Neural representational geometry correlates with behavioral differences between monkeys

Valeria Fascianelli[1,2], **Fabio Stefanini**[1,2], **Satoshi Tsujimoto**[3], **Aldo Genovesio**[4*], **and Stefano Fusi**[1,2,5,6*]

[1] Center for Theoretical Neuroscience, Columbia University, New York, NY 10027, USA
[2]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027, USA
[3]SixthFactor Pte. Ltd, Singapore
[4]Department of Physiology and Pharmacology, Sapienza University of Rome, Rome, Italy
[5]Department of Neuroscience, Vagelos College of Physicians and Surgeons, Columbia University Irving Medical Center, New York, NY 10027, USA
[6]Kavli Institute for Brain Science, Columbia University, New York, NY 10027, USA
[*]Correspondence: aldo.genovesio@uniroma1.it (A.G.), sf2237@columbia.edu (S.F.)

Abstract

Animals likely use a variety of strategies to solve laboratory tasks. Traditionally, combined analysis of behavioral and neural recording data across subjects employing different strategies may obscure important signals and give confusing results. Hence it is important to develop techniques that can infer strategy at the single-subject level. We analyzed an experiment in which two monkeys perform a visually cued rule-based task. From the analysis of their performance there is no indication that they used a different strategy. However, when we examined the geometry of stimulus representations in the state space of the neural activities recorded in dorsolateral prefrontal cortex, we found striking differences. Our purely neural results predict behavioral differences that we observed by analyzing the reaction times. These analyses provide strong support that the animals employed different strategies. Finally, we used a modeling study to correlate these strategies with the amount of training that the animals received.

Keywords: representational geometry; abstraction; disentangled representations; individual differences; behavioral differences; strategy; dorsolateral prefrontal cortex

Introduction

Although the tasks designed in a laboratory are relatively simple and they are performed in highly controlled situations, different animals can still adopt different strategies to solve the same task. It is surprisingly difficult to reproduce the exact same behavior in different laboratories, even when the training protocol, the experimental hardware, software, and procedures are standardized [1]. In many situations, it is also possible that the behavioral performance is the same, but the strategy used to perform the task is different. Consider, for example, a task in which multiple stimulus properties must be mapped onto appropriate behavioral responses. Such a task can be accomplished by rote learning of this map, but if the task involves structure across stimulus attributes, such as irrelevant stimulus features, learning can be simplified by adopting more "intelligent" strategies that exploit this structure. All these strategies may produce the same level of task performance, so how can we distinguish among them?

Here we show that this can be done by examining the geometry of stimulus representations in the state space of recorded neural activities. The recorded neural responses are typically very diverse and seemingly disorganized [2, 3, 4, 5]. However, when the neural activity is analyzed at the population level, it is often possible to identify interesting and informative "structures". In particular, the analysis of the geometry of the neural representations has recently revealed that some variables are represented in a special format which enables generalization to novel situations [6]. The representational geometry is defined by the set of distances between points that represent different experimental conditions in the neural activity space. The set of points of all the conditions of the experiment defines an object that has specific computational properties [7] and it is often preserved across subjects [8]. For example, if the points define a high dimensional object (in this article we always consider the embedding dimensionality [9] when we speak about dimensionality), then a linear decoder can separate the points in a large number of different ways, permitting a downstream neuron to perform many different tasks [2, 3]. If instead the points define a low dimensional object, the representations allow a simple linear decoder of one variable to generalize across the values of other variables [6]. These representations have been called abstract because of their generalization properties and they are known as disentangled representations in the machine learning community [10, 11]. Abstract representations have been observed in several brain areas [6, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22] but it is still unclear whether they correlate with behavior. As pointed out by Krakauer et al. [23], a new conceptual framework that meaningfully maps the neural data to behavior is necessary to better understand the brain-behavior relationship, and to accomplish that, the analysis of the behavior should be as fine-grained as the analysis performed on the neural data.

Here we show that differences between neural representational geometries across subjects predict significant differences in their behavior, providing evidence that the aspects of the representational geometry that we typically study can affect behavior. Thanks to these correlations, the analysis of the representational geometry is also an important tool for reliably predicting and interpreting individual differences in the behavior.

More specifically, we analyzed the activity of neurons recorded in dorsolateral prefrontal cortex (PFdl) of two monkeys performing a visually cued rule-based task [24]. The task required choosing between two spatial targets based on the rule cued by a visual stimulus, either staying with the same response as in the previous trial (after a stay cue) or shifting to the alternative response (after a shift cue). The task average performance was the same for the two monkeys.

