ARTICLES
https://doi.org/10.1038/s41593-019-0400-9

nature neuroscience

Effective learning is accompanied by high-dimensional and efficient representations of neural activity

Evelyn Tang[1,2], Marcelo G. Mattar[3], Chad Giusti[1,4], David M. Lydon-Staley[1], Sharon L. Thompson-Schill[3] and Danielle S. Bassett[1,5,6,7,8]*

A fundamental cognitive process is to map value and identity onto the objects we learn about. However, what space best embeds this mapping is not completely understood. Here we develop tools to quantify the space and organization of such a mapping in neural responses as reflected in functional MRI, to show that quick learners have a higher dimensional representation than slow learners, and hence more easily distinguishable whole-brain responses to objects of different value. Furthermore, we find that quick learners display more compact embedding of their neural responses, and hence have higher ratios of their stimuli dimension to their embedding dimension, which is consistent with greater efficiency of cognitive coding. Lastly, we investigate the neurophysiological drivers at smaller scales and study the complementary distinguishability of whole-brain responses. Our results demonstrate a spatial organization of neural responses characteristic of learning and offer geometric measures applicable to identifying efficient coding in higher-order cognitive processes.

Essential to human cognition is the ability to group stimuli into meaningful identities. The emergence of such identities is accompanied by the development of a mapping that is encoded in the activity patterns of neural circuitry[1]. Exactly how new information about objects is mapped into the correct groups, such that relevant information becomes associated, is not completely understood. Furthermore, it is not generally known how far apart such groups should be and what kind of space efficiently embeds such a mapping. These concepts and questions are reminiscent of studies of coding efficiency in neural responses to low-level sensory stimuli[2,3]—a notion quantifying a system's information processing given biophysical and metabolic constraints. An open question is whether similar principles of efficiency play a role in higher-level processes such as cognition[4]. What goals and constraints must be balanced to enable such cognitive coding efficiency[5–7]? And how might such efficiency support accurate perceptions and decisions?

To formalize intuitive notions of space and organization in neural activity during the building of such mental maps, we use a geometric perspective adapted from machine learning[8,9]. Specifically, we represent distributed neural responses as points in a multidimensional space. Applied to neuron-level data, such representations have been shown to be very effective in isolating an intrinsic low-dimensional subspace relevant to ongoing cognitive processes[8,10,11]. In this study, we extend these tools to the examination of large-scale neural responses in humans as they integrate information across many areas to form representations of novel objects, appreciate the abstract properties of those objects[12], and both prepare and execute associated motor responses[9]. Despite our growing understanding of the regions activated by such learning[12], a notable gap in knowledge lies in delineating how spatial patterns of neural responses in these activated regions allow for effective behavioral choices. Our approach complements multivoxel pattern analysis and related techniques, which enable a local quantification of regional representations of objects or concepts[9], by offering tools that synthesize information across all brain regions simultaneously.

Fundamentally, these tools allow us to hypothesize that the dimension of a geometric representation of neural responses is related to the effective identification of stimuli and corresponding learned values. The simple intuition behind this hypothesis is that a higher dimension allows for an easier grouping of neural responses according to different objects in the geometric space. To test this hypothesis, we examine blood oxygen level-dependent (BOLD) magnitudes at the regional and voxel levels, in a cohort of 20 healthy adult humans as they learn the values of 12 novel objects over the course of 4 consecutive days for a total of 80 experimental imaging sessions. Motivated by a desire to study parsimonious representations and also by recent work decoding object identity[13], stimulus response[14] and markers of emotional and affective processing[15] from coarse-scale measurements across the brain, we spatially average these indirect measurements of neural activity in 83 regions of interest (ROIs) defined by a whole-brain anatomical parcellation. Next, we use a generalized linear model (GLM) to deconvolve the hemodynamic response function to obtain approximate neural responses to each stimulus at the time point when it was presented. We ask how the dimension of the geometric representation of such neural responses reflects the speed with which participants learn the objects' monetary values[16]. To answer this question, we study three aspects of the geometric organization of these neural

[1]Department of Bioengineering, School of Engineering & Applied Science, University of Pennsylvania, Philadelphia, PA, USA. [2]Department of Living Matter Physics, Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany. [3]Department of Psychology, College of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA. [4]Department of Mathematical Sciences, College of Arts and Sciences, University of Delaware, Newark, DE, USA. [5]Department of Electrical & Systems Engineering, School of Engineering & Applied Science, University of Pennsylvania, Philadelphia, PA, USA. [6]Department of Physics & Astronomy, College of Arts & Sciences, University of Pennsylvania, Philadelphia, PA, USA. [7]Department of Neurology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [8]Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. *e-mail: dsb@seas.upenn.edu

responses: the stimulus dimension; the embedding dimension; and label assortativity.

We demonstrate that fast learners have higher dimensional stimulus representations, allowing for an easier development of boundaries between neural responses to different stimuli. However, a potential disadvantage of using a high-dimensional representation is that the brain might use more resources to embed the information. To assess the presence or absence of this potential trade-off, we study the embedding dimension: the geometric representation of each individual's neural responses, with the map between stimulus and neural response shuffled uniformly at random. We find that the embedding dimension of a fast learner is more compact than that of a slow learner, suggesting that their neural responses form a more contained underlying subspace within the higher dimensional ROI space. The large ratio between the stimulus and the embedding dimensions is indicative of efficient coding and is observed most commonly in individuals who learned rapidly. To enhance our understanding of the anatomy driving these observations, we identify brain regions that most contribute to the emergence of high-dimensional patterns in quick learners; we further implement a voxel-level analysis to examine a finer-scale structure in neural responses. Lastly, we use the complementary metric of label assortativity to characterize the distinguishability of neural responses. Confirming our prior results, we find that fast learners have more distinguishable neural responses than slow learners. Taken together, our approach provides an insight into the geometry of neural responses supporting learning and offers a suite of computational heuristics to intuitively describe cognitive processes more generally.

Results

Quick learners develop higher dimensional stimulus representations of neural responses. We sought to understand how the neural responses of individuals are distributed according to task-relevant stimuli and how this distribution reflects their learning ability. The dimensionality of the functional magnetic resonance imaging (fMRI) BOLD evoked responses (Fig. 1a,b) can be estimated based on the performance of a linear classifier in distinguishing assigned binary labels on the data[8]. Intuitively, a given spatial arrangement of these responses will make it easier for any of the stimuli to be distinguished from the others, when the data are arranged in a higher dimensional manner. Specifically, for n stimuli there are 2^n ways to assign binary labels to these stimuli. When the data are arranged in a low-dimensional manner, some binary assignments will result in poorer separability, whereas in a higher dimension, these binary assignments will result in higher separability on average (see Fig. 1c). By exhaustively examining all 2^n-2 choices of binary labeling and by recording the resulting separability, the average performance over this combinatorial number of assignments yields the separability dimension of stimulus representation.

We applied this method and assortativity (see Fig. 1d) to the evoked neural responses of participants learning the value of 12 shapes (see Fig. 2a; ref. [16]). Each shape was assigned a distinct mean monetary value. Learning phases and a value judgment task were repeated daily for 4 d (see Fig. 2b). As the sessions progressed, participants improved in their abilities to select the shape with the higher expected value. By the conclusion of the second day of practice, all participants reached a generally high level of performance (see Fig. 2b). During the learning phase, participants were shown a pair of shapes simultaneously and asked to select which shape had the higher value, after which they received feedback based on their response (see Fig. 2c). This portion of the experiment was followed by a value judgment task, where participants were shown individual shapes and asked to indicate if the shape was one of the six least or six most valuable shapes (see Fig. 2d).

We sought to explore how the geometric representation of each individual's neural responses is related to their learning effectiveness. To obtain this geometric representation, we used data from the value judgment task when shapes were presented one at a time and we applied a GLM to obtain the neural response to each shape (see Fig. 1b) for every ROI. For each shape, the neural responses across all regions contribute a point in the ROI space. Hence, 140 shape presentations in one session jointly form a point cloud or geometric representation (see Fig. 3a), whose dimension we quantified. However, for $n=12$, calculating the 2^n-2 binary assignments is computationally expensive. Thus, in practice we chose a subset of $m=4$ stimuli over which to calculate this separability dimension. To ensure that our results did not depend on the particular subset of stimuli chosen, we repeated the calculation on 20 different combinations (roughly 7%) out of the $\binom{n}{m}$ available choices, while ensuring that each shape was represented a roughly equal number of times throughout these sets. As a measure of an individual's learning effectiveness, we used their response accuracy from each of these daily value judgment sessions.

We found that individuals with a higher than average dimension of their neural representation also exhibit higher than average learning accuracies, using a multilevel model to analyze the behavior and neural data across all days ($P<0.043$; see Methods and Supplementary Table 1). To better understand this effect, we assessed the correlation among behavioral accuracy scores across participants, averaged across days (mean $r=0.58$, s.d. of the mean $=0.03$). This high correlation indicates that individuals who display higher accuracy than average on the first day, also tend to display higher accuracy than average on the other 3 d. We also assessed the correlation in separability dimension across participants, averaged across days, and found that they were highly variable (mean $r=-0.02$, s.d. of the mean $=0.10$). This low correlation suggests that values of separability dimension varied appreciably over time within an individual during the learning experiment. Hence, we next treated the behavioral accuracy scores as repeated measures and sought to understand whether they could better predict an individual's separability dimension earlier versus later in the experiment. We found evidence for an approximately linear increase in the correlation between an individual's behavioral accuracy and their separability dimension each day, as we moved from day 1 to day 4 (see Supplementary Fig. 1). This result suggests that the neural representations on day 4 most strongly reflect the effects of learning.

