during the active feature working memory task (see Supplementary Table S1).

Spatial stimuli are geometrically arranged in a low-dimensional subspace

We applied Principal Component Analysis (PCA) on the covariance matrix of the zero-centered data, and selected eigenvectors sorted in decreasing order in terms of explained variance. In agreement with previous studies suggesting that prefrontal neurons represent stimuli in a low-dimensional manifold[4,5], we found that, on average, 73% of the variance across brain regions and training phases could be explained by the first three principal components across brain regions and training phases (Fig. S1A and Table S2). This result validated our approach of visualizing and measuring geometry structure in low-dimensional space.

Previous studies have also revealed that spatial information is arranged in a well-structured geometry resembling the physical appearance of visual stimuli in working memory[6,15], however, it is not clear whether task training may be responsible for the development of this geometry. Our analysis revealed that the geometrical arrangement of spatial stimuli in the neural population space exhibited a ring-like pattern (Fig. 2). This observation was made within the subspace defined by the first two principal components, which preserved the relative position of stimuli in the low-dimensional subspace (Fig. 2A). In contrast, shape stimuli did not demonstrate this pattern (see Fig. S12).

While no task training was required for the establishment of this geometric structure, we investigated if training could enhance its order regularity. More specifically, we explored whether training could increase the order similarity between the empirical geometry order of the eight stimulus conditions and the natural order of eight vertices in an octagon (Fig. 2A). To quantify this similarity, we calculated the unexplained variance between the empirical geometry order and the 'standard' order that maximizes the distance between diametric locations. The order regularity was defined by the difference in the unexplained variance between the empirical dataset and an order-shuffled dataset. By comparing different prefrontal subdivisions, we found that for the cue period, the anterior and middle regions of PFC already exhibited a strong alignment with the 'standard' order before training, while for the first delay period, all areas showed little to no order regularity before the animals were trained for working memory tasks. We also found that in general, training increased the "order regularity" for mid-posterior regions, for both the cue and the delay period. This was most pronounced for the delay period in the mid-dorsal and posterior-dorsal subdivisions (Fig. 2B, nonparametric bootstrap test, for pre- vs. post-training order regularity, MD, PD, and PV, $p < 0.001$ in each case).

Some subspace transformations are task independent

Since low-dimensional stimulus representations in neuronal population activity are transformed during cognitive operations[7–9], we sought to test how such transformations differ systematically before and after training to perform a cognitive task. Previous research suggested that the rotation under different contexts is beneficial for the task, since when two subspaces are orthogonal variation in one subspace will have near zero variation in the other (Fig. 3A2). For example, when one plane in a 3-dimentional space is viewed from the space spanned by another plane, it would have a much larger projection if the two planes are parallel to each other, whereas if they are orthogonal the first plane would only project as a line to the other. We thus quantified transformations based on the rotation of subspaces in different epochs. This was done by measuring the primary angle between the low-dimensional subspaces that accounted for the most variance in different contexts (Fig. 3A1). Measured rotation angles were compared to a baseline rotation, to control for the difference in response variation and selectivity for various PFC subdivisions. This baseline rotation was measured by calculating the rotation between the empirical dataset, and a synthetic dataset made of simulated units, of equal mean firing rate to the real units for each stimulus condition (see Methods: Dimensionality reduction and rotation of subspaces). Alternatively, the empirical data were split into two halves and the rotation angle was quantified between these two halves. The two methods gave qualitatively similar results (Figs. 3, 4 vs. Fig. S2), and we present results from the first method in most following figures. We examined specifically the geometry of our stimulus set in neuronal activity during the cue presentation and the delay period that followed it; during the cue and match presentations; and during the match and nonmatch presentations. (see "Dimensionality reduction and rotation of subspaces" in Methods).

Strikingly, even before any task training, representations of the same stimuli during the cue and match period already exhibited significant rotation angles in multiple brain areas, for both the spatial and feature sets (Fig. 3B1, B2). Our analysis further revealed that the pattern of transformation between the cue and match stimuli was highly area-

https://doi.org/10.1038/s41467-024-50717-y

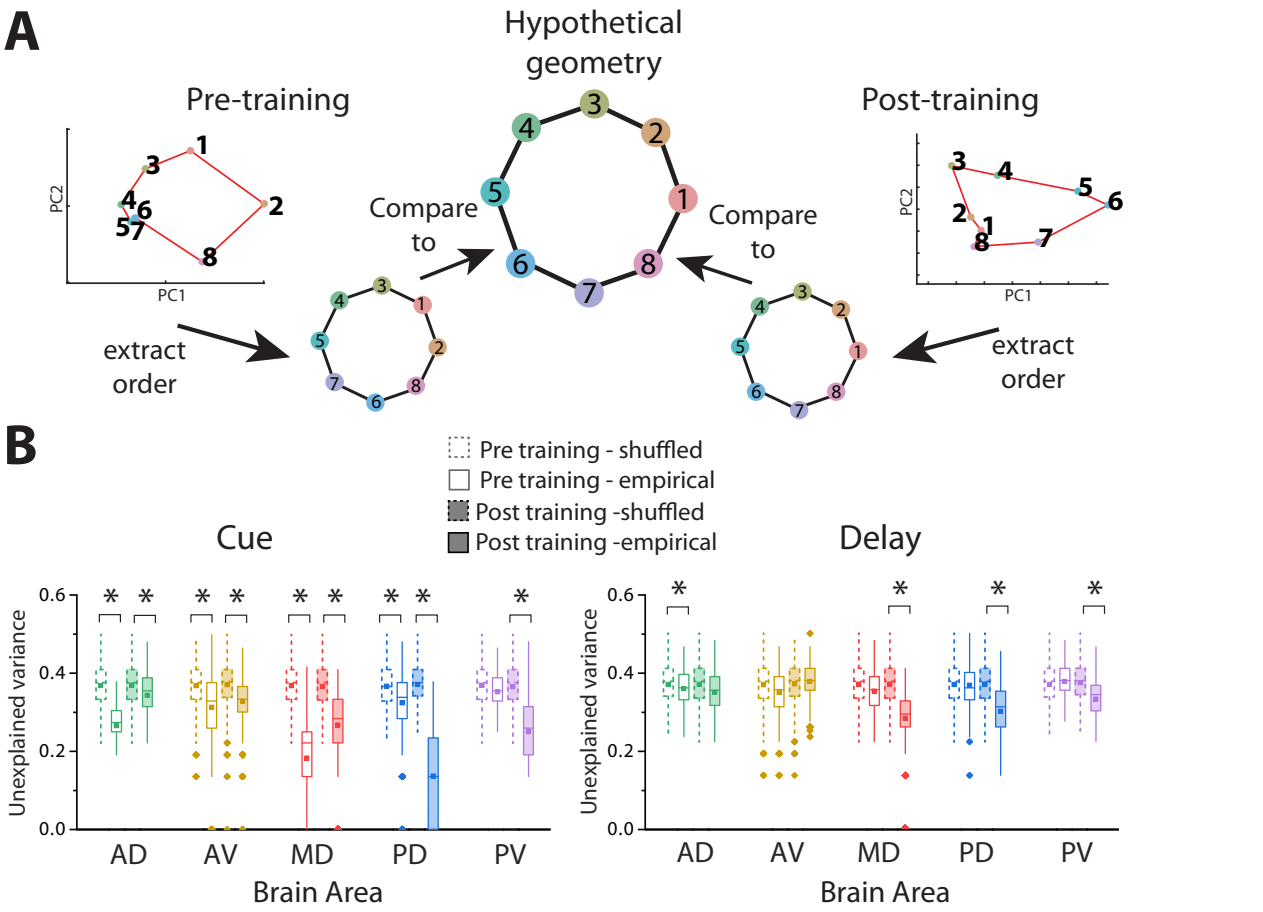


Fig. 2 | Comparison of within-subspace geometry for the spatial task. **A** Graphical illustration of how representational geometry order was compared. Geometries were calculated from cue period activity of the spatial task, pre- and post-training. The geometries were compared to a hypothetical geometry that maximizes the distance between match and nonmatch stimulus pairs (see Methods: Geometry of low-dimensional representation). **B** Training increased the orderly structure in the delay period of the spatial task. Geometrical similarity to the hypothetical configuration for both tasks in the cue and the first delay period was quantified by the unexplained variance after fitting to a standard geometry. To quantify the contribution of the factor of representation arrangement order (e.g. 1, 2, 3, 4 vs. 3, 2, 1, 4 on a ring) independent of the factor of geometrical shape (deviation from vertices of an octagon), the difference in unexplained variance between the empirical and an identity-label-shuffled dataset was compared. Color filled boxes represent post-training data; empty bars represent pre-training data. Solid line boxes represent empirical data and dashed line bars represent the similarity measure after shuffling location/shape labels, as a control (100 control and 100 empirical bootstrapped measurements were compared using a nonparametric shuffling test to derive statistics, n shuffle = 1000). Dot and horizontal line in the box plot represent the mean and median, respectively; The bottom and top edge of each box are the 25th and 75th percentiles of the sample. The whisker represents 1.5 times the interquartile range. A larger difference between the filled and the dashed boxes indicates a higher order similarity between the standard hypothetical geometry order and the empirical measurement. Test statistics for difference between this order similarity from the pre- and post- training phase in Table S3. Statistically significant differences detected by a two-sided nonparametric test are indicated by an asterisk (*) above the corresponding bars. Source data are provided as a Source Data file.