We studied systematically specific aspects of the geometry of the neural representations. First, we looked at the ability of a decoder to classify all the task relevant variables (the shape of the visual cue, the rule, the current and the past responses). Then we tried to decode the variables that correspond to all the possible ways of dividing the conditions into two groups of equal size (balanced dichotomies). Some of these dichotomies correspond to obvious task relevant variables, while some others are still interpretable but do not have a simple label. For all these dichotomies, we also computed the cross-condition generalization performance (CCGP) by training a decoder on a subset of conditions and testing on a different subset. These other conditions are completely novel for the decoder, and hence a high CCGP means that the geometry allows for generalization. Studying which variables have an elevated CCGP allowed us to identify which variables were represented in an abstract format and therefore describe another important aspect of the representational geometry [6]. This set of measures revealed that the representational geometry is strikingly different for the two monkeys. This finding brought us to reanalyze the behavior, and we discovered that the reaction times actually reflect the different geometries.

Our study shows that it is possible to find individual differences in the strategy used to perform a task by examining the representational geometry. Moreover, the fact that the geometry is related to the observed behavior indicates that the geometry is probably important for performing the task.

Results

We analyzed single unit recordings in dorsolateral prefrontal cortex (PFdl) of two male rhesus monkeys. As the main message of this work is that the representational geometry can explain the differences in behavior of the two monkeys, we will present the neural and behavioral results for each monkey separately. We refer to them as Monkey 1 and Monkey 2.

Both monkeys were trained to perform a visually cued rule-based task (Figure 1A). The task was to choose one of two targets, with a saccadic movement, according to the rule instructed in each trial by one of four possible visual cues (Figure 1B). Two cues instructed the monkey to “stay” with the target chosen in the previous trial, while the other two cues instructed to “shift” to the alternative target. In each trial, the visual cue was randomly chosen. At the time of the recordings, both animals were already trained and they were performing the task with the same high accuracy.

When we analyzed the geometry of the neural representations recorded during the task, we found significant differences between the two monkeys. The representational geometry is defined by the set of distances between the points in the firing rate space that represent different conditions (see e.g. [25]). This is a relatively large set of variables, which are not defined in a unique way as there are several reasonable measures of distances in the presence of noise. We found significant differences between the two monkeys by focusing on two particular aspects of the geometry that also have the advantage of being cross-validated and interpretable: the first is the set of linear decoding accuracies for the task relevant variables and all the other variables that correspond to balanced dichotomies of the conditions (i.e. all the possible ways of dividing the conditions into two equal groups). The task relevant variables are the previous response, the rule, the current response, and the shape of the visual cue (Figure 1D). The latter identifies whether the visual cue is a rectangle or a square, although the cue differs also because the rectangles are grey and the squares are colored (yellow and purple). The second aspect of the geometry is related to the ability of a linear classifier to generalize across conditions when trained to decode the balanced dichotomies (cross-condition generalization performance or CCGP [6]).

The decoding accuracy is directly related to the distance between two groups of points, and in this respect, it is a geometrical measure. It is better than the average distance because it is cross-validated and it takes into account the structure of the noise, similarly to the Mahalanobis distance [26]. Moreover, it is interpretable because it tells us something about the variables that are represented. The second quantity, the CCGP, is more sensitive to the angles between coding directions, another aspect of the representational geometry: the ability of a linear classifier to generalize depends on the parallelism of the coding directions [6]. Say we consider two binary variables x and y, and we train a decoder to report the value of variable x from the patterns of neural activity. If this decoder is trained only in the situations in which $y = y_1$, it is not guaranteed that it would work right away for a different value of y, say $y = y_2$. In order to generalize to $y = y_2$, it is necessary that the coding direction of x (i.e. the direction from the points corresponding to neural activities when $x = x_1$ to $x = x_2$) for $y = y_1$ is approximately the same as for $y = y_2$. CCGP also takes into account the noise structure and it is cross-validated. Moreover, if a variable has high CCGP it means that the variable is encoded in a special format, that we formerly defined as “abstract”[6]. The variable is encoded in an abstract format (or simply it is abstract) because the coding direction does not depend on the specific instance. This guarantees special generalization properties (cross-condition), which are the hallmark of abstraction.