To simplify the presentation of the results, we used a single day's behavioral accuracy as the predictor for the separability dimension on day 4. We chose day 1 for two reasons. First, we theoretically hypothesized that the effectiveness of early learning would predict the crystallization of neural representations in the future. Second, the behavioral accuracy scores on day 1 were characterized by the greatest variance to skewness ratio, indicating significant and homogeneous variability across participants (furthest from the performance ceiling)—unlike later days when behavioral accuracy became more skewed, tending to be high and similar in value across most participants with a few outliers displaying low accuracy. With this simplification, we found that the response accuracy of participants on the first day was significantly correlated with their separability dimension on the last day (Pearson's correlation, $r=0.56$ and $P<0.001$; see Fig. 3b). The statistical significance of this relationship was assessed using a null model permuting the object labels of the neural responses uniformly at random and calculating the separability dimension of these permuted data. Across participants, we calculated the correlation between their response accuracy and the dimension of these null data in 1,000 bootstrapped samples (gold bars in Fig. 3c). We observed that the true stimulus-based data fell significantly outside this distribution, with non-parametric $P<0.001$. This finding suggests that participants who learn more quickly display a larger stimulus separability dimension of their representations, which allows for easier distinguishability between

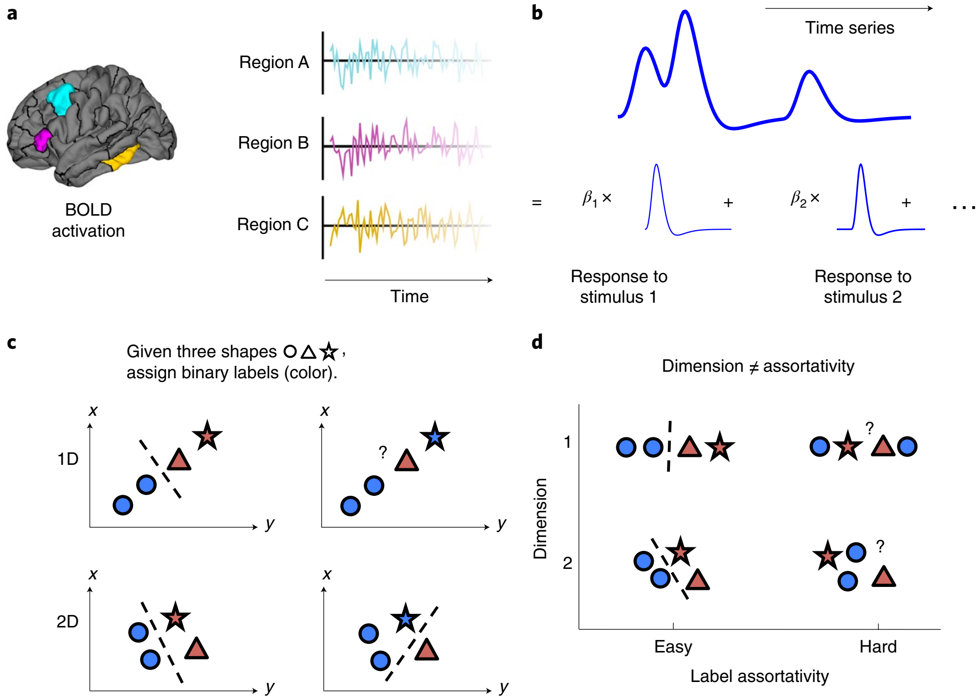

Fig. 1 | Neural responses from the fMRI data: separability dimension and assortativity. a, We measured regional fMRI BOLD activation over 1 h of task practice. **b**, Using a GLM to deconvolve the hemodynamic response function from the BOLD time series, we obtained the approximate neural responses, β_i, to each stimulus at the time point when it was presented. **c**, We assigned binary labels (denoted by color) to the neural data (denoted by shapes). When data are arranged in a low-dimensional manner (top row), some binary assignments result in poorer separability, whereas in a higher dimension these binary assignments can be more easily separated. The average performance of separability over different possible binary assignments gives the separability dimension. When applied to the neural responses to each shape, this procedure yields the dimension of the stimulus representation, where the *x* and *y* axes denote an ROI measurement space. 1C, one-dimensional; 2D, two-dimensional. **d**, Label assortativity does not depend strictly on dimension and can measure a different geometric aspect of the same data. In **c** and **d**, the dashed lines represent a classifier boundary, while the question marks illustrate the difficulty of finding a clean boundary.

stimuli associated with different values. Following this exploratory analysis, we further examined the separability dimension of neural responses from the other days and we verified that this correlation is strongest with data from the final day (see Supplementary Fig. 2). This suggests that these higher dimensional representations emerge most clearly over time and the course of the experiment. This result also survives multiple hypothesis testing for data from all 4 d.

Quick learners have a lower embedding dimension and hence overall more efficient representations. Intuitively, a high-dimensional response per se (independent of its coding information) provides flexibility in coding for stimuli but naturally uses more resources and has more potential to be distorted by errors. In contrast, a low-dimensional response per se uses fewer resources and has less potential for distortion in the encoding process. How might quick learners potentially balance these two competing factors—the use of minimal resources to encode information while ensuring that the encoding is maximally informative of relevant stimuli—to develop efficient neural responses? To address this question, we extended our calculations across a range of values of m, the cardinality of the subset of shapes from which the dimension is estimated. We calculated the correlation between separability dimension and response accuracy for the true stimulus-based data and for the null data in 100 bootstrapped samples, up to $m=11$ (see Fig. 3d and Methods). First, we noticed that the true data were consistently positively correlated (red points) and fell far outside the error bars of the null data (gold points), confirming that, across a range of m, the true data reflect quick learners having a higher stimulus dimension of their representations. In fact, the results at large m are particularly instructive since the combinatorics of 2^m-2 averaged over for each calculation lead to a strong convergence of the results, as reflected in small error bars. Lastly, we noted that while the positive correlation between stimulus dimension and learning accuracy holds over a range of m values, $m=4$ provides the strongest signal and is relatively computationally feasible to calculate in large quantities; further investigations into the stimulus dimension were done with $m=4$.

Across participants, we further observed that the correlation between separability dimension and learning accuracy is negative in the null data, particularly for large m values (see Fig. 3d). Intuitively, these data are the geometric distribution of neural responses per se, independent of stimulus information, and thus reflect the embedding space of neural activity during the task. Therefore, we refer to the separability dimension of these null data as the embedding dimension. Surprisingly, the negative correlation between participants' learning accuracy and embedding dimension shows that fast learners have a lower embedding dimension, complementing their higher stimulus dimension. This large stimulus to embedding dimension ratio for fast learners suggests an efficient cognitive coding: the use of a smaller amount of embedding resources from which a more informative set of task-relevant features can be constructed. We provide a low-dimensional schematic comparing such geometric arrangements in Fig. 3e. While the use of efficiency as a construct in cognitive science has been debated[4], in this study we provide a mathematical definition that contrasts the coding for meaningful content with the neural activity involved per se, via the ratio between stimulus and embedding dimensions.

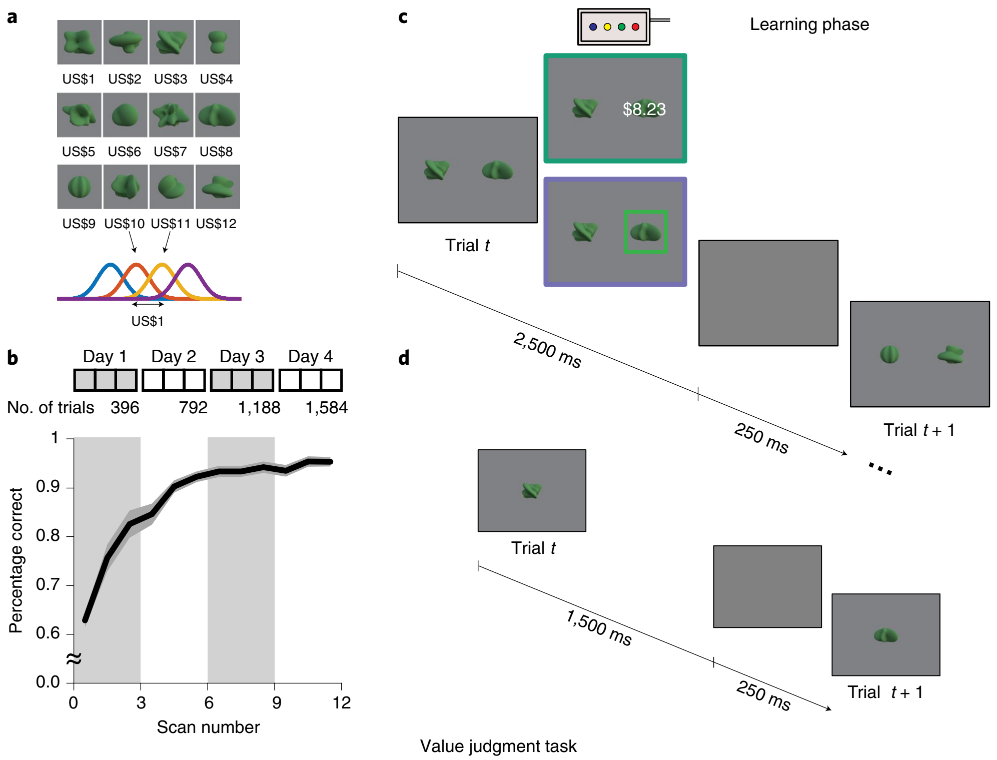

Fig. 2 | Experimental protocol and behavioral results. a, Stimulus set and corresponding values. Twelve abstract shapes were computer-generated and an integer value between $1 and $12 was assigned to each. On each trial, the empirical value of each shape was drawn from a Gaussian distribution with a fixed mean and an s.d. of $0.50. **b**, The experiment was conducted over 4 consecutive days, with learning phases (3 experimental scans or 396 trials each day, for a total of 1,584 trials) and a daily value judgment task in which stimuli were presented singly. Participants' accuracy in selecting the shape with the higher expected value improved steadily over the course of the experiment, increasing from chance level in the first few trials to approximately 95% in the final few trials. Error clouds are centered at the mean and show the standard error ($n=20$). **c**, During the learning phase, participants were presented with two shapes side by side on the screen and asked to choose the shape with the higher monetary value. Once a selection was made, feedback on their selection was provided. Each trial lasted 2.75 s (250 ms interstimulus interval). **d**, During the value judgment task, participants were presented with a single shape and asked if the shape was one of the six least or six most valuable shapes. No feedback was provided; each trial lasted 1.75 s (250 ms interstimulus interval).

Quick learners show high-dimensional stimulus representations within local brain regions. To better understand the main effects reported in the previous sections, we first sought to determine which regions contribute most to the higher stimulus dimension observed in quick learners. To address this question, we conducted a virtual lesioning analysis where we removed brain regions one at a time; then, we recalculated the separability dimension of the modified representation. The regions whose absence caused the largest change in the observed correlation between separability dimension and response accuracy across participants were the left hippocampus and right temporal pole, respectively (magnitude of z-score > 2 or $P<0.023$, uncorrected; see Fig. 4). A possible explanation for these results is that learning to perform this task requires effective separability of stimulus dimensions mediated by these regions. Such an interpretation is in line with the known role of the hippocampus in the rapid learning of stimulus associations[17], and the role of the temporal pole in representing information about abstract conceptual properties of objects (such as object value)[18].

Up to this point, we have studied neural activity across the whole brain and the separability dimension of such neural activity. It is natural to ask if this relationship between learning ability and the dimension of neural responses can also be found in the multivoxel patterns of single brain regions hypothesized to be relevant for task performance. To address this question, we adapted our approach to examine 10 ROIs composed of 300 (or fewer) voxels (see Methods and Table 1). Following the prior analyses, we examined the correlation between separability dimension in the neural data in each local region and the participants' learning accuracy. Overall, we noted that none of the regions show a negative correlation between their separability dimension and learning accuracy. Moreover, we found that three regions showed a significant positive correlation, greater in magnitude than expected in the null model of shuffled data (non-parametric $P\leq 0.05$; see Fig. 5): the left anterior cingulate (ACC) and primary visual cortices, as well as the right posterior fusiform cortex. We noted that only the non-parametric test for the left ACC displayed $P\leq 0.05$ after correcting for multiple comparisons. Notably, the ACC is thought to play a role in reward-based learning[19], while the V1 areas of the visual cortex and the posterior fusiform cortex are involved in the representation of lower- and higher-level features of objects, respectively[20]. Therefore, our findings suggest that these regions are comparatively more engaged in the creation of a value-related heuristic at a local level.