specific. In other words, rotation angles differed across regions, while similar angles were seen in the same subdivision before and after training (Fig. 3B1, B2, Fig. S3, left). Our analysis of plane rotations was replicated in higher-order subspaces as well (Fig. S1B, C), with qualitatively similar results, though for area PV, the rotation angle increased considerably post-training when we considered higher dimensions. Similar to the cue vs match trials, we found that substantial rotations between the cue and nonmatch representations already were evident in naïve animals (Fig. S4C), and the rotation angles remained quite consistent across different subdivisions and training status.

To further substantiate the subspace orthogonality observed in the angle measure, we calculated the variance accounted for (VAF) ratio between the cue and the match period (Fig. S5A, B, top panel). The VAF, with a value range from 0 to 1, is a measurement of subspace alignment. Higher VAF values indicate better alignment, while VAF values close to 0 indicate orthogonality (see Methods: variance accounted for ratio - VAF). The results in Fig. S5 qualitatively agree with the analysis of rotation angles (Figs. 3 and 4), with higher angle measurements consistently corresponding to lower VAF ratios. Once again, we found that the relative order of values across subdivisions was preserved between the pre- and post-training phases. This result indicates that populations of neurons in different subregions of the prefrontal cortex transform matching stimuli in a stereotypical fashion, independent of training and task execution.

We were also interested in how single-cell properties contributed to the rotation phenomenon observed at the population level. We examined repetition suppression, the phenomenon of decreased response to a stimulus that is repeated in a trial (match, in the context of our task) over a stimulus that is not repeated (nonmatch)[16]. This turned out not to have a substantial impact on population rotation measurements; removing the most suppressed cells still yielded results that were highly correlated with the original dataset (Fig. S6). Similarly, we investigated whether the number of cells contributing to the low-dimensional subspace that accounted for most variance

https://doi.org/10.1038/s41467-024-50717-y

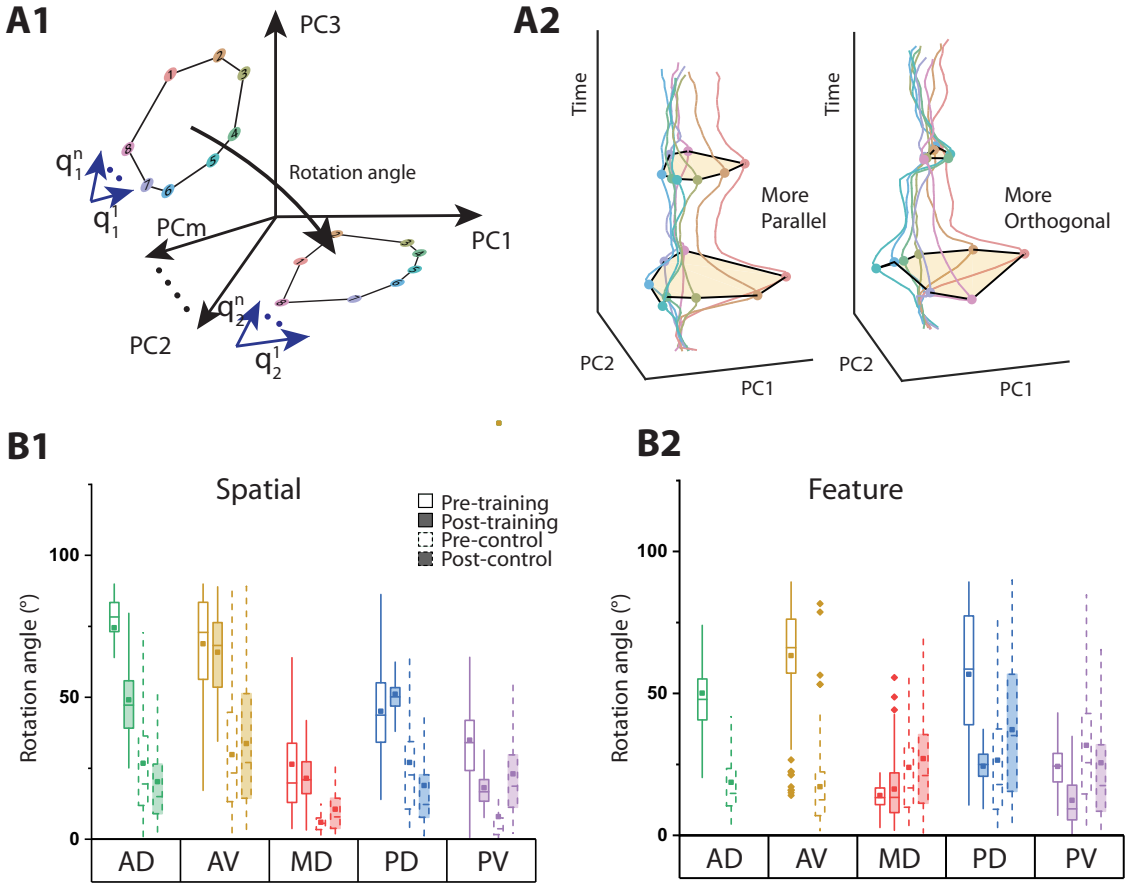


Fig. 3 | Subspace rotation between the cue and match epochs. A1 Illustration of subspace rotations. Neural responses of different conditions were projected onto a m- dimensional PCA-space ($m = 3$ in Figs. 3–5, m = 3-15 in Fig. S1), where the rotation and geometrical similarity were calculated (see Methods: Geometry of low-dimensional representation for details). (A2) Neural trajectory projected in the cue subspace when two representations were more parallel (MD cue vs. match) vs. more orthogonal (MD cue vs. nonmatch). When the second representation was parallel (left) to the first one, the projection on the first subspace was larger, compared to the condition of two representations being orthogonal (right). **B** Solid line bars: angles between the cue and the match condition in different PFC subdivisions pre- (empty boxes) and post-training (color-filled boxes) for the spatial (**B1**) and the feature task (**B2**). Dashed line bars (control): angles from the control data obtained by generating surrogate data (100 control and 100 empirical bootstrapped measurements were compared using nonparametric shuffling test to derive statistics, n shuffle=1000). Dot and horizontal line in the box plot represent the mean and median, respectively; The bottom and top edge of each box are the 25th and 75th percentiles of the sample, and the whisker represents 1.5 times the interquartile range. Pre vs post, two-sided nonparametric test: spatial AD $p = 0.277$; spatial AV $p = 0.652$; spatial MD $p = 0.932$; spatial PD $p = 0.901$; spatial PV $p = 0.122$; feature MD $p = 0.676$; feature PD $p = 0.480$; feature PV $p = 0.651$. All empirical angles were significantly larger than the corresponding control angle measurements ($p < 0.01$, to correct for multiple comparisons, Table S4), except for the post-training spatial and pre-training feature comparisons in the PV subdivision. Source data are provided as a Source Data file.

(participation ratio) changed across different contexts, or if the tuning of individual neurons to stimuli changed across different contexts. Our analysis showed that the former, difference in participation ratio in various contexts, better explained the rotation angle: areas with highest rotation angles typically exhibited the greatest difference in participation ratio (Fig. S7B, C top).