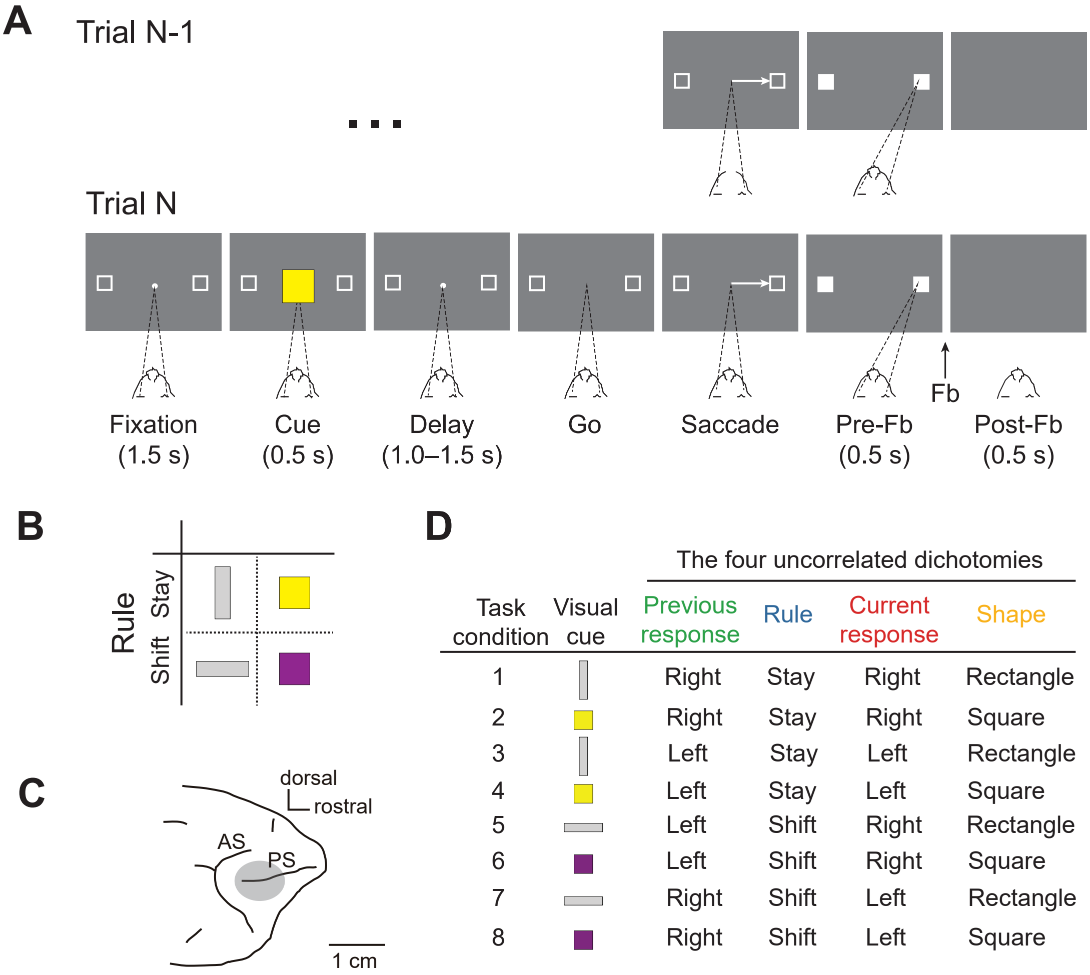
Differences between the representational geometries of the two monkeys

Our neural database consists of 289 and 262 neurons from Monkey 1 and Monkey 2, respectively. To investigate which task variables can be decoded, we built pseudo simultaneous trials (pseudo trials) for each monkey separately (see Methods). We defined the pseudo trial as the combination of spike counts randomly sampled from different trials of the same task condition [2]. For each neuron the spike count was estimated in a $200ms$ time bin. We considered only neurons recorded for at least 5 complete and correct trials in each task condition for a total of 205/289 (71%) neurons for Monkey 1 and 188/262 (72%) neurons for Monkey 2.

We found that, in Monkey 1, almost all the dichotomies can be linearly decoded during the cue presentation (Figure 2A), but not all of them are in an abstract format, i.e. with a high CCGP (Figure 2B). Shape is the variable with the highest CCGP, followed by the current response, while the previous response and the rule can be decoded but do have a CCGP at chance and hence are not in an abstract format. In Monkey 2, almost all the dichotomies can be linearly decoded during the cue presentation, except for the shape and the previous response (Figure 2C). The CCGP analysis reveals that, in Monkey 2, the rule is in an abstract format with the highest CCGP, differently from Monkey 1 (Figure 2D). In both monkeys, instead, during the cue presentation, the current response is in an abstract format, while the previous response is not.