Quick learners develop more assortative representations. Besides separability dimension, a complementary geometric measure is that of label assortativity, which simply identifies how easily distinguishable neural responses are from each other according to all labels, not just according to binarized labels. While data that are more assortative are typically also higher dimensional, it is also possible for these metrics to vary independently (see Fig. 1d). Data can be arranged in a high- or low-dimensional manner but still be easily classifiable (Fig. 1d, left) or data can be arranged in a high- or

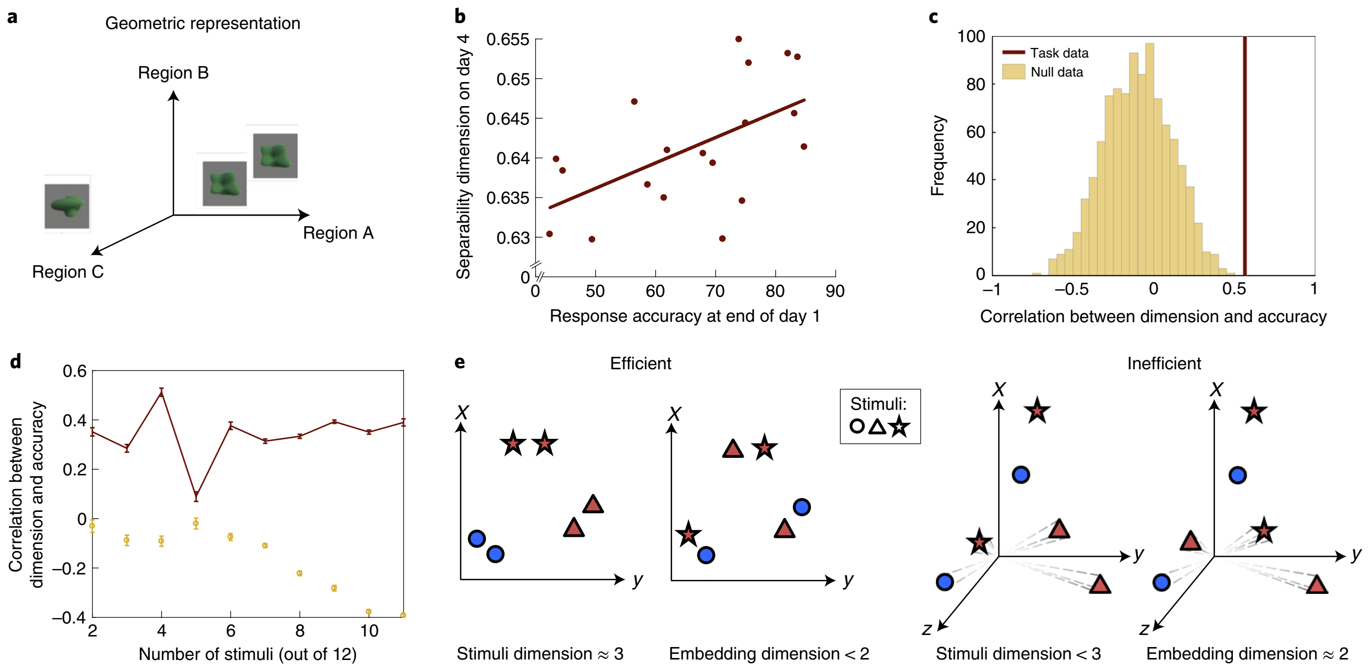

Fig. 3 | Quick learners show higher dimensional and more efficient representations. a, Schematic of how the presentation of each shape evokes neural responses across brain regions to contribute a data point in the ROI space. Many shape presentations in a task session jointly form a point cloud or geometric representation, whose dimension we quantified. **b**, Relationship between the stimulus separability dimension ($m = 4$) and learning accuracy across 19 participants (Pearson's correlation, $r = 0.56$). We compared this correlation value with that observed in a permutation-based null model in which object labels are shuffled uniformly at random and the separability dimension is recalculated. We found that the true correlation is significantly greater than that expected under this null model with a one-sided, non-parametric $P < 0.001$. **c**, Histogram of 1,000 bootstrapped estimates of the value of the across-participant correlation coefficient between participant-specific response accuracy and the dimension of participant-specific null data (gold bars). The correlation value estimated from the true stimulus-based data is shown in red. **d**, Relationship between separability dimension and learning accuracy across participants for m from 2 to 11. The true data are shown in red while the null data are shown in gold (same color scheme as in **c**). The estimates become more reliable as m increases. The true data display a positive correlation with a magnitude far outside the error bars of the null model, which by contrast displays a negative correlation. These observations jointly suggest that fast learners have a large stimulus but small embedding dimensions, overall forming an efficient representation of neural responses. The error bars are centered at the mean and show the s.e.m. based on 20 draws for the true data (red) and 100 bootstrapped samples for the null data (gold). **e**. Schematic of (left) an efficient representation with high stimulus and low embedding dimensions and (right) an inefficient representation with comparatively lower stimulus and higher embedding dimensions. Our findings suggest that fast learners possess efficient neural representations, as manifested by a larger ratio of stimulus to embedding dimension. The x, y and z axes denote an ROI measurement space.

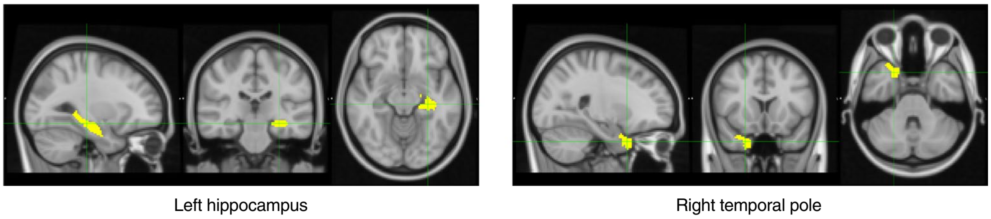

Fig. 4 | Regional drivers of the relationship between representation and behavior. A virtual lesioning experiment shows the brain regions that most weaken the correlation between separability dimension and learning accuracy on removal (z-score < –2; see Methods). We found that removal of the left hippocampus and right temporal pole, respectively, caused the largest decreases in the observed correlation. In other words, in individuals that learn quickly, the left hippocampus and right temporal pole seem to contribute to a higher separability dimension and vice versa. In contrast, removal of regions such as the left rostral middle frontal cortex and left supramarginal gyrus most strongly enhance the observed correlation, suggesting that their activity is orthogonal to, or does not directly contribute to, the large separability dimension that characterizes quick learners.

low-dimensional manner but be difficult to classify (Fig. 1d, right). Hence, the analysis of both metrics provides distinct and potentially independent information regarding the organization of the data. We hypothesized that quick learners would show a more assortative representation, in addition to having a higher stimulus dimension (see Fig. 6a). In the current study, we calculated assortativity using a linear support vector machine (SVM), chosen because of its simple interpretability. When examining the same neural data from the value judgment session at the end of the fourth day, we found a positive correlation between assortativity and the response accuracy of participants on the first day ($r = 0.55$, non-parametric $P = 0.012$ estimated from a null model where labels are randomly permuted; see Fig. 6b). Intuitively, these data suggest that participants who learn more quickly have a more assortative pattern of neural responses

Table 1 | Brain regions where a higher dimensional representation is correlated with learning ability.

No. of voxels	Brain region	Hemisphere	*r*	*P*
300	ACC	Left	0.543	0.003*
300	ACC	Right	0.306	0.063
300	Primary visual cortex	Left	0.500	0.016
300	Primary visual cortex	Right	0.090	0.390
300	Posterior fusiform cortex	Left	0.085	0.396
300	Posterior fusiform cortex	Right	0.608	0.050
300	Lateral occipital cortex	Left	0.415	0.092
300	Lateral occipital cortex	Right	0.0591	0.465
140	Orbitofrontal cortex	Left	0.142	0.291
140	Orbitofrontal cortex	Right	0.357	0.103

The *r* and non-parametric *P* values are given from comparison with the null model. The left ACC passes the non-parametric $P < 0.005$ threshold corrected for multiple comparisons (marked with an asterisk).

than participants who learn less quickly. To verify that the metrics of separability dimension and label assortativity do not have a strict overlap, we noted that one metric explains approximately $r^2 = 34\%$ of the variance of the other metric, where r is Pearson's correlation.

Discussion

In this study, we developed and applied a computational framework to reveal how the high-dimensional neural responses of quick learners allow for greater distinguishability of meaningful stimuli while requiring fewer informational resources. Our observations were enabled by emerging methods from machine learning and data science[8,9], which can be used to estimate the intrinsic dimension of a representation despite pervasive measurement noise. We extended the metric of the stimulus dimension[8] to study a complex cognitive task in whole-brain neural data; we also introduced the new idea of the embedding dimension. In a cohort of 20 healthy adult humans learning the value of new objects over the course of 4 d, we found that participants who learn most quickly display uniquely optimized neural responses to encode the cognitive processes associated with the task. The joint profile of the stimulus and embedding dimensions allows us to quantify a concept of cognitive coding efficiency, based on the ratio between these two dimensions for each individual. We complemented this examination with supporting studies of finer neuroanatomy (assessing multivoxel patterns) and computation (assessing local assortativity). Broadly, our work provides a suite of tools to characterize response geometry, thereby offering a simple and intuitive explanation for how individuals learn to successfully distinguish between relevant stimuli in their environment over time.

A notion of cognitive coding efficiency. The concept of coding efficiency has been exercised at smaller spatial scales to characterize the (often unexpectedly low) dimension of neural representations. For example, neuronal spiking patterns measured in the lateral intraparietal area as macaques engage in a visual spatial attention task map onto a one-dimensional dynamical trajectory[10]. The simplicity and low-dimensionality of these dynamics mark disparate cognitive processes from decision-making and attentional shifting, to biased representations that arise from associative learning[11]. Indeed, such low-dimensionality is almost ubiquitous in neuronal measurements[21], although this often saturates the low-dimensional bound set by the limited complexity of neural tasks commonly used today[22], or their autocorrelation structure[23]. Within this low-dimensional manifold, temporal variation in this 'effective' dimension of neural activity can also indicate temporal variation in behavior[24]. For example, as macaques engage in a recall task, the estimated stimulus dimension from neural spiking activity in the prefrontal cortex is higher during correct responses than during incorrect responses[8].

Extending previous methods, we introduced two complementary types of dimension (stimulus and embedding) that allow insight into learning capacity and cognitive flexibility. Our results are consistent with the notion that the substantially different use of these two types of dimension allows the efficient encoding of contextually relevant data, potentially supporting optimal learning strategies. The compression of a large amount of information or content into a restricted number of channels has been studied in other cognitive domains such as sensory processing[2,3]. In light of these historical contributions, our results suggest that similar principles of geometric efficiency may extend to higher-order cognitive processes in humans. Further work could directly investigate commonalities in such principles across different scales of space and time. Such an investigation is in principle made possible by the fact that while the absolute value of these geometric metrics depends on the particular measurement technique, relative changes in value could be used to compare between data collected across wholly different measurement techniques.