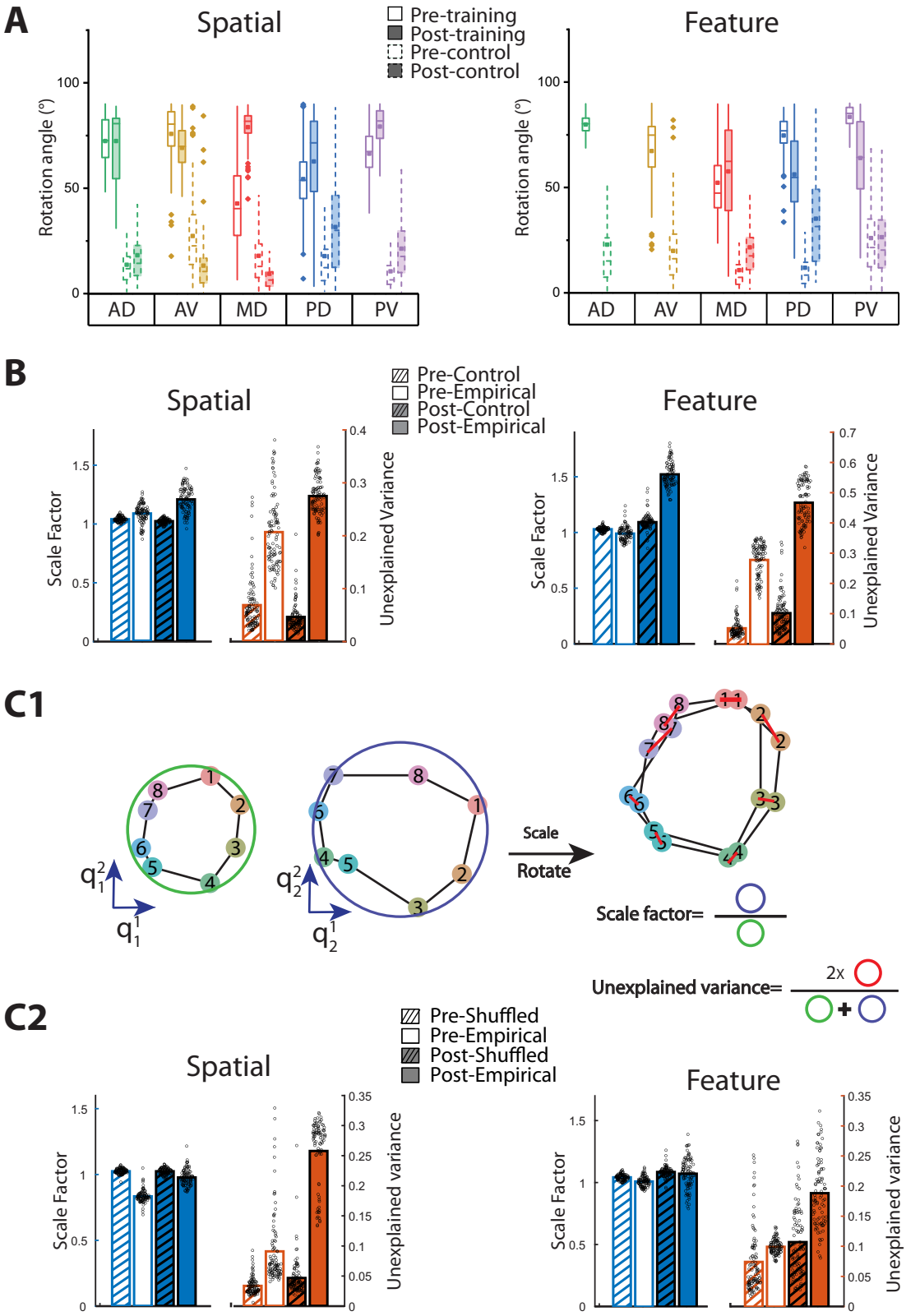
Although we have emphasized so far the task-independent nature of rotations, some representation transformations were only observed after training, for example involving the representation of stimuli in match and nonmatch conditions. In the spatial task, the angle of rotation between the match and nonmatch representations changed considerably after training (Fig. 4, left) in the most spatially selective, mid-dorsal region[17]. Consequently, across areas, there was little correlation between the plane angles observed before and after training across areas, Fig. S3 middle). The VAF ratio measurement agreed with angle measurement in this case too, evidenced by a major decrease for the mid-dorsal region in the spatial task (Fig. S5A middle). Similar training effects were observed for the subspace angle between the cue and the first delay period, with anterior prefrontal areas exhibiting increased rotation, particularly for the spatial task (Fig. S4A).

Task-dependent and independent representation geometry

In addition to subspace transformations, we also used two indexes to examine the within-subspace geometry in different contexts. The first index employed is the unexplained variance, which quantified geometrical structure similarity after controlling for differences in scale and orientation of representation in their respective subspaces. The second index we used was the scale factor, which intuitively represents the ratio of the areas covered by two geometries in respective subspace (i.e. the ratio between the blue and green circle in Fig. 4C1). A within epoch measure was used as a control (i.e., comparing the geometry of data from the same dataset to itself, see Methods: Dimensionality reduction and rotation of subspaces). For the unexplained variance, greater similarity between two conditions would result in measurements closer to 0, while for the scale factor, greater similarity between two conditions would result in values closer to 1.

In the geometry similarity analysis, we found that the effect of training for the most selective subdivision, the mid-dorsal PFC, was a decrease in similarity between epochs. This was true for the cue vs match comparison: the spatial unexplained variance increased from a

https://doi.org/10.1038/s41467-024-50717-y



pre-training value of 0.097 ± 0.06 (mean ± std), to a post-training value of 0.25 ± 0.06 while the feature pre-training unexplained variance similarly increased from a pre-training value of 0.11 ± 0.01, to a post-training value of 0.22 ± 0.06 (Fig. 4C2). The same relationship held between the cue stimulus and first delay period epochs: the spatial pre-training unexplained variance increased from 0.10 ± 0.003 to a post-training value of 0.22 ± 0.007 (Fig. S4B).

The representation scale measured across the cue and match epochs aggregated around 1 both before and after training (Fig. 4C, and Fig. S8A). This means that the stimulus representation generally spanned roughly the same amount of area in the respective subspace, either when compared between the cue and match, or between the match and nonmatch conditions (Fig. 4B and Fig. S8B), thus indicating a stable magnitude of population responses. For the cue vs. delay

Fig. 4 | Subspace rotation between the match and nonmatch conditions. **A** Primary angles between the match and nonmatch in different PFC subdivisions for the spatial and the feature task (pre vs post, two-sided nonparametric test: spatial AD $p = 0.24$; spatial AV $p = 0.384$; spatial MD $p < 0.001$; spatial PD $p = 0.822$; spatial PV $p = 0.071$; feature MD $p = 0.023$; feature PD $p = 0.254$; feature PV $p = 0.305$). The bottom and top edge of each box are the 25th and 75th percentiles of the sample, and the whisker represents 1.5 times the interquartile range. As a control we generated surrogate data from a particular task epoch (empirical: solid bars; control: dashed bars). In this comparison, the control angles were measured by fitting a Poisson distribution to the firing rates in the match trials. All empirical angles were significantly larger than the corresponding control angle measurements (two-sided nonparametric test, $p < 0.01$, to correct for multiple comparisons, Table S6) 100 control and 100 empirical bootstrapped measurements were compared using nonparametric shuffling test to derive statistics, n shuffle=1000. **B** Geometrical similarity between match and nonmatch in MD. Solid bars represent results from the match-nonmatch dataset, while hatched bars are results from the within-match period control group. Bar height indicates mean; overlaid black circles represent individual measurements from each bootstrap sampling. The difference of the two reflects the change in population geometry between two periods. Scale factor and unexplained variance measurement for other subdivisions are presented in Fig. S8B. Statistical test results of pre vs. post-training phase, in Table S7. **C1** Graphical summary of subspace geometrical similarity. The scale factor is the ratio between within-subspace to total variance. The unexplained variance quantifies geometrical structure similarity after controlling for scale and orientation differences in respective subspaces. **C2** Geometrical similarity in the MD subdivision. Solid bars represent results from the cue-match dataset; hatched bars are results from the within-cue period control group. Bar height indicates mean, overlaid black circles represent individual measurements from each bootstrap sampling. Difference between the two reflects the change in population geometry between two periods. Scale factor and unexplained variance measurement for other subdivisions presented in Fig. S8A. Number of bootstrap resampling=100 for **C1** and **C2**. Statistics test results of pre vs. post-training phase presented in Table S5. Source data are provided as a Source Data file.

comparison, the scale factor measure fell below 1 (Fig. S4B) partially due to the fact that even when accounting for persistent delay activity, the mean delay period firing rate was lower compared to the stimulus presentation epochs. The stability of the scale factor during the stimulus presentation periods implies that any changes in the correlation structure induced by training did not result in an overall increase of stimulus-driven firing rate deviation from the mean. Instead, the amplitude of stimulus representation remained relatively stable at the population level.