To better highlight the differences in the representational geometry, we focused our analysis on the $300ms$ time window in which the differences are large (from $200ms$ after the cue onset until the cue offset, grey vertical shade in Figure 2). The beeswarm plots in Figure 3A show the decoding accuracy and CCGP for all the possible dichotomies in the $300ms$ time window for Monkey 1 and Monkey 2. It is evident that Monkey 1, during the cue presentation, represents the shape of the visual cue in an abstract format (highest CCGP), while rule is not abstract (CCGP at chance), even though it can be decoded (Figure 3A, left). Rule becomes abstract only later in the delay period after the cue offset (Figure 2B). Instead, for Monkey 2, the rule is the variable with the highest CCGP, while the shape of the visual cue is not abstract (CCGP at chance) (Figure 3A, right). Moreover, both monkeys represent the current response in an abstract format, but not the previous response. Interestingly, in both monkeys the current response is not abstract from the time when it can be decoded, but only slightly later (see Figure 2). These results suggest that Monkey 1 is grouping together the cues with the same shape, and hence it is using a strategy based on the identity of individual visual stimuli. Instead, Monkey 2 is using a more “cognitive” strategy because rule is the variable with the highest decoding accuracy and CCGP, and hence Monkey 2 is grouping together the visual cues that correspond to the same rule, despite the fact that they are visually very different.

Shape cannot be decoded in Monkey 2 using a linear classifier. We were wondering whether it is not encoded at all, or it could be decoded using other decoders. We decided to consider pairs of conditions separately, which is equivalent to consider non-linear decoders for all the



The four uncorrelated dichotomies

Task condition	Visual cue	Previous response	Rule	Current response	Shape
1	▯	Right	Stay	Right	Rectangle
2	🟨	Right	Stay	Right	Square
3	▯	Left	Stay	Left	Rectangle
4	🟨	Left	Stay	Left	Square
5	▭	Left	Shift	Right	Rectangle
6	🟪	Left	Shift	Right	Square
7	▭	Right	Shift	Left	Rectangle
8	🟪	Right	Shift	Left	Square

Fig. 1: **Behavioral task, visual cues, recording site, and task conditions. A)** Example of two consecutive trials of the visually cued rule-based task with temporal ordering of task events from left to right. The dark gray rectangle represents the video screen as viewed by the monkey. The target of the monkey's gaze is indicated by dashed lines. In this example, the trial N is a stay trial instructed by the yellow square, requiring the monkey to choose the same right target chosen in the previous trial N-1. Fb, Feedback. **B)** Visual cues presented to the monkey. Each visual cue instructed the rule to be applied: the vertical and yellow square instructed the stay rule; the horizontal rectangle and purple square instructed the shift rule. **C)** Recording area in dorsolateral prefrontal cortex. AS, Arcuate Sulcus; PS, Principal Sulcus. **D)** List of the eight task conditions defined as the combination of the four main uncorrelated dichotomies: previous response (green), rule (blue), current response (red), shape (orange). The color code of the four dichotomies is conserved across all the figures.

points. Indeed, if two conditions are sufficiently separated, i.e. the distance between the corresponding points is large enough compared to the noise, then a linear decoder should work. This is true even when the dichotomy is not linearly separable, for example in the case of XOR for four points that define a low dimensional object like a square: a linear decoder would not be able to separate the two points on the diagonal from the other two, but it would separate all pairs of points, if taken one pair at the time. In addition to considering pairs of points, we denoised the data by projecting the neural activity of a single pseudo trial into a lower dimensional space (3D) using the multi-dimensional scaling technique described in the Methods. Using this procedure, we found that the shape can be decoded in both monkeys. In particular, in Monkey 1 (Figure 3B, left) the shape can be decoded for both rule conditions with high accuracy. This was expected as the shape was already linearly decodable for all the points without denoising (decoding accuracy in Figure 3A, left). Shape could be decoded also in Monkey 2, in both rule conditions (Figure 3B, right). These results show that both monkeys PFdl neurons encode the shape of the visual stimulus as required to perform the task with high accuracy, but with different geometries, making the shape linearly separable and in an abstract format only in one of the monkeys.

To visualize the different geometries of the two monkeys, we used Multi-Dimensional Scaling