Complex cognitive tasks require new models of cognitive coding efficiency. Recent theoretical studies have used biologically plausible models to demonstrate that complex tasks such as image recognition or sensory processing are supported by high- dimensional representations, which in turn allow for an accurate readout of stimulus identity[25]. These and other theoretical developments show that the two types of dimension (stimulus and embedding dimensions) may have very different advantages and behavior, even within the same experiment or within the same neural network[26]. An efficient balance between these two types of dimension may control a generalization–discrimination trade-off[25]; new models accounting for these two dimensions are necessary especially for the fundamental understanding of complex cognitive tasks. In a separate line of work, the concept of efficiency has been applied to large-scale human neuroimaging data, predominantly to describe situations where the behavior of participants appears similar but neural activation is greater for one group (taken to be the 'less efficient' group) than for the other[6,7]. For instance, in an experiment involving working memory, less neural activity was needed for trained items compared to new ones[27]. The authors interpreted this difference as a correlate of a gain in neural efficiency, and that training causes a more efficient neural representation. However, it has been pointed out that this interpretation does not shed light on the relationship between these two facts[4]. In our study, we show that a more compact dimension of neural activation is simultaneously tied to larger information content in the same neural activation, leading to the idea of efficiency in the representation itself. This notion is more akin to how the concept of efficiency is used in other contexts in the neuroscience literature, such as in studies of efficient coding in sensory systems[2,3] or in studies of network efficiency[28,29], where a maximal amount of information is conveyed through a fixed (or smaller) feature or basis set. That the efficient cognitive coding we observed also appears differentially in individuals who learn faster is consistent and intuitive, but is not in itself required for our definition of efficiency. Hence, our calculations of the dimension of representation provide a rigorous framework for quantifying and reasoning about the efficiency of cognitive coding, which can be measured and compared in other cognitive processes.

In our experiment, participants were presented with a set of shapes designed to have no visual features that correlated with their monetary value (see Methods). Each participant was required to flexibly reassign new values to these shapes through the course of the experiment. In general, humans can be guided to act according

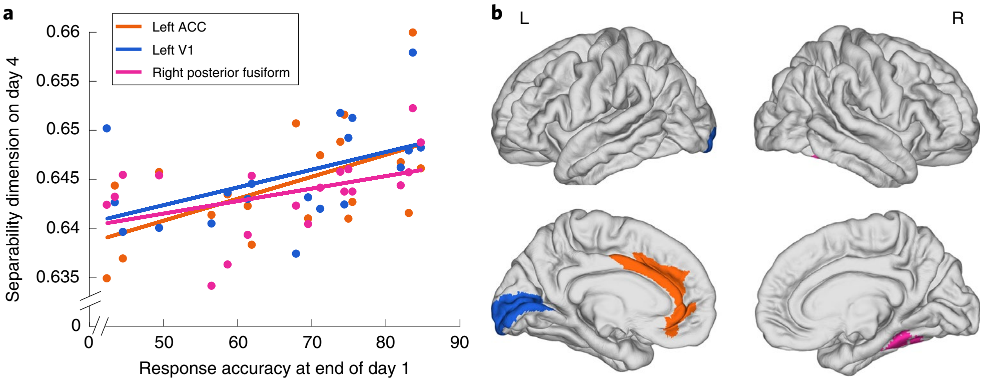

Fig. 5 | Quick learners show a larger stimulus dimension of responses in certain task-relevant regions at the voxel level. a, We studied the regions of 300 (or fewer) voxels that we hypothesized were involved in the processing of value and learning of shapes. Three regions showed a positive correlation between learning accuracy and separability dimension, with non-parametric $P \leq 0.05$ compared to the null model of shuffled data in a one-sided permutation test ($n = 1,000$). **b**, Topographical representation of these three regions on the surface of the brain: the left (L) ACC, left primary visual area and right (R) posterior fusiform. The laterality of this latter effect is consistent with prior work demonstrating that the right and left posterior fusiform exhibit differential responses during object recognition[46-48].

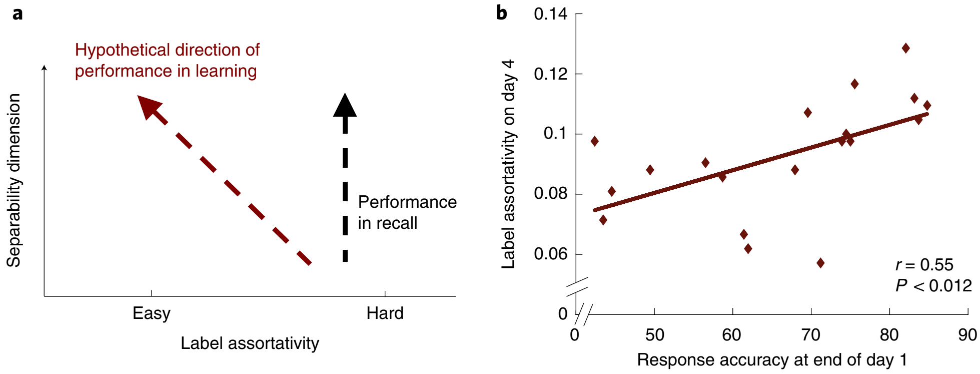

Fig. 6 | Dimension and assortativity provide a geometric depiction of neural data. a, Since different cognitive processes can exhibit typified geometric changes in the neural responses to various stimuli, we hypothesized that learning performance is associated with both higher dimension and higher assortativity. **b**, The distinct metric of label assortativity (according to all labels; see Fig. 1d) across the whole brain shows that quick learners display a higher assortativity (Pearson's $r = 0.55$; red markers) compared to the shuffled data in gold with non-parametric $P < 0.012$, in a one-sided permutation test ($n = 1,000$).

to what has been previously reinforced or to move toward promising sources of future reward[30]. Our work examines the neural basis that supports this flexible identification of new value to existing objects, and how such objects become distinguished from each other in the representation of neural activity according to stimulus cues. Indeed, on investigating data from sessions where participants were asked to evaluate the size and not the value of each shape, fast learners showed no particular difference in the dimension of their neural responses (see Fig. 7). Our results complement previous investigations into the relevance of cognitive flexibility for effective learning[31] and the underlying processes of executive function[32,33], while illuminating the emergent geometric architecture of the neural responses of effective learners. Future work could determine whether efficient geometric representations arise in individuals who exhibit higher degrees of cognitive flexibility and dynamic reorganization of their neural responses.

Changes in neural representations during learning and practice. In seeking to decipher the rules of adaptation, learning and development, it is common to examine how neural mechanisms support or foster behavioral patterns. In a complementary perspective, one can examine how temporally localized decisions or short-term behaviors can drive adaptation or change in neural circuitry. Indeed, developmental systems theory has long viewed behavior as the leading edge of adaptation[34]. From this perspective, behavior serves an integrative function, inducing changes in intraorganismic activity (encompassing brain structure and function) in response to changes in extraorganismal activity (encompassing the environment within which the individual is situated). Crucially, the timescale of change differs across levels of analysis such that the various systems (for example, behavior, brain) are differentially open for change over different periods of time. It is in response to persistent (rather than transient) behavior change that activity within other systems eventually becomes aligned with behavior by virtue of bidirectional activity across the levels of analysis. This perspective is also supported by computational modeling studies using deep networks, which show that network weights continue to change long after the networks demonstrate low thresholds of learning errors[35,36]. Indeed, after learning errors have reduced to a regime where the deep networks are performing accurately, it is common practice to continue updating and changing the weights as the representations continue to improve

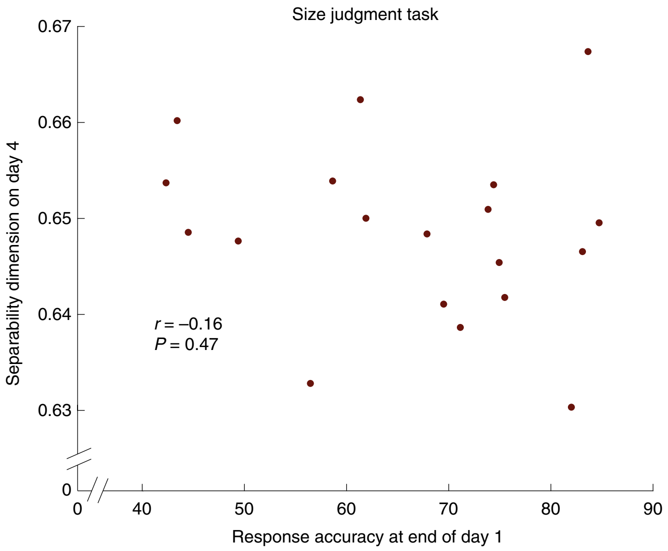

Fig. 7 | Dimension of neural data from the size judgment session. The separability dimension of data from the size judgment task on the last day does not show significant differences between quick and slow learners, suggesting that the cognitive task or effort of judging value itself is necessary for this emergence of a larger dimensional neural response. Pearson's correlation is shown in a one-sided test ($n = 19$).

(in compression and generalization), even after the learning errors have stopped decreasing.

Such artificial neural network models are consistent with observations in human neuroscience, where different phases in adaptation across multiple timescales are observed. Fast improvements tend to be seen during the early phase of learning; slower improvements when automaticity develops are observed in a later phase of learning[37]. The corresponding changes in neural representations outlast a training session and can be observed up to months or a year later[37]. Some notable studies identified changes in fMRI measurements of brain activity following video game playing, which were associated with improvements in visuospatial and attention-related skills[38], as well as in the rate of regional subcortical glucose metabolism[39]. Changes from tasks involving spatial navigation and visuomotor coordination have also been identified in structural brain properties, with the effects outlasting even a short intensive gaming period[40]. In our study, as participants learned to associate rewards of different magnitudes with novel stimuli, it is likely that new neural representations would emerge over days to represent these distinct groups. The emergence of structured neural representations over long (rather than short) timescales is particularly expected in tasks whose stimuli have complex similarities and differences[41] or multiple layers of content[42], where recurrent retrieval might foster variability in the representations in early learning caused by the altering or adding of underlying memory representations, thereby leading to more effective memories in the future[43]. Additional early modulators of later neural representations include consolidation[41], insight[44] and the nature of the learning environment[45].

Role of single regions within a broader whole-brain geometry. Geometry and topology can be investigated across multiple scales of any complex system or its emergent dynamics. While some systems can display heterogeneity in geometric principles across spatial and temporal scales, others display greater scale invariance, with the principles at one scale being recapitulated at other scales. Applying our methods at different scales, we found that neural activity patterns elicited by value judgments of learned stimuli display similar geometric principles whether assessed at the level of the whole brain, or at the level of multivoxel patterns in single brain areas. Our choice to begin with an analysis of ROIs across the whole brain complements prior studies that often focused on fine-grained voxel patterns, and captures global organization that would be relevant during value learning. On a smaller scale, we found that the left primary visual cortex, ACC and right posterior fusiform cortex of quick learners display a differential increase in dimension. While we focused on just ten local regions, each of which are hypothesized to play an important role in the cognitive processes elicited by this task, it would be of interest to expand the study to additional regions or sets of regions defined with other methods. Then, using the computational techniques we have introduced, one could begin to bridge the regional drivers of whole-brain simplicity and complexity in response geometry.