We also compared the dynamic trajectory of population activity before and after training. As expected, pre-training activity was largely confined to a narrow subspace during the delay period, while post-training activity became increasingly dynamic in an expanded space (Fig. S9A, B). To further quantify this change, we randomly sampled half of the trials from all periods and calculated PCA space as the base to project the remaining trials (see Methods: Dimensionality reduction and rotation of subspaces). The mean ratios of the decoding space area between fixation, first delay and second delay period (using the fixation period as reference) were 1: 1.07: 1.26 for the pre-training data and 1: 8.62: 1.83 for the post-training data (Fig. S9C–E).

Subspace rotation and dynamics correlates with behavior performance

To test whether the post-training population rotation in the match relative to the nonmatch condition was linked to improved task performance, we compared the subspace rotation angle in correct and error trials in a subset of neurons with sufficient error trials across conditions ($n = 295$ in the spatial working memory task, and $n = 201$ in the feature working memory task), pooling data across all areas (see Methods: Analysis of error trials). In correct trials (Fig. 5A, C), the low dimensional PC subspaces of match and nonmatch epochs were nearly orthogonal to each other (79.9 ± 5.4 and 78.0 ± 14.5 degrees for the spatial and feature stimuli, respectively), consistent with the post-training findings of the full population (Fig. 4). In error trials, this rotation was significantly reduced: Fig. 5B, D, 19.7 ± 9.0 and 18.8 ± 14.5 respectively (t-test for spatial: $t_{100} = 57.4$, $p < 0.0001$, feature: $t_{100} = 28.9$, $p < 0.0001$). To determine how population responses evolved across the length of the trial, we investigated representation dynamics for match and nonmatch trials in the reduced PCA space. Example trajectories for a single stimulus condition for the mid-dorsal area are plotted in Fig. 5E. These also resembled the trajectories in state space described previously for neuronal activity in animals trained to perform perceptual decision tasks[8]. In correct trials, as expected, match and nonmatch trajectories stayed close to each other until the sample period and then diverged during the sample presentation and the delay period that followed it. In error trials, an abnormal rotation was already present in the cue period and match and nonmatch trajectories were less-distinguishable (Fig. S10). Importantly, in error trials, the neural pattern we observed was not the reverse of that of correct trials. Instead, when the animal wrongly reported the matching status, the match and nonmatch trajectories were similar to each other but distinct from either the match or nonmatch condition in correct trials (Fig. S11). The result suggests that incomplete rotation of the stimulus subspace is more likely to result in errors. However, this aberrant trajectory emerged early in the trial and was not specifically dependent on incomplete orthogonalization of the nonmatch stimulus, which, as we showed here, was present even in naïve animals.

Discussion

Computation in neural activity is characterized by the transformation of representations across stages of processing[12]. Our results demonstrate that neuronal populations transform stimulus representations and exhibit dynamics shown to represent cognitive operations[5,8–10], even in the absence of task execution, in animals naïve to any such training. Our results in the same dataset from trained animals corroborated previous findings: in some divisions of the PFC, sequential cue and match stimuli from the same trials were represented in orthogonal planes; match and nonmatch stimuli exhibited large state space separations; and error trials exhibited smaller plane rotations than correct ones. Our results indicate that some dynamic transformations of population-based stimulus representations are not caused by task operations requiring explicit training.

Neural basis of cognitive flexibility

Natural plasticity allows for significant training-induced improvements of working memory, particularly through changes in PFC activity, over long periods of training[18–22]. Representation of stimuli in neural activity has also been shown to be dynamic at much faster timescales, during the time course of a single behavioral session[23] or, depending on trial events, on a moment-by-moment basis[1]. The transformation of stimulus representations has been shown to be indicative of flexible task execution, e.g. rotating representations may minimize interference between multiple stimuli, thus allowing additional stimuli to be encoded, or learning of associations between stimuli, otherwise presented passively[9]. Our current results suggest that the prefrontal cortex may perform these operations automatically, even in the absence of task execution, or learning novel stimuli. Furthermore, the scale of stimulus representation was largely unchanged after training. The idea of stable population information under shifting neural responses over days is an increasingly appreciated phenomenon in studies involving post-learning plasticity[24,25].

Regardless of transformations, data across prefrontal areas and task conditions were generally well fit by low dimensional

https://doi.org/10.1038/s41467-024-50717-y

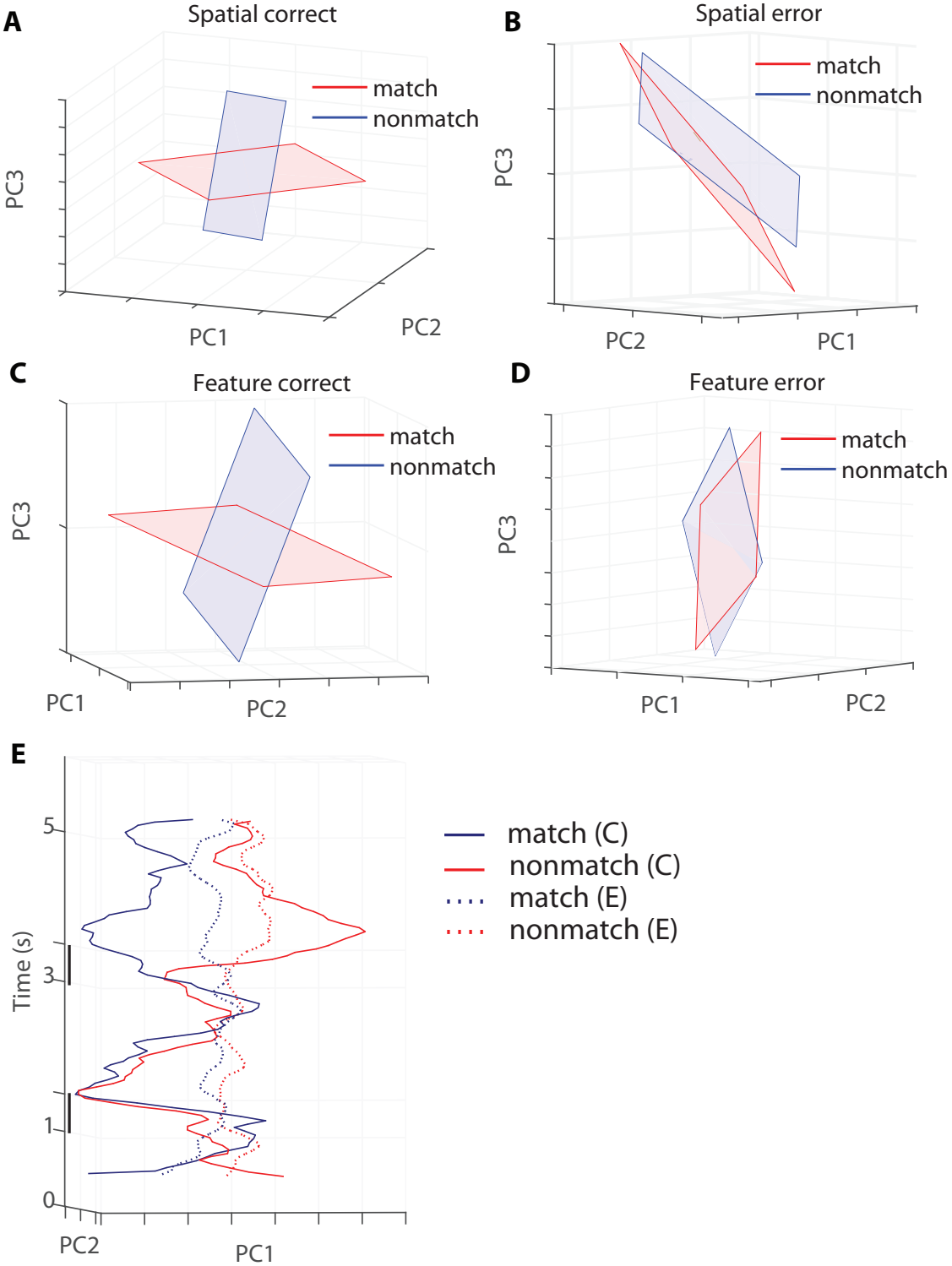


Fig. 5 | Population representations in correct and error trials. A Representation of spatial stimuli during the match and nonmatch period in correct trials ($n = 295$ neurons). Rotation $\varphi = 76°$. **B** Representation of spatial stimuli during the match and nonmatch period in the error trials (same population as in panel A). Rotation $\varphi = 15°$. **C** Representation of feature stimuli during the match and nonmatch period in correct trials ($n = 201$ neurons). Rotation $\varphi = 84°$. **D** Representation of feature stimuli during the match and nonmatch period in error trials (same population as in panel **C**). Rotation $\varphi = 7°$. **E** Dynamics of stimulus representations in correct and error trials, for one example spatial location. Solid blue: match in correct trials, solid red: nonmatch in correct trials, dashed blue: match in error trials, dashed red: nonmatch in error trials. The same cells are plotted across all conditions ($n = 295$ neurons). Source data are provided as a Source Data file.

representations. Low dimensionality in working memory representations is considered a hallmark of generalizability, while high dimensionality is correlated with better discriminability[26]. In our experiment, stimuli were highly discriminable from each other from the outset, and task training involved incorporating them into the context of a new task rather than learning fine features of the stimuli[27], suggesting that an increase in dimensionality was not required to perform the task after training.

Mechanisms and interpretations of rotations

The mechanisms of the population rotation phenomenon are still under debate. Sequential activation of individual neurons in the motor cortex can create an apparent rotation at the population level[28], although the rotation observed experimentally cannot fully be accounted by this phenomenon[29]. In the prefrontal cortex, neuronal activation is less sequential in nature compared to that of the motor cortex and the orthogonality we observed between contexts could be caused by the activation of largely nonoverlapping coding populations or a change of tuning to stimuli (Fig. S6). Similarly, the interpretation of the functional significance of this low-dimensional rotation, is multifaceted. Previous research has suggested that one of the functions of this rotation may be to reduce the interference between representations, and notably such rotations may appear as mice learn the structure of a stimulus set, without needing to perform an actual comparison based on working memory[9]. Our finding supports this conclusion by comparing rotations in response to the same stimuli before and after they were incorporated into a working memory task. The monkeys in our cohort were familiar with the stimulus sets and structure of the passive "task" (i.e. sequence of two stimulus presentation and their relative timing) by the time recordings in the "naïve" state began. It is therefore possible that the rotations emerged after the animals had discovered this regularity. Future studies may address this question. In any case, our results show that rotations in some PFC regions are task-independent and track information related to the temporal structure of the task, by low dimensional rotation, even when that information is not behaviorally relevant. This automatic tracking of variables changing in the natural world without explicitly prompt could be essential for learning.

Regional Specialization

Prior studies have debated if and how different types of information may be represented across the dorsal-ventral and anterior-posterior axes of the prefrontal cortex[30–33]. Converging evidence, however, suggests that anterior subdivisions of the prefrontal cortex generally represent more abstract information, and their activation depends to a greater extent on the task subjects have been trained to perform[34]. Consistent with this idea, we observed that anterior areas consistently exhibited high rotation values for match and nonmatch stimuli relative to the cue (and relative to each other), both before and after training. Our results raise the possibility that this innate stimulus transformation endows anterior areas of the prefrontal cortex with greater capacity for cognitive flexibility.

Methods

Data obtained from six male rhesus monkeys (*Macaca mulatta*), ages 5–9 years old, weighing 5–12 kg, were analyzed in this study. None of these animals had any prior experimentation experience at the onset of our study. Monkeys were either single-housed or pair-housed in communal rooms with sensory interactions with other monkeys. All experimental procedures followed guidelines set by the U.S. Public Health Service Policy on Humane Care and Use of Laboratory Animals and the National Research Council's Guide for the Care and Use of Laboratory Animals and were reviewed and approved by the Wake Forest University Institutional Animal Care and Use Committee under protocol numbers A06-033, A09-002, A14-196 and A17-139.

Monkeys sat with their heads fixed in a primate chair while viewing a monitor positioned 68 cm away from their eyes with dim ambient illumination and were required to fixate on a 0.2° white square appearing in the center of the screen. During each trial, the animals were required to maintain fixation while visual stimuli were presented either at a peripheral location or over the fovea, in order to receive a liquid reward (typically fruit juice). Any break of fixation immediately terminated the trial and no reward was given. Eye position was monitored throughout the trial using a non-invasive, infrared eye position scanning system (model RK-716; ISCAN, Burlington, MA). The system achieved a < 0.3° resolution around the center of vision. Eye position was sampled at 240 Hz, digitized and recorded. The visual stimulus display, monitoring of eye position, and synchronization of stimuli with neurophysiological data was performed with in-house software implemented in the MATLAB environment (Mathworks, Natick, MA), utilizing the Psychophysics Toolbox[35].

Pre-training task

Following a brief period of fixation training and acclimation to the stimuli, monkeys were required to fixate on a center position while stimuli were displayed on the screen. The stimuli shown in the pre-training, passive, spatial task consisted of white 2° squares, presented in one of nine possible locations arranged in a 3 × 3 grid with 10° of distance between adjacent stimuli. Only the eight peripheral locations are analyzed here, as the center location never appeared as a nonmatch. The stimuli shown in the pre-training passive feature task consisted of white 2° geometric shapes drawn from a set comprising a circle, a diamond, the letter H, the hashtag symbol, the plus sign, a square, a triangle, and an inverted Y-letter. These stimuli could be presented in one of the possible locations of the spatial grid.

The presentation began with a fixation interval of 1 s where only the fixation point was displayed, followed by a 500 ms stimulus presentation (referred to hereafter as cue), followed by a 1.5 s "delay" epoch where, again, only the fixation point was displayed. A second stimulus (sample) was subsequently shown for 500 ms. In the spatial task, this sample would be either identical in location to the initial stimulus, or diametrically opposite. In the feature task, the sample would appear in the same location as the cue and would either be an identical shape or the corresponding nonmatch shape (each shape was paired with one nonmatch shape). Only one nonmatch sample was paired with each possible cue, so that the number of match and nonmatch trials were balanced in each set. In both the spatial and feature task, this sample stimulus display was followed by another "delay" period of 1.5 s where only the fixation point was displayed. The location and identity of stimuli were of no behavioral relevance to the monkeys during the "pre-training" phase, as fixation was the only necessary action for obtaining rewards.

Post-training task

Three monkeys, the data of which are analyzed here, were trained to perform active working memory tasks that involved the presentation of identical stimuli as the spatial and feature tasks during the "pre-training" phase. Monkeys were required to remember the spatial location and/or shape of the first presented stimulus, and report whether the second stimulus was identical to the first or not, via saccading to one of two target stimuli (green for matching stimuli, blue for nonmatching). Each target stimulus could appear at one of two locations orthogonal to the cue/sample stimuli, pseudo-randomized in each trial.

Surgery and neurophysiology

The animals were initially implanted with a headpost device. Surgical anesthesia was induced with an intramuscular injection of ketamine (5 mg/kg) and maintained with inhalant isoflurane (1-3%). Opioid analgesics were administered after the surgery and the animals were allowed to recover for at least three weeks before behavioral sessions began. A second surgery was subsequently performed, under the same anesthetic and analgesic regimen, in which a 20 mm diameter craniotomy over the PFC was performed and a recording cylinder was implanted over the site. A 20 mm diameter craniotomy was performed over the PFC and a recording cylinder was implanted over the site. The location of the cylinder was visualized through anatomical magnetic resonance imaging (MRI) and stereotaxic coordinates post-surgery. For two of the four monkeys that were trained to complete active spatial and feature working memory tasks, the recording cylinder was

moved after an initial round of recordings in the post-training phase to sample an additional surface of the PFC.

Anatomical localization

Each monkey underwent an MRI scan prior to neurophysiological recordings. Electrode penetrations were mapped onto the cortical surface. We identified six lateral PFC regions: a posterior-dorsal region that included area 8 A, a mid-dorsal region that included areas 8B and 9/46, an anterior-dorsal region that included area 9 and area 46, a posterior-ventral region that included area 45, an anterior-ventral region that included area 47/12, and a frontopolar region that included area 10. However, the frontopolar region was not sampled sufficiently to be included in the present analyses.

Neuronal recordings

Neural recordings were carried out in the aforementioned areas of the PFC both before and after training in each WM task. Extracellular recordings were performed with multiple microelectrodes that were either glass- or epoxylite-coated tungsten, with a 100-250 µm diameter and 1–4 MΩ impedance at 1 kHz (Alpha-Omega Engineering, Nazareth, Israel). A Microdrive system (EPS drive, Alpha- Omega Engineering) advanced arrays of up to 8 microelectrodes, spaced 0.2–1.5 mm apart, through the dura and into the PFC. The signal from each electrode was amplified and band-pass filtered between 500 Hz and 8 kHz while being recorded with a modular data acquisition system (APM system, FHC, Bowdoin, ME). Waveforms that exceeded a user-defined threshold were sampled at 25 µs resolution, digitized, and stored for offline analysis. Neurons were sampled in an unbiased fashion, collecting data from all units isolated from our electrodes, with no regard to the response properties of the isolated neurons. A semi-automated cluster analysis relied on the KlustaKwik algorithm, which applied principal component analysis of the waveforms to sort recorded spike waveforms into separate units. To ensure a stable firing rate in the analyzed recordings, we identified recordings in which a significant effect of trial sequence was evident at the baseline firing rate (ANOVA, $p < 0.05$), e.g., due to a neuron disappearing or appearing during a run, as we were collecting data from multiple electrodes. Data from these sessions were truncated so that analysis was only performed on a range of trials with stable firing rates. Less than 10% of neurons were corrected in this way. Identical data collection procedures, recording equipment, and spike sorting algorithms were used before and after training in order to prevent any analytical confounds.