Methodological considerations. Several methodological considerations are pertinent to our study. First, while the GLM extracts neural responses from the time series averaged across entire regions, it could also be useful to perform this extraction on time series at the voxel level before averaging; this may decrease the noise from irrelevant signals. Second, dimension and assortativity constitute starting points for a deeper analysis and further work could identify the exact topology of the response. Third, the broad geometric methods that we developed and used in this study could be complemented by a dynamic study to assess how this geometry evolves across time. Fourth, while our cohort of 20 individuals already demonstrate significant evidence for geometric features that distinguish quick from slow learners, these results could well be verified across larger samples. Fifth, in our work, we found a significantly higher stimulus dimension in the neural responses of quick learners on the last day (see Supplementary Figs. 1–3), suggesting that this higher dimensional and more efficient representation emerges most clearly over time and training. However, our results remain correlative and cannot suggest a causal link between this high-dimensional representation and effective learning. Sixth, our results can be replicated using a different whole-brain parcellation (Supplementary Fig. 4), alternative measures of differences in stimulus representation (Supplementary Fig. 5) and behavioral metric (Supplementary Fig. 6), as well as cross-validation data partition (Supplementary Fig. 7). Finally, we studied a single cognitive task; future work could extend these notions to other cognitive domains during different experiments, or as different cognitive processes are engaged. In a previous experiment examining recall performance in trained macaques, the two estimates of dimension and decoding accuracy (analogous to assortativity) were differentially related to behavior[8]. Specifically, while the stimulus dimension of the macaque's neural representation was predictive of the macaque's performance, the decoding accuracy of the same neural data instead remained constant in both error and correct trials. These observations raise fundamental questions about whether different cognitive processes can exhibit typified geometric changes in neural responses. In humans, a particularly interesting context in which to study such differences is the mental states engendered by 'explore' versus 'exploit' behaviors common in general human experience, which are thought to give rise to diffuse versus structured neural representations.

Conclusion. In the current study, we offer a computational framework for quantifying and understanding the geometry of neural responses in humans. The tools we have developed and exercised hold promise for the analysis of other complex cognitive tasks due to their general applicability to non-invasive neuroimaging and notable robustness to noise. We illustrate the utility of these tools in characterizing the organization of neural activity associated with effective cognitive performance and efficient cognitive coding during the learning of abstract values associated with novel objects. Our results suggest that effective learners are marked by a type of cognitive coding efficiency characterized by high-dimensional geometric representations in concert with a compact embedding of the stimulus information. Our observations motivate future work in cognitive and clinical neuroscience examining the generalizability of this notion of efficiency and its relevance for disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41593-019-0400-9.

Received: 6 December 2017; Accepted: 29 March 2019;
Published online: 20 May 2019

References

1. Freedman, D. J., Riesenhuber, M., Poggio, T. & Miller, E. K. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* **291**, 312–316 (2001).
2. Barlow, H. in *Sensory Communication* (ed. Rosenblith, W. A.) Ch. 13 (MIT Press, 1961).
3. Olshausen, B. A. & Field, D. J. Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Res.* **37**, 3311–3325 (1997).
4. Poldrack, R. A. Is efficiency a useful concept in cognitive neuroscience? *Dev. Cogn. Neurosci.* **11**, 12–17 (2015).
5. Buzsáki, G. & Watson, B. O. Brain rhythms and neural syntax: implications for efficient coding of cognitive content and neuropsychiatric disease. *Dialogues Clin. Neurosci.* **14**, 345–367 (2012).
6. Gold, B. T., Kim, C., Johnson, N. F., Kryscio, R. J. & Smith, C. D. Lifelong bilingualism maintains neural efficiency for cognitive control in aging. *J. Neurosci.* **33**, 387–396 (2013).
7. Heinzel, S. et al. Working memory load-dependent brain response predicts behavioral training gains in older adults. *J. Neurosci.* **34**, 1224–1233 (2014).
8. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
9. Diedrichsen, J., Wiestler, T. & Ejaz, N. A multivariate method to determine the dimensionality of neural representation from population activity. *Neuroimage* **76**, 225–235 (2013).
10. Ganguli, S. et al. One-dimensional dynamics of attention and decision making in LIP. *Neuron* **58**, 15–25 (2008).
11. Fitzgerald, J. K. et al. Biased associative representations in parietal cortex. *Neuron* **77**, 180–191 (2013).
12. Grill-Spector, K. & Malach, R. The human visual cortex. *Annu. Rev. Neurosci.* **27**, 649–677 (2004).
13. Waskom, M. L., Kumaran, D., Gordon, A. M., Rissman, J. & Wagner, A. D. Frontoparietal representations of task context support the flexible control of goal-directed cognition. *J. Neurosci.* **34**, 10743–10755 (2014).
14. Bzdok, D. et al. Characterization of the temporo-parietal junction by combining data-driven parcellation, complementary connectivity analyses, and functional decoding. *Neuroimage* **81**, 381–392 (2013).
15. Chang, L. J., Yarkoni, T., Khaw, M. W. & Sanfey, A. G. Decoding the role of the insula in human cognition: functional parcellation and large-scale reverse inference. *Cereb. Cortex* **23**, 739–749 (2013).
16. Mattar, M. G., Thompson-Schill, S. L. & Bassett, D. S. The network architecture of value learning. *Netw. Neurosci.* **2**, 128–149 (2018).
17. Squire, L. R. Memory and the hippocampus: a synthesis from findings with rats, monkeys, and humans. *Psychol. Rev.* **99**, 195–231 (1992).
18. Peelen, M. V. & Caramazza, A. Conceptual object representations in human anterior temporal cortex. *J. Neurosci.* **32**, 15728–15736 (2012).
19. Bush, G. et al. Dorsal anterior cingulate cortex: a role in reward-based decision making. *Proc. Natl Acad. Sci. USA* **99**, 523–528 (2002).
20. Grill-Spector, K. The neural basis of object perception. *Curr. Opin. Neurobiol.* **13**, 159–166 (2003).
21. Cunningham, J. P. & Yu, B. M. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* **17**, 1500–1509 (2014).
22. Gao, P. et al. A theory of multineuronal dimensionality, dynamics and measurement. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/214262v2 (2017).
23. Fusi, S., Miller, E. K. & Rigotti, M. Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.* **37**, 66–74 (2016).
24. Sadtler, P. T. et al. Neural constraints on learning. *Nature* **512**, 423–426 (2014).
25. Barak, O., Rigotti, M. & Fusi, S. The sparseness of mixed selectivity neurons controls the generalization–discrimination trade-off. *J. Neurosci.* **33**, 3844–3856 (2013).
26. Chung, S., Lee, D. D. & Sompolinsky, H. Classification and geometry of general perceptual manifolds. *Phys. Rev. X* **8**, 031003 (2018).
27. Zimmer, H. D., Popp, C., Reith, W. & Krick, C. Gains of item-specific training in visual working memory and their neural correlates. *Brain Res.* **1466**, 44–55 (2012).
28. Bullmore, E. & Sporns, O. The economy of brain network organization. *Nat. Rev. Neurosci.* **13**, 336–349 (2012).
29. Bassett, D. S. et al. Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS Comput. Biol.* **6**, e1000748 (2010).
30. Simon, H. A. A behavioral model of rational choice. *Q. J. Econ.* **69**, 99–118 (1955).
31. Bassett, D. S. et al. Dynamic reconfiguration of human brain networks during learning. *Proc. Natl Acad. Sci. USA* **108**, 7641–7646 (2011).
32. Shine, J. et al. The dynamics of functional brain networks: integrated network states during cognitive task performance. *Neuron* **92**, 544–554 (2016).
33. Braun, U. et al. Dynamic reconfiguration of frontal brain networks during executive cognition in humans. *Proc. Natl Acad. Sci. USA* **112**, 11678–11683 (2015).
34. Gariepy, J.-L. in *Developmental Science*, 3rd edn, Vol. 4 (eds Cairns, R. B. et al.) Ch. 8 (Cambridge University Press, 1996).
35. Tishby, N. & Zaslavsky, N. Deep learning and the information bottleneck principle. In *Proc. 2015 IEEE Information Theory Workshop* (ed. Xing, C.) 1–5 (ITW, 2015).
36. Goldt, S. & Seifert, U. Thermodynamic efficiency of learning a rule in neural networks. *New J. Phys.* **19**, 113001 (2017).
37. Ruitenberg, M. F. L. et al. Neural correlates of multi-day learning and savings in sensorimotor adaptation. *Sci. Rep.* **8**, 14286 (2018).
38. Gorbet, D. J. & Sergio, L. E. Move faster, think later: women who play action video games have quicker visually-guided responses with later onset visuomotorrelated brain activity. *PLoS One* **13**, e0189110 (2018).
39. Haier, R. J. et al. Regional glucose metabolic changes after learning a complex visuospatial/motor task: a positron emission tomographic study. *Brain Res.* **570**, 134–143 (1992).
40. Momi, D. et al. Acute and long-lasting cortical thickness changes following intensive first-person action videogame practice. *Behav. Brain Res.* **353**, 62–73 (2018).
41. Tompary, A. & Davachi, L. Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. *Neuron* **96**, 228–241.e5 (2017).
42. Mason, R. A. & Just, M. A. Physics instruction induces changes in neural knowledge representation during successive stages of learning. *Neuroimage* **111**, 36–48 (2015).
43. Karlsson Wirebring, L. et al. Lesser neural pattern similarity across repeated tests is associated with better long-term memory retention. *J. Neurosci.* **35**, 9595–9602 (2015).
44. Milivojevic, B., Vicente-Grabovetsky, A. & Doeller, C. F. Insight reconfigures hippocampal-prefrontal memories. *Curr. Biol.* **25**, 821–830 (2015).
45. Dunsmoor, J. E., Kragel, P. A., Martin, A. & LaBar, K. S. Aversive learning modulates cortical representations of object categories. *Cereb. Cortex* **24**, 2859–2872 (2014).
46. Vuilleumier, P., Henson, R. N., Driver, J. & Dolan, R. J. Multiple levels of visual object constancy revealed by event-related fMRI of repetition priming. *Nat. Neurosci.* **5**, 491–499 (2002).
47. Koutstaal, W. et al. Perceptual specificity in visual object priming: functional magnetic resonance imaging evidence for a laterality difference in fusiform cortex. *Neuropsychologia* **39**, 184–199 (2001).
48. Simons, J. S., Koutstaal, W., Prince, S., Wagner, A. D. & Schacter, D. L. Neural mechanisms of visual object priming: evidence for perceptual and semantic distinctions in fusiform cortex. *Neuroimage* **19**, 613–626 (2003).

Acknowledgements

We thank B. Falk for helpful discussions and S. Solomon for helpful comments on an earlier version of this paper. This work was supported by the J.D. and C.T. MacArthur, A.P. Sloan, ISI and Paul G. Allen Family Foundations, the US Army Research Laboratory (grant no. W911NF-10-20022), the Army Research Office (grant nos. Bassett-W911NF-141-0679, Grafton-W911NF-16-1-0474 and DCIST-W911NF17-2-0181), the Office of Naval Research, the National Institute of Mental Health (grant nos. 2-R01-DC-009209-11, R01MH112847, R01-MH107235 and R21-M MH-106799), the Eunice Kennedy Shriver National Institute of Child Health and Human Development (grant no. 1R01HD086888-01), the National Institute of Neurological Disorders and Stroke (grant no. R01 NS099348) and the National Science Foundation (grant nos. BCS-1441502, BCS-1430087, NSF PHY-1554488 and BCS-1631550). The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding agencies.