Statistics and reproducibility

Analysis involved measures of firing rate from neurons recorded before and after training. Due to the nature of the study, the Investigators were not blinded to allocation during experiments and outcome assessment. All available recordings were used for analysis, provided they had at least 12 trials at each stimulus condition (so that bootstrap analyses could be performed using subsets of trials from each neuron), and at least 4 but less than 400 spike events available between the cue onset and end of the trial when these were averaged across trials from all experiment conditions (to avoid outliers with either very low or very high firing rate and balance the pre- and post-training datasets in terms of firing rate range). A minimum of 16 available trials per condition was required for some analyses. No statistical method was used to predetermine sample size. Analysis methods relied on bootstrap tests and surrogate data (described in more detail below) to ensure that results and conclusions obtained were robust and reproducible.

Dimensionality reduction and rotation of subspaces

We applied principal components analysis (PCA) to visualize the neural population activity manifolds during the spatial and feature tasks. PCA was performed on the mean firing rate of neurons across different prefrontal regions, for data collected both before and after training. We examined the submanifolds of neural activity as a function of the spatial and feature stimulus sets.

In the spatial task, we collected data from five prefrontal regions before and after the monkeys were trained to perform the task. In each region, trials were collected when the cue stimuli appeared at $L = 8$ locations. For each neuron, the average firing rate of the match trials (and nonmatch trials) during a given period at the eight locations formed a column entry (8×1 vector) for the population activity matrix A_1 (A_2 for nonmatch trials). During each task epoch we defined the corresponding activity matrix A_1 (A_2) for each prefrontal region as an $8 \times N$ matrix, where 8 is the number of locations where the cue occurred and N is the number of neurons. To find the rotation angle between low-dimensional representations in two task epochs, we aligned the activity matrix into one single matrix B, which is a $16 \times N$ matrix with the first eight rows containing the activities in one period and the rest containing another period. Then we normalized B by subtracting the mean across each column to guarantee the matrix was zero-centered. PCA was applied to the centered data using singular value decomposition (SVD). We selected the first three eigenvectors of the covariance matrix of the zero-centered data and sorted them by decreasing order in terms of explaining the variance. The first three principal components explained an average of 73% of the response matrix variance across all examined periods and locations. Similar procedures were applied to visualize the responses in the feature task, where we grouped the data according to the shape of the cue stimulus. Specifically, there were eight different shapes in the feature task. Therefore, the corresponding activity matrix had a dimension of $8 \times N$, in which N is the number of neurons, and each row stands for a shape in the feature task (see Fig. 1B). For each pair of comparison for the rotations, we aligned the data in the same way as described above.

To calculate the subspace rotation between different contexts, in three or higher dimensions, we projected the trial averaged activity matrix into an N-dimensional PCA space (N = 3 in Figs. 2–5; N = 3-15 in Fig. S1). After projection, the population representations of the eight spatial locations (and eight shapes in the feature task) in a particular epoch (e.g. the cue period) formed eight points represented by an 8xN matrix. Next, we determined the 2D subspaces (2xN matrix) defined by the first two PCA axes of the 8xN matrix. We repeated that process for two conditions under comparison (e.g. cue and match), and defined the corresponding 2D subspaces as u and w. We performed singular value decomposition of the matrix uw^T, and got singular values a and b. The angle between two 2D planes under comparison was defined as arccos(min[a,b]). Geometrically, the procedure above calculates the principal angles between the pairs of the canonical basis of two subspaces. When N = 3, one of the angles is zero because, two sets of 2D subspaces necessarily share one axis. When we reduced the dimensionality to 3, a high proportion of variance was explained by the fitted plane, as demonstrated in Table S2 (e.g. for the PD area in the cue presentation epoch pre-training: 89.1% ± 3.3%; post-training PD 85.1% ± 3.6%). This validated our hypothesis that the reduced space could effectively be presented in 2D.

We defined the angles between different conditions based on the best-fit planes that we constructed above. Given the vectors spanning the best-fit plane (P_1) for one period are $\vec{v}_1$ and $\vec{v}_2$ (obtained from the PCs of the reduced eight points), and the best-fit plane (P_2) for another period in the same region are $\vec{v}_3$ and $\vec{v}_4$, the angle between P_1 and P_2 was calculated as follows:

$$\langle P_1, P_2 \rangle = \cos^{-1}\left(\frac{|(\vec{v}_1 \times \vec{v}_2) \cdot (\vec{v}_3 \times \vec{v}_4)|}{|\vec{v}_1 \times \vec{v}_2| \cdot |\vec{v}_3 \times \vec{v}_4|}\right) \quad (1)$$

where "$\langle P_1, P_2 \rangle$" denotes the angle between P_1 and P_2, $\cos^{-1}$ is the inverse cosine function, and "$\vec{v}_1 \times \vec{v}_2$" is the cross product that finds the vector that is perpendicular to the plane spanned by $\vec{v}_1$ and $\vec{v}_2$, the "$\cdot$"

sign stands for the dot product, and "|x|" returns the absolute value (for a scaler) or length (for a vector) of x.

To distinguish between an authentic large rotation of stimulus representation and an incidental rotation arising in a population of neurons with low selectivity and high variability, we employed two methods to create a control condition for each rotation-angle measurement. In the first method, we calculated the mean firing rate from n trials for each cell and each stimulus class across different task epochs. For example, we computed the mean firing rate during the cue period from the 16 trials of cell #1 located in the anterior dorsal area during the pre-training spatial task and stored this mean as λ. Subsequently, we generated a Poisson distribution with the same mean (λ) and randomly produced n sample values from it, simulating n trials from the cue period of this specific cell in the anterior dorsal area during the pre-training spatial task. This control dataset mirrored the firing rate and selectivity statistics of the empirical dataset, thus forming a baseline for rotation within each epoch, and providing a reference for angle and geometry measurements. Control conditions for rotation angle and geometry measurements are plotted with dashed boxes, whereas empirical data are indicated by solid frame boxes, as illustrated in Fig. 3B1, B2 and Fig. 4A. Only cells with at least 6 match and 6 nonmatch trials for each stimulus condition were included in the rotation angle analysis. We assessed variability by randomly drawing 80% of the cells from each prefrontal subdivision over 1000 iterations.

With the second method (Fig. S2), we randomly split all trials into two halves, and then, as the control, we calculate the rotation angle between the split trials of the same task condition. For instance, to find the control of rotation angles between the cue period and sample match in the pre-training spatial task, we randomly partitioned trials in all cells into two halves. The rotation angle between the two subspaces formed by the cue period from these halves was then calculated as one observation of the control for cue. A similar approach was applied to determine the control for the rotation angle of the sample match. We repeated the random division (into halves) 100 times to establish the distribution of the control angles for cue and sample match. This method halves the sample available, therefore we only used neurons with at least 16 trials per condition. As some areas did not have enough trials for comparison, we pooled data into three areas: anterior (comprising the AD and AV subregions), middle (comprising the MD subregion) and posterior (comprising the PD and PV subregions).

We implemented a nonparametric bootstrap method to assess the statistical significance of the differences between empirical angles at different training phases (pre vs. post), as well as between empirical angles and control rotation angles. First, the underlying true angle difference was calculated using the angle difference derived from all trials and cells. Then we conducted the bootstrap procedure. In each iteration, we shuffled the training status labels of all cells under comparison, followed by the recalculation the angle difference post-shuffle. This process was repeated 1000 times. The p-value for the observed difference was estimated based on the proportion of values in the shuffled data that exceeded true empirical value.

Variance accounted for ratio (VAF)

To cross-check the rotation angle between the 2D subspace of different task epochs, we measured the Variance Accounted For (VAF) ratio for each angle measurement[15]. The VAF ratio for epoch subspace pair (1,2) was defined as follows:

$$VAF_{1,2} = \frac{Var(v_2 v_2^T v_1 L_1^T)}{Var(v_1 L_1^T)} \tag{2}$$

Where v_i (i = 1,2) is a 3 × 2 matrix representing two subspace axes, and L_1 represents the stimulus projection in the first subspace. VAF ranges from 0 to 1; large values indicate better alignment of subspaces, while values close to 0 suggest orthogonality. Only cells with at least 6 match and 6 nonmatch trials for each stimulus category were included in the VAF analysis.