Author contributions

E.T. developed the theory, performed the computational modeling and wrote the manuscript. E.T., M.G.M. and D.S.B. designed the study. D.S.B. M.G.M. and D.L.S. revised the manuscript. C.G. contributed intellectually to theory development through discussions. D.L.S. performed the statistical analyses and contributed to data interpretation. M.G.M. developed the experiment in collaboration with S.T.-S. and D.S.B. M.G.M. also collected and preprocessed the data. S.T.-S. acquired funding to support data collection and contributed to data interpretation. D.S.B. acquired funding to support theory development and data analysis, and contributed to theory and data interpretation.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-019-0400-9.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to D.S.B.

Journal peer review information: Nature Neuroscience thanks Stefano Fusi and other anonymous reviewer(s) for their contribution to the peer review of this work.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Dimension estimation. Given several types of data such as the shapes in Fig. 1c and given that there can be several measurements for the same shape, we can assign a binary label to each shape, represented by the color. In our case, each shape represents a neural response to one of n stimuli. Given n stimuli or shapes, there are 2^n ways to assign binary labels to these data. We can then ask how separable these binary groups are across all $2^n - 2$ relabelings[8]. Note that the additional -2 is because 2 cases out of the 2^n assign the same label to all of the data; thus, it is clearly unnecessary to calculate separability in those cases. When the data are arranged in one dimension, it becomes hard to separate the binary groups in all but one of the binary assignments. When the data are in a higher dimension, it is easier to separate these binary groups. Hence, the average binary separability over different assignments estimates the separability dimension of this geometric representation of the data, that is, a higher value indicates that the data effectively live in a larger dimensional space.

All simulations were performed in MATLAB (MathWorks). Each calculation of linear separability was cross-validated across partitions of the data. Specifically, to calculate linear separability on the binary categories, we used the default validation scheme within the Classification Learner application, retaining the default option of fivefold cross-validation. This algorithm partitions the data into five disjoint sets or folds, chosen randomly but with roughly equal size. For each fold, the algorithm trains a linear SVM using the out-of-fold observations and then assesses model performance using infold data. Next, the average test error is calculated over all folds to yield the separability for each binary assignment, which is a number between 0 and 1 (with the chance level being 1/2). We repeated this process over all binary assignments to obtain the separability in each case, and we took the average separability over all $2^n - 2$ assignments for n types of data (or object identities). An advantage of this process of averaging over many separating hyperplanes is a robustness of the results to noise; while the result in any particular hyperplane might be sensitive to perturbation, the average result will be stable.

The resulting average is the separability dimension, which has a monotonic relation with the cardinal dimension. Separability dimension is hence a useful proxy for cardinal dimension and is sufficient to show relative differences between individuals, which was the purpose of our study. Note that the cardinal dimension, which is more intuitively familiar and ≥1, could be inferred by counting N_c, the number of successful binary assignments above a threshold, and relating that to the cardinal dimension d using $d = \log_2 N_c$ (ref. [8]). However, the number of data points needed to extrapolate this cardinal dimension (approximately 4,000 in Rigotti et al.[8]) exceeds the amount of measured data available from typical experiments, and this number increases with task complexity. Hence, estimating the cardinal dimension often requires the introduction of additional data resampling techniques, which we chose not to use.

We performed this analysis on m subsets of the stimuli. That is, for m stimuli out of the 12 there are $\binom{12}{m}$ ways to assign binary labels. We chose 20 draws out of the different possible combinations in a uniform way, such that each stimulus is represented a similar number of times. This can be done for $m = 2,\ldots, 10$, where $\binom{12}{m} > 20$; for $m = 11$ we used all 12 possible draws. To preserve statistical rigor we did not study $m = 12$ since there would be only one draw for $m = 12$. For most calculations, we chose to use $m = 4$ as a midsize subset due to computational tractability, except in Fig. 3c where we show results for all $m < 12$ to verify that the conclusions remain similar.

Linear SVM and cross-validation. In calculating binary separability, the MATLAB linear SVM is used with cross-validation by partitioning the data in five folds. For each fold, a model was trained using the out-of-fold observations, after which model performance was assessed using infold data. The average test error is calculated over all folds to provide an estimate of the predictive accuracy of the final model and is used as the measure of binary separability. A similar cross-validation procedure is used to calculate label assortativity. In this case the MATLAB linear SVM is also used with the data retaining all $n = 12$ distinct labels.

Effects of noise on the measure of dimensionality. This method is robust to generic noise that artificially enlarges the dimension of the data in all directions. This is because of the cross-validation approach that uses different samples for training and testing where the realization of the noise is different. This is the typical problem that a readout in the brain has to solve. The representational variance is not consistent across repetitions of the same trial; hence, the geometry of the set of points is not inherently different from the noiseless case. Note that this would not be true for a simple application of principal component analysis. However, a low signal-to-noise ratio, or when noise is large along the relevant directions, will cause a collapse of dimensionality[8] or the ability to distinguish between different objects. This situation is interesting and illustrates cases where a readout in the brain will be unable to resolve object identity. Indeed, we expect that the cognitive state of an individual (modulated by attention or perception) plays a significant role in the dimensionality of the ensuing representations. As with other measures that consider the broad geometric or topological properties of neural data[8,49,50], such findings would be well complemented by finer-scale experiments into the link between physiological state and altered neural representations.

Value learning experiment. *Participants.* Twenty participants (9 female; ages 19–53 years; mean age = 26.7 years) with normal or corrected vision and no history of neurological disease or psychiatric disorders were recruited for this experiment. No statistical methods were used to predetermine sample sizes; however, our sample sizes are similar to those reported in previous publications[7,9,16]. All participants volunteered and provided informed consent in writing in accordance with the guidelines of the Institutional Review Board of the University of Pennsylvania (no. 801929). Participants had no prior experience with the stimuli or with the behavioral paradigm. Notably, there were no specific conditions in our experiment, except for the condition of the day of practice. However, it is impossible for human experimenters to be ignorant of the day; thus, data collection and analysis were not performed blind to the conditions of the experiment.

Stimulus design. The novel stimuli were three-dimensional shapes generated with a custom-built MATLAB toolbox (http://github.com/saarela/ShapeToolbox) and rendered with RADIANCE[51]. ShapeToolbox allows the generation of three-dimensional radial frequency patterns by modulating basis shapes, such as spheres, with an arbitrary combination of sinusoidal modulations in different frequencies, phases, amplitudes and orientations. A large number of shapes were generated by selecting combinations of parameters at random. From this set, we selected 12 that were considered to be sufficiently distinct from one another. A different monetary value, varying from US$1.00 to US$12.00 in integer steps, was assigned to each shape (Fig. 2a). These values were uncorrelated with any parameter of the sinusoidal modulations, so that visual features were not informative of value.

Experimental paradigm. Participants learned the monetary value of 12 novel visual stimuli over the course of 4 consecutive days[16]. Each day included the following phases: (1) a size judgment task; (2) a learning phase; (3) a repetition of the size judgment task; (4) a value judgment task. A 10-minute resting-state session preceded the experiments on each day. In the main text, we report data only from the value judgment task.

Learning phase. On each trial of the experiment, participants were presented with two shapes side by side on the screen and asked to choose the shape with the higher monetary value in an effort to maximize the total amount of money in their bank. Feedback (explicit or implicit) was given based on their response (Fig. 2b). The shape values on a given trial were independently drawn from a Gaussian distribution with the mean equal to the true monetary value and the s.d. = US$0.50 (Fig. 2a). This variation in the trial-specific value of a shape was incorporated to ensure that participants thought about the shapes as having worth, as opposed to simply associating a number or label with each shape. The average accuracy in selecting the shape with the highest mean value at each trial gradually improved over the course of the experiment, increasing from approximately 50% (chance) in the first few trials to approximately 95% in the final few trials.

Value judgment task. The value judgment task scans consisted of consecutive presentations of shapes drawn from the set (1,500 ms presentation and 250 ms interstimulus interval) as participants indicated whether the shape was one of the six least or six most valuable shapes. No feedback was given in this task.

We analyze the data from the value judgment scans (both the BOLD data and participants' response accuracy) in the main text of the paper. We focus specifically on these data because the presentation of single stimuli in these sessions allows for the isolation of neural responses to each shape, which would be harder to disentangle from the simultaneous presentation of two shapes characteristic of the task used in the learning sessions. The fMRI time series were poorly recorded for one participant in the value judgment session of the first day, due to a lack of synchronization between computer and scanner. Hence, this participant was excluded from the analyses, with the other 19 participants contributing data for the main analyses described in this article.

The behavioral data reported in the main text is the accuracy in this task (specifically, the accuracy at the end of the first day), while the neural data reported in the main text is measured from this task (specifically, based on day 4).

Size judgment task. The size judgment task scans consisted of consecutive presentations of shapes drawn from the set and presented with a ±10% size modulation (1,500 ms presentation and 250 ms interstimulus interval) as participants indicated whether the shape was presented in a slightly larger or smaller variation.

Image acquisition. We collected BOLD fMRI data from each participant as they performed the task.

Learning phase. A total of 12 scan runs over 4 d were completed by each person (3 scans per session), totaling 1,584 trials (Fig. 2c). Participants completed 20 min of the main task protocol on each scan session, learning the values of the 12 shapes through feedback. The sessions consisted of three scans of 6.6 min each, starting with 16.5 s of a blank gray screen, followed by 132 experimental trials (2.75 s each), and ending with another period of 16.5 s of a blank gray screen. Stimuli were back-projected onto a screen viewed by the participant through a mirror mounted on the head coil and subtended 4 degrees of visual angle, with 10 degrees separating the center of the two shapes. Each presentation lasted 2.5 s (250 ms interstimulus interval) and, at any point within a trial, participants entered their responses on a 4-button response pad indicating their shape selection with a leftmost or rightmost button press. Stimuli were presented in a pseudorandom sequence with every pair of shapes presented once per scan.

Value and size judgment tasks. A total of 4 scan runs over 4 d were completed by each person (one scan per session) for the value judgment task, while a total of 8 scan runs over 4 d were completed by each person (two scans per session) for the size judgment tasks. Each scan lasted 5 min and 22 s (184 trials). Stimuli were back-projected onto a screen viewed by the participant through a mirror mounted on the head coil and subtended 4 degrees of visual angle. Each presentation lasted 1.75 s (250 ms interstimulus interval); at any point within a trial, participants entered their responses on a four-button response pad indicating their shape selection with a leftmost (least valuable) or rightmost (most valuable) button press, during the value judgment tasks. During the size judgment task, these leftmost and rightmost button presses corresponded to smaller and larger shapes, respectively. Stimuli were presented in a counterbalanced sequence.

MRI data collection and preprocessing. MRI images were obtained at the Hospital of the University of Pennsylvania using a Siemens Tim Trio 3.0T MRI scanner equipped with a 32-channel head coil. T_1-weighted structural images of the whole brain were acquired on the first scan session using a three-dimensional magnetization-prepared rapid acquisition gradient echo pulse sequence (repetition time (TR) 1,620 ms; echo time (TE) 3.09 ms; inversion time 950 ms; voxel size 1 × 1 × 1 mm; matrix size 190 × 263 × 165). A field map was also acquired at each scan session (TR 1,200 ms; TE1 4.06 ms; TE2 6.52 ms; flip angle 60°; voxel size 3.4 × 3.4 × 4.0 mm; field of view 220 mm; matrix size 64 × 64 × 52) to correct geometric distortion caused by magnetic field inhomogeneity. In all experimental runs with a behavioral task, T_2*-weighted images sensitive to BOLD contrasts were acquired using a slice-accelerated multiband echo-planar pulse sequence (TR 2,000 ms; TE 25 ms; flip angle 60°; voxel size 1.5 × 1.5 × 1.5 mm; field of view 192 mm; matrix size 128 × 128 × 80).

In all resting-state runs, T_2*-weighted images sensitive to BOLD contrasts were acquired using a slice-accelerated multiband echo-planar pulse sequence (TR 500 ms; TE 30 ms; flip angle 30°; voxel size 3.0 × 3.0 × 3.0 mm; field of view 192 mm; matrix size 64 × 64 × 48).

Cortical reconstruction and volumetric segmentation of the structural data was performed with the FreeSurfer Software Suite (version 5.3)[52]. Boundary-based registration between the structural and mean functional images was performed with FreeSurfer bbregister[53]. Preprocessing of the resting-state fMRI data was carried out using FEAT (fMRI Expert Analysis Tool) v.6.00, part of the FMRIB Software Library v.6.0 (https://www.fmrib.ox.ac.uk/fsl). The following pre-statistics processing was applied: echo-planar imaging distortion correction using FUGUE[54]; motion correction using MCFLIRT[55] ; slice-timing correction using Fourier-space time series phase-shifting; non-brain removal using BET[56]; grandmean intensity normalization of the entire four-dimensional dataset by a single multiplicative factor; high-pass temporal filtering (Gaussian-weighted least-squares straight line fitting, with $\sigma = 50.0$ s).

Nuisance time series were regressed voxelwise from the preprocessed data. Nuisance regressors included: (1) three translation (X, Y, Z) and three rotation (pitch, yaw, roll) time series derived by retrospective head motion correction (R = (X, Y, Z, pitch, yaw, roll)), together with expansion terms ((R,R2,Rt−1,Rt−1)), for a total of 24 motion regressors[57]; (2) the first 5 principal components of non-neural sources of noise, estimated by averaging signals within white matter and cerebrospinal fluid masks, obtained with the FreeSurfer segmentation tools and removed using the anatomical CompCor method (aCompCor)[58]; and (3) an estimate of a local source of noise, estimated by averaging signals derived from the white matter region located within a 15 mm radius from each voxel, using the ANATICOR method[59]. The global signal was not regressed out of the voxel time series[60–62]. Instead, we followed recent guidelines by removing the local white matter signal and other non-neural sources[63,64].

GLM to extract stimulus responses from BOLD time series. From the BOLD time series of 0.5 Hz, we interpolated the data to obtain a time series corresponding to the frequency of presentation of stimuli during the value judgment session (at 1.75 s intervals). We then used a GLM to obtain the static responses to each of these stimuli, β_i, for 184 stimuli in each sequence (see Fig. 1b). This procedure is repeated from each ROI, such that each stimulus has a β_i from each of the 83 ROIs. Hence, each stimulus can be embedded as a point in the 83-dimensional ROI space. From here we kept the results for the first 140 stimuli shown in each session out of all 184 stimuli, which jointly form a 140-point data cloud or geometric representation in this ROI space (see Fig. 3a). This choice to truncate the data past 140 trials was dictated by the fact that MRI acquisition does not continue past the length of the hemodynamic response for several of the last stimuli, thus providing inadequate data for GLM decoding.

Whole-brain parcellation. For the whole-brain analyses, we subdivided participants' gray matter volume into 83 cortical and subcortical areas in both hemispheres, based on regions assigned from the Lausanne atlas[65]. For a replication of our results on a different whole-brain parcellation, please see the Supplementary Information.

Voxel-level study of brain regions and virtual lesioning approach. We examined ten brain regions: posterior fusiform cortex, ACC, orbitofrontal cortex, lateral occipital cortex and primary visual cortex, each from the left and right hemispheres. We used the group-constrained subject-specific method to define the regions[66]. For each region, a large parcel is defined based on an existing parcellation[67], within which a maximum of 300 voxels with highest object versus scrambled *t*-statistic contrast from an independent localizer were selected. For the lateral occipital and posterior fusiform cortices, the parcels were downloaded from https://web.mit.edu/bcs/nklab/GSS.shtml. This procedure allowed the selection of ROIs that exhibited univariate responses to objects in a participant-specific manner.

We conducted an exploratory analysis using a virtual lesioning approach where we removed brain regions one at a time and then recalculated the separability dimension of the modified representation. In the current study, we report the regions whose absence causes the largest change in the observed correlation between separability dimension and response accuracy across participants (magnitude of z-score > 2 or $P < 0.023$, uncorrected). Since this is a ranking procedure where regions contribute the most, we simply report the regions with the largest deviation from the distribution of contributions from each region[67]. This analysis does not lend itself to a correction for multiple comparisons and is commonly used in examining which brain regions most strongly drive a particular effect[68–70].

Statistics. In general, and throughout the main text, we used Pearson's correlation coefficients to assess the relationships between two variables. To evaluate the statistical significance of these coefficients, we used non-parametric statistical tests rather than assuming or testing for normality. This choice was motivated by the fact that the sample size is such that we are underpowered to make strong claims about normality. The one exception to this non-parametric approach was our use of a multilevel statistical model to assess whether the average dimension of neural representations was related to average learning accuracies. We describe this approach in greater detail in the next section; additional details can be found in the accompanying Nature Research Reporting Summary.

Multilevel statistical model. A multilevel model framework was adopted to accommodate the nested nature of the intensive repeated measures data (4 occasions nested within 19 persons). Repeated measures data contain information on both within-person and between-person information that must be disaggregated appropriately to make both within-person and between-person inferences[71]. This disaggregation was achieved by parameterizing the separability dimension variable into time-invariant (between-person) and time-varying (within-person) versions of the separability dimension variable. We calculated a time-invariant, between-person variable of usual separability dimension as the grand mean-centered individual mean score of the separability dimension, respectively, across all days in the study. Participants with positive values on this between-person variable had greater than usual levels of separability dimension throughout the study compared with other participants in the sample. Participants with negative values on this variable had lower levels of separability dimension. We calculated a time-varying, within-person version of the separability dimension variable as deviations from the between-person mean; thus, zero on this within-person variable indicated days of usual levels of separability dimension, negative values indicated days of fewer than usual levels of separability dimension and positive values indicated days of more than usual levels of separability dimension for each individual.

At level 1 (day-level variables), we constructed the following formal model equation (equation (1)):

$$\begin{aligned}\text{Accuracy}_{it} &= \beta_{0i} + \beta_{1i}\ \text{Day's Separability Dimension}_{it} \\ &\quad + \beta_{2i}\ \text{Linear Time}_{it} + \beta_{3t}\ \text{Quadratic Time}_{it} + e_{it}\end{aligned} \tag{1}$$

where Accuracy$_{it}$ is accuracy for person i on day t; β_{0i} indicates the expected accuracy on day 1; β_{1i} indicates within-person differences in accuracy associated with differences in the day's separability dimension; β_{2i} and β_{3i} test for linear and quadratic slopes of time, respectively; and e_{it} are day-specific residuals that are allowed to be autocorrelated (AR(1)).

Person-specific intercepts and associations (from level 1) are specified (at level 2) in equations (2–5) as:

$$\beta_0 = \gamma_{00} + \gamma_{01}\ \text{UsualDimension}_i + u_{0i} \tag{2}$$

$$\beta_1 = \gamma_{10} + u_{1i} \tag{3}$$

$$\beta_2 = \gamma_{20} \tag{4}$$

$$\beta_3 = \gamma_{30} \tag{5}$$

where γ denotes a sample-level parameter and u denotes residual between-person differences that may be correlated, but are uncorrelated with e_{it}. Parameter γ_{01} indicates how between-person differences in the usual level of brain separability across the 4 d was associated with the usual level of accuracy. The multilevel model was fitted with the 'nlme' package in R using maximum likelihood estimation; incomplete data were treated using assumptions of being missing at random. Statistical significance was evaluated at $\alpha = 0.05$.

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

Data availability

The datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. The code used for the statistical analysis and modeling has been provided as Supplementary Software.

References

49. Giusti, C., Pastalkova, E., Curto, C. & Itskov, V. Clique topology reveals intrinsic geometric structure in neural correlations. *Proc. Natl Acad. Sci. USA* **112**, 13455–13460 (2015).
50. Saggar, M. et al. Towards a new approach to reveal dynamical organization of the brain using topological data analysis. *Nat. Commun.* **9**, 1399 (2018).
51. Ward, G. J. The radiance lighting simulation and rendering system. In *Proc. 21st Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1994* (eds. Schweitzer, D. Glassner, A. & Keeler, M.) 459–472 (ACM, 1994).
52. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis. I. segmentation and surface reconstruction. *Neuroimage* **9**, 179–194 (1999).
53. Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage* **48**, 63–72 (2009).
54. Jenkinson, M. Improving the registration of b0-distorted EPI images using calculated cost function weights. In *Proc. Tenth International Conference on Functional Mapping of the Human Brain* 459–472 (2004).
55. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825–841 (2002).
56. Smith, S. M. Fast robust automated brain extraction. *Hum. Brain Mapp.* **17**, 143–155 (2002).
57. Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S. & Turner, R. Movement-related effects in fMRI time-series. *Magn. Reson. Med.* **35**, 346–355 (1996).
58. Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for bold and perfusion based fMRI. *Neuroimage* **37**, 90–101 (2007).
59. Jo, H. J., Saad, Z. S., Simmons, W. K., Milbury, L. A. & Cox, R. W. Mapping sources of correlation in resting state fMRI, with artifact detection and removal. *Neuroimage* **52**, 571–582 (2010).
60. Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B. & Bandettini, P. A. The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? *Neuroimage* **44**, 893–905 (2009).
61. Saad, Z. S. et al. Trouble at rest: how correlation patterns and group differences become distorted after global signal regression. *Brain Connect.* **2**, 25–32 (2012).
62. Chai, X. J., Castañón, A. N., Öngür, D. & Whitfield-Gabrieli, S. Anticorrelations in resting state networks without global signal regression. *Neuroimage* **59**, 1420–1428 (2012).
63. Power, J. D., Schlaggar, B. L. & Petersen, S. E. Recent progress and outstanding issues in motion correction in resting state fMRI. *Neuroimage* **105**, 536–551 (2015).
64. Murphy, K. & Fox, M. D. Towards a consensus regarding global signal regression for resting state functional connectivity MRI. *Neuroimage* **154**, 169–173 (2017).
65. Daducci, A. et al. The connectome mapper: an open-source processing pipeline to map connectomes with MRI. *PLoS One* **7**, e48121 (2012).
66. Julian, J. B., Fedorenko, E., Webster, J. & Kanwisher, N. An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *Neuroimage* **60**, 2357–2364 (2012).
67. Glasser, M. F. et al. A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
68. Honey, C. J., Kötter, R., Breakspear, M. & Sporns, O. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proc. Natl Acad. Sci. USA* **104**, 10240–10245 (2007).
69. van den Heuvel, M. P. & Sporns, O. Network hubs in the human brain. *Trends Cogn. Sci.* **17**, 683–696 (2013).
70. Hagmann, P. et al. Mapping the structural core of human cerebral cortex. *PLoS Biol.* **6**, e159 (2008).
71. Curran, P. J. & Bauer, D. J. The disaggregation of within-person and between-person effects in longitudinal models of change. *Annu. Rev. Psychol.* **62**, 583–619 (2011).