Single cell basis of coding subspace

To characterize the geometric relationship between single-cell axes and the low-dimensional subspace in certain contexts, we aimed to measure the alignment of single cells' selectivity with the low dimensional subspace. Geometrically, better alignment corresponds to a smaller angle between a vector and a subspace. This also implies that variance in the vector would yield significant variance within the subspace. Intuitively, a better-aligned vector will have a larger projection onto the subspace[15]. In accordance with this concept, we projected a unit vector i from the neuron under question onto the coding subspace of a specific context. The resultant projection vector onto the subspace is denoted by A, and the angle between A and a specific stimulus vector (we chose the first location/shape) is φ, so A and φ satisfy:

$$(q^1)^T i = A\cos(\varphi) \tag{3}$$

$$(q^2)^T i = A\sin(\varphi) \tag{4}$$

in which q^1 and q^2 represent the first two PCs of the subspace. Thus, A measures the degree of alignment, or the strength of contribution from the cell, and φ indicates the turning direction of the cell in the coding subspace. We also employed a quantity called normalized participation ratio (PR)[15] to quantify how distributed the subspace is across the population, as follows:

$$PR = \frac{\left(\sum_{i=1}^{N} A^2\right)^2}{N\sum_{i=1}^{N} A^4} \tag{5}$$

This quantity ranges from 0 to 1, with 0 indicating very sparse coding and 1 indicating evenly distributed coding across the whole population.

To examine the contribution of repetition suppression on the observed subspace rotation, cells with prominent repetition suppression were removed from the database, and the rotation angle was recalculated. Specifically, the top 10% of cells with the largest reduction in firing rate from the cue to the match epoch was excluded from this analysis.

Geometrical order in low-dimensional representation. A matrix L of size 8 × 2 was used to represent the projection of 8 locations or shapes on the 2D plane, with each row representing a stimulus in the 2D PCA space. To compare the geometric order with a hypothetical order on an octagon (Fig. 2, Fig. S12), we first projected the 8 stimuli onto a circle. Specifically, we aligned the 8 points uniformly on an imaginary circle by equalizing the arrangement as follows

$$D = S\begin{bmatrix} 1 & 0 \\ 0 & \frac{\max(S_{(:,1)}) - \min(S_{(:,2)})}{\max(S_{(:,2)}) - \min(S_{(:,2)})} \end{bmatrix} \tag{6}$$

$$T = CD^{-1} \tag{7}$$

Here, S represents the score matrix and C the coefficient matrix, both obtained from the PCA of L. Here, $S_{(:,i)}$ is the ith column of S, and T is the new matrix representing the 8 stimuli in the reduced space. Subsequently, we determined the circular order of the stimuli by computing the angles of vectors from the center to each point, with the angle for the ith stimulus defined as $\theta_i = \text{arccot}(I)$, where I denotes the vector from the center to the ith point.

Focusing solely on the sequential order, we aligned the stimuli to the vertices of an ideal octagon in the order obtained from the circular projection. This realignment is defined as $N(o_{rank},:) = M(o_{hypothetical},:)$, where o_{rank} and $o_{hypothetical}$ are vectors containing the measured stimulus order in the low dimensional space and the hypothetical order, respectively. M contains coordinates of a unit-size octagon, and N is a low-dimensional representation of the 8 stimuli, reshaped to a standard octagon.

We quantified "order regularity" by the unexplained variance determined by the Kabsch algorithm[36], which finds the optimal rotation matrix R that minimizes the root mean squared deviation between M and N. The Kabsch algorithm was implemented as follows

$$R = V\begin{bmatrix} 1 & 0 \\ 0 & d \end{bmatrix}U^T \tag{8}$$

$$d = sign\left(\det\left(VU^T\right)\right) \tag{9}$$

where U and V come from the singular value decomposition of the covariance matrix of M and N. Unexplained variance between two set of coordinates, referred to as C_{first} and C_{second}, is defined as

$$\text{Unexplained variance} = \frac{2 \times \sqrt{\sum\sum (C_{first} - C_{second})^2}}{\sqrt{\sum_p \sum_q C_{first}^2(p,q)} + \sqrt{\sum_p \sum_q C_{second}^2(p,q)}} \tag{10}$$

A lower value of unexplained variance indicates a higher congruence, signifying greater "order regularity."

When comparing the empirical geometry with a hypothetical geometry (Fig. 2, Fig. S12), we hypothesized that the distance between each cue and its nonmatch pair should be maximized. For the spatial task, the match-nonmatch structure is symmetric (e.g. when A is cue, B is nonmatch; and when B is cue A is nonmatch). These two assumptions should lead to a predicted arrangement similar to the physical appearance of the spatial stimuli, where the nonmatch pair always perfectly falls on two ends of a diagonal line.

For geometrical comparisons between two task epochs (Fig. 4), the geometries from the same neural population were used, considering both the order and the location variation. To quantify the amplitude of stimuli representations across two conditions, we used a scale factor defined as the variance ratio between two geometrical arrangements, calculated as follows

$$Scale\ factor = \frac{\sqrt{\sum_p \sum_q C_{second}^2(p,q)}}{\sqrt{\sum_p \sum_q C_{first}^2(p,q)}} \tag{11}$$

$$C_{(p,q)} = L_{(p,q)} - \bar{L} \tag{12}$$

Where C_{first} and C_{second} represent mean-subtracted projections on a 2D plane of the first and second condition in a comparison, respectively, and where p=1, …,8 and q = 1,2, represent two coordinates for eight shapes in a 2D subspace. Only cells with at least 6 match and 6 nonmatch trials for each stimulus category were included in the geometrical similarity analysis. Standard error was measured by random drawing 80% of cells in each subdivision over 100 iterations.

Time course of subspace rotation

In addition to visualizing the rotation between different phases of the WM task, we investigated the dynamics of presentations in the reduced PCA space. To construct the trajectory, we discretized the whole task with a timestep of $\Delta_t = 50ms$ and an interval of $l_t = 250ms$. For each timestep k, we recorded the mean firing rates for the match and nonmatch trials in the interval $[t_0 + (k-1)\Delta_t - 250, t_0 + (k-1)\Delta_t]$ as the entries for the population activity matrix A_1 and A_2 (defined in the above section). Similar as the procedure of constructing the PCA space, we projected the neural responses for each condition into their first three principal components. To plot the trajectories, we started from 250 milliseconds after the start of fixation, i.e., set the initial time $t_0 = 250ms$. Consequently, the first interval we considered spans from 0 to 250 milliseconds in fixation. We also used the first three PCAs in the match trials during the cue period as the common basis to calculate the coordinates. In other words, we projected the activity matrix into the same common subspace spanned by the PCA space of the match trials. As illustrated in Fig. 3A2, we explored the dynamics of representation for each location (shape) in the spatial (feature) task, where we plot every six points from the dynamics, i.e., we set an increment of $dt = 300ms$ for the purpose of visualization.

To further investigate the changes in decoding spaces, we randomly sampled half of the trials from all periods (fixation, cue, delay1, sample and delay2) and calculated its PCA space as the base to project the remaining half trials. The area of the projection for a given period indicates how much information the state space contains from all eight locations, which we defined as the decoding subspace (S_t). For example, in the reduced space formed by PC1 and PC2, the eight spatial task locations were presented by eight points. To quantify the size of our decoding subspace, we first found the boundaries of the eight points, which was a set of points representing a single conforming to a 2-D boundary around the eight points. Then we calculated the areas of the polygons defined by the boundary points, which we defined as the areas of the decoding space (A_t). We calculated the ratios of the areas of decoding spaces based on the fixation period (A_{fix}). For example, the ratios (R) between fixation, delay1 and delay2 were defined as follows:

$$R = \frac{A_{fix}}{A_{fix}} : \frac{A_{delay1}}{A_{fix}} : \frac{A_{delay2}}{A_{fix}} \tag{13}$$

Note that we divided each area by that computed at the fixation period to normalize the data so that we could compare the ratios from different samples. We repeated the random sampling across 10,000 iterations, and the statistics of the ratios were reported in the results section.

Significance testing of rotations with bootstrap

Our previous analysis was based on all data recorded from neurons that had more than 16 trials for all cue conditions and in each region in Fig. 1C. To test the significance and robustness of the representation rotations, we applied a bootstrap method to sample 100 iterations from the original data. On each iteration, 80% of the cells were drawn with replacement. The angles were calculated following the same procedure as above. Mean and median angles from 100 iterations obtained from bootstrap were represented by dots and horizontal bars within the box plots, while the interquartile range and five percent extreme values were indicated by box border and whiskers, respectively.

Analysis of error trials

In the post-training spatial task, most cells did not have error trials for all conditions. Therefore, the corresponding activity matrix (as defined above) for the error trials was very sparse. In the spatial task, there were 295 neurons where error trials presented at more than half of the eight locations, i.e., 295 rows of the population activity matrix had more than four non-empty entries. We selected those 295 neurons for the error trial analysis in the spatial task. Following the same criterion, we selected 201 neurons containing error trials in the feature task. We first aligned the activity matrix for both cases in the same order, i.e., the same neuron occupied the same row in both matrices. In most

cases, for a given stimulus, there were more correct trials than error trials, which led to a larger number of samples in most entries of the activity matrix. Each entry of the activity matrix was calculated from the mean activity across all trials. We thus matched the number of trials in each entry to allow comparison between correct and error trials. Specifically, for each empty entry in the error matrix, we removed the corresponding entry in the correct matrix. Additionally, we selected the same number (minimum of the corresponding entries) of trials in both cases to ensure the same number of samples in both the correct and error trials.

Dimensionality reduction was performed on the error trials using PCA. Given that there are several possible versions of PCA for a sparse matrix, we relied on a probabilistic principal component (PPCA) with variational Bayesian learning[37] as this method worked best for our neural data.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data supporting the individual figures are provided in the Source Data Files. The full data set used for analysis has been made available at: https://codeocean.com/capsule/2817512/tree/v2. Source data are provided with this paper.

Code availability

Analysis code for the current project is available at: https://codeocean.com/capsule/2817512/tree/v2.

References

1. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
2. Rainer, G., Asaad, W. F. & Miller, E. K. Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature* **393**, 577–579 (1998).
3. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
4. Ebitz, R. B. & Hayden, B. Y. The population doctrine in cognitive neuroscience. *Neuron*, (2021).
5. Cueva, C. J. et al. Low-dimensional dynamics for working memory and time encoding. *Proc. Natl Acad. Sci. USA* **117**, 23021–23032 (2020).
6. Murray, J. D. et al. Stable population coding for working memory coexists with heterogeneous neural dynamics in prefrontal cortex. *Proc. Natl Acad. Sci. USA* **114**, 394–399 (2017).
7. Tang C., Herikstad R., Parthasarathy A., Libedinsky C. & Yen S. C. Minimally dependent activity subspaces for working memory and motor preparation in the lateral prefrontal cortex. *Elife* **9**, (2020).
8. Okazawa, G., Hatch, C. E., Mancoo, A., Machens, C. K. & Kiani, R. Representational geometry of perceptual decisions in the monkey parietal cortex. *Cell* **184**, 3748–3761 e3718 (2021).
9. Libby, A. & Buschman, T. J. Rotational dynamics reduce interference between sensory and memory representations. *Nat. Neurosci.* **24**, 715–726 (2021).
10. Minxha J., Adolphs R., Fusi S., Mamelak A. N. & Rutishauser U. Flexible recruitment of memory-based choice representations by the human medial frontal cortex. *Science* **368**, (2020).
11. Panichello, M. F. & Buschman, T. J. Shared mechanisms underlie the control of working memory and attention. *Nature* **592**, 601–605 (2021).
12. Kriegeskorte, N. & Wei, X. X. Neural tuning and representational geometry. *Nat. Rev. Neurosci.* **22**, 703–718 (2021).
13. Meyer, T., Qi, X. L., Stanford, T. R. & Constantinidis, C. Stimulus selectivity in dorsal and ventral prefrontal cortex after training in working memory tasks. *J. Neurosci.* **31**, 6266–6276 (2011).
14. Riley, M. R., Qi, X. L., Zhou, X. & Constantinidis, C. Anterior-posterior gradient of plasticity in primate prefrontal cortex. *Nat. Commun.* **9**, 3790 (2018).
15. Xie, Y. et al. Geometry of sequence working memory in macaque prefrontal cortex. *Science* **375**, 632–639 (2022).
16. Grill-Spector, K., Henson, R. & Martin, A. Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn. Sci.* **10**, 14–23 (2006).
17. Dang, W., Jaffe, R. J., Qi, X. L. & Constantinidis, C. Emergence of Nonlinear Mixed Selectivity in Prefrontal Cortex after Training. *J. Neurosci.* **41**, 7420–7434 (2021).
18. Dux, P. E. et al. Training improves multitasking performance by increasing the speed of information processing in human prefrontal cortex. *Neuron* **63**, 127–138 (2009).
19. Klingberg, T. et al. Computerized Training of Working Memory in Children with ADHD - a Randomized, Controlled Trial. *J. Am. Acad. Child Adolesc. Psychiatry* **44**, 177–186 (2005).
20. Jaeggi, S. M., Buschkuehl, M., Jonides, J. & Perrig, W. J. Improving fluid intelligence with training on working memory. *ProcNatlAcadSciUSA* **105**, 6829–6833 (2008).
21. Constantinidis, C. & Klingberg, T. The neuroscience of working memory capacity and training. *Nat. Rev. Neurosci.* **17**, 438–449 (2016).
22. Qi, X. L. & Constantinidis, C. Neural changes after training to perform cognitive tasks. *Behav. Brain Res* **241**, 235–243 (2013).
23. Asaad, W. F., Rainer, G. & Miller, E. K. Neural activity in the primate prefrontal cortex during associative learning. *Neuron* **21**, 1399–1407 (1998).
24. Rule M. E. et al. Stable task information from an unstable neural population. *Elife* **9**, (2020).
25. Rule M. E. & O'Leary T. Self-healing codes: How stable neural populations can track continually reconfiguring neural representations. *Proc. Natl. Acad. Sci. USA* **119**, (2022).
26. Bernardi, S. et al. The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell* **183**, 954–967.e921 (2020).
27. Meyers, E. M., Qi, X. L. & Constantinidis, C. Incorporation of new information into prefrontal cortical activity after learning working memory tasks. *Proc. Natl Acad. Sci. USA* **109**, 4651–4656 (2012).
28. Lebedev, M. A. et al. Analysis of neuronal ensemble activity reveals the pitfalls and shortcomings of rotation dynamics. *Sci. Rep.* **9**, 18978 (2019).
29. Elsayed, G. F. & Cunningham, J. P. Structure in neural population recordings: an expected byproduct of simpler phenomena? *Nat. Neurosci.* **20**, 1310–1318 (2017).
30. Wilson, F. A., Scalaidhe, S. P. & Goldman-Rakic, P. S. Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science* **260**, 1955–1958 (1993).
31. Badre, D. & D'Esposito, M. Is the rostro-caudal axis of the frontal lobe hierarchical? *Nat. Rev. Neurosci.* **10**, 659–669 (2009).
32. Rao, S. C., Rainer, G. & Miller, E. K. Integration of what and where in the primate prefrontal cortex. *Science* **276**, 821–824 (1997).
33. Owen, A. M. et al. Functional organization of spatial and nonspatial working memory processing within the human lateral frontal cortex. *Proc. Natl Acad. Sci. USA* **95**, 7721–7726 (1998).
34. Constantinidis, C. & Qi, X. L. Representation of Spatial and Feature Information in the Monkey Dorsal and Ventral Prefrontal Cortex. *Front Integr. Neurosci.* **12**, 31 (2018).
35. Meyer, T. & Constantinidis, C. A software solution for the control of visual behavioral experimentation. *J. Neurosci. Methods* **142**, 27–34 (2005).
36. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr* **A32**, 922 (1976).

37. Ilin A. & Raiko T. Practical Approaches to Principal Component Analysis in the Presence of Missing Values. *Journal of Machine Learning Research*, 1957-2000 (2010).

Acknowledgements

Research reported in this paper was supported by the National Eye Institute of the National Institutes of Health under award number R01 EY017077 to CC. We wish to thank Rye Jaffe, Junda Zhu, and Zhengyang Wang for helpful comments on the manuscript.

Author contributions

C.C. conceived and designed the experiments. C.C. and X. L. Q. performed experiments. S.P, W.D., and C.C. performed data analysis. C.C., S. P. and W.D. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interest.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-50717-y.

Correspondence and requests for materials should be addressed to Christos Constantinidis.

Peer review information *Nature Communications* thanks Tatiana Engel, Liping Wang and the other, anonymous, reviewer for their contribution to the peer review of this work. A peer review file is available.