natureresearch

Corresponding author(s): Danielle S. Bassett

Last updated by author(s): Mar 5, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
□	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Cortical reconstruction and volumetric segmentation of the structural data was performed with the Freesurfer image analysis suite. Boundary-Based Registration between the structural image and the mean functional image was performed with Freesurfer bbregister. Preprocessing of the resting state fMRI data was carried out using FEAT (FMRI Expert Analysis Tool) Version 6.00, part of FSL (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl). The following pre-statistics processing was applied: EPI distortion correction using FUGUE; motion correction using MCFLIRT; slice-timing correction using Fourier-space time series phase-shifting; non-brain removal using BET; highpass temporal filtering (Gaussian-weighted least-squares straight line fitting, with sigma=50.0s).

Data analysis

Simulations were performed in Matlab_R2018b (MathWorks). To calculate linear separability on the binary categories, we used the default validation scheme within the Classification Learner application, retaining the default option of 5-fold cross-validation. This algorithm partitions the data into 5 disjoint sets or folds, chosen randomly but with roughly equal size. For each fold, the algorithms trains a linear SVM using the out-of-fold observations, and then assesses model performance using in-fold data. Next, the average test error is calculated over all folds, to yield the separability for each binary assignment, which is a number between 0 and 1. We repeat this process over all binary assignments to obtain the separability in each case, and take the average separability over all assignments for n types of data (or object identities). The resulting average is the separability dimension.

In the multilevel model, a subject's learning accuracy forms the independent variable. This model was fit with nlme in R version 3.5.1 using maximum likelihood estimation, and incomplete data was treated using assumptions of being missing at random. Statistical significance was evaluated at α = 0.05.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated and analysed during the current study are available from the corresponding author on reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Twenty human participants with normal or corrected vision and no history of neurological disease or psychiatric disorders were recruited for this experiment. No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those reported in previous publications (see Methods). All participants volunteered and provided informed consent in writing in accordance with the guidelines of the Institutional Review Board of the University of Pennsylvania (IRB #801929).
Data exclusions	In the value judgement session of the first day for one participant, the fMRI time series was poorly recorded due to a lack of synchronization between the computer and scanner. Hence this participant was excluded from the analyses, with the other 19 subjects contributing data for the main analyses described in this paper. This exclusion criteria on the grounds of poor data quality was pre-established.
Replication	The study was conducted over twenty human participants for statistical and comparison purposes. No further replication of the experiment was performed, however extensive replication of the analysis was conducted with different modeling choices that confirmed our findings.
Randomization	Participants were randomly assigned to two groups determining the type of feedback that they would receive, however this assignment is not relevant to our study where we treated data from all participants equally.
Blinding	Blinding is not relevant to our study as we treat data from all participants equally.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology
☒	□	Animals and other organisms
□	☒	Human research participants
☒	□	Clinical data

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
□	☒	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Twenty human participants (nine female; ages 19-53years; mean age = 26.7 years) with normal or corrected vision and no history of neurological disease or psychiatric disorders were recruited for this experiment.
Recruitment	Participants were recruited from the departmental participant pool and advertisements in the community. All participants volunteered and provided informed consent in writing. Participants had no prior experience with the stimuli or with the behavioral paradigm.

Ethics oversight

Institutional Review Board of the University of Pennsylvania (IRB #801929)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Magnetic resonance imaging

Experimental design

Design type

Subjects learned the monetary value of 12 novel visual stimuli over the course of four consecutive days. Each day comprised of the following phases: (i) a size judgment task; (ii) a learning phase; (iii) a repetition of the size judgment task; (iv) a value judgment task. A 10-minutes resting-state session preceded the experiments on each day.

Design specifications

Learning phase: A total of 12 scan runs over 4 days were completed by each person (three scans per session), totaling 1584 trials. The sessions were comprised of three scans of 6.6 minutes each, starting with 16.5 seconds of a blank gray screen, followed by 132 experimental trials (2.75 seconds each), and ending with another period of 16.5 seconds of a blank gray screen. Each presentation lasted 2.5 seconds (250 ms inter-stimulus interval).
Value judgement and size judgement tasks: A total of 4 scan runs over 4 days were completed by each person (one scan per session) for the value judgement task, while a total of 8 scan runs over 4 days were completed by each person (two scans per session) for the size judgement tasks. Each scan lasted 5 minutes and 22 seconds (184 trials). Stimuli were back-projected onto a screen viewed by the participant through a mirror mounted on the head coil and sub-tended 4 degrees of visual angle. Each presentation lasted 1.75 seconds (250 ms inter-stimulus interval).

Behavioral performance measures

Learning phase: At any point within a trial, participants entered their responses on a 4-button response pad indicating their shape selection with a leftmost or rightmost button press. The average accuracy in selecting the shape with the highest mean value at each trial gradually improved over the course of the experiment, increasing from approximately 50% (chance) in the first few trials to approximately 95% in the final few trials.
Value and size judgement tasks: At any point within a trial, participants entered their responses on a 4-button response pad indicating their shape selection with a leftmost (least valuable) or rightmost (most valuable) button press, during the value judgement tasks. During the size judgement tasks, these leftmost and rightmost but- ton presses corresponded to smaller and larger shapes respectively.

Acquisition

Imaging type(s)

Blood oxygen level dependent (BOLD) functional MRI data was collected from each participant as they performed the task.

Field strength

3.0 T

Sequence & imaging parameters

T1-weighted structural images of the whole brain were acquired on the first scan session using a three-dimensional magnetization-prepared rapid acquisition gradient echo pulse sequence (repetition time (TR) 1620 ms; echo time (TE) 3.09 ms; inversion time 950 ms; voxel size 1 mm X 1 mm X 1 mm; matrix size 190 X 263 X 165). A field map was also acquired at each scan session (TR 1200 ms; TE1 4.06 ms; TE2 6.52 ms; flip angle 60 degrees; voxel size 3.4 mm X 3.4 mm X 4.0 mm; field of view 220 mm; matrix size 64 X 64 X 52) to correct geometric distortion caused by magnetic field inhomogeneity. In all experimental runs with a behavioral task, T2*-weighted images sensitive to blood oxygenation level-dependent contrasts were acquired using a slice accelerated multi-band echo planar pulse sequence (TR 2,000 ms; TE 25ms; flip angle 60 degrees; voxel size 1.5 mm X 1.5 mm X 1.5mm; field of view 192 mm; matrix size 128 X 128 X 80).

Area of acquisition

A whole brain scan was used.

Diffusion MRI □ Used ☒ Not used

Preprocessing

Preprocessing software

Cortical reconstruction and volumetric segmentation of the structural data was performed with the Freesurfer image analysis suite. Boundary-Based Registration between the structural image and the mean functional image was performed with Freesurfer bbregister. Preprocessing of the resting state fMRI data was carried out using FEAT (FMRI Expert Analysis Tool) Version 6.00, part of FSL (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl). The following pre-statistics processing was applied: EPI distortion correction using FUGUE; motion correction using MCFLIRT; slice-timing correction using Fourier-space time series phase-shifting; non-brain removal using BET; highpass temporal filtering (Gaussian-weighted least-squares straight line fitting, with sigma=50.0s).

Normalization

Grand-mean intensity normalization of the entire 4D dataset by was done with a single multiplicative factor.

Normalization template

MNI152

Noise and artifact removal

Nuisance time series were voxelwise regressed from the preprocessed data. Nuisance regressors included (i) three translation (X; Y;Z) and three rotation (pitch; yaw; roll) time series derived by retrospective head motion correction (R = [X; Y;Z; pitch; yaw; roll]), together with expansion terms ([R,R2,Rt-1,R2t-1]), for a total of 24 motion regressors); (ii) the first five principal components of non-neural sources of noise, estimated by averaging signals within white matter and cerebrospinal fluid masks, obtained with Freesurfer segmentation tools and removed using the anatomical CompCor method (aCompCor); and (iii) an estimate of a local source of noise, estimated by averaging signals derived from the white matter region located within a 15 mm radius from each voxel, using the ANATICOR method. Global signal was not

	regressed out of voxel time series. Instead, we follow recent guidelines by removing local white-matter signal and other non-neural sources.
Volume censoring	No volume censoring was performed.

Statistical modeling & inference

Model type and settings	We used data science methods (linear classifiers) for analysis, as well as non-parametric permutation tests, a multilevel statistical model and an ANOCOVA test to test the global findings.
Effect(s) tested	Linear separability of the neural data according to stimuli labels were calculated and the statistical significance of individual variability in the results; no factorial designs were used.

Specify type of analysis: □ Whole brain □ ROI-based ☒ Both

Anatomical location(s)	We use the Group-Constrained Subject-Specific (GSS) method for defining the regions. For each region, a large parcel is defined based on an existing parcellation, within which a maximum of 300 voxels with highest object-versus-scrambled t-statistic contrast from an independent localizer were selected. For lateral occipital and posterior fusiform, the parcels were down-loaded from http://web.mit.edu/bcs/nklab/GSS.shtml). These are described in Methods.

Statistic type for inference (See Eklund et al. 2016)	Voxel-wise and cluster-wise inference was not used.
Correction	Permutation tests were used for non-parametric statistical testing, and multiple comparisons was used to evaluate the resulting p-values where there were multiple regions or days considered.

Models & analysis

n/a	Involved in the study
☒	□ Functional and/or effective connectivity
☒	□ Graph analysis
□	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis	Simulations were performed in Matlab (MathWorks). To calculate linear separability on the binary categories, we used the default validation scheme within the Classification Learner application, retaining the default option of 5-fold cross-validation. This algorithm partitions the data into 5 disjoint sets or folds, chosen randomly but with roughly equal size. For each fold, the algorithms trains a linear SVM using the out-of-fold observations, and then assesses model performance using in-fold data. Next, the average test error is calculated over all folds, to yield the separability for each binary assignment, which is a number between 0 and 1. We repeat this process over all binary assignments to obtain the separability in each case, and take the average separability over all assignments for n types of data (or object identities). The resulting average is the separability dimension. In the multilevel model, a subject's learning accuracy forms the independent variable. This model was fit with nlme in R using maximum likelihood estimation, and incomplete data was treated using assumptions of being missing at random. Statistical significance was evaluated at α = 0.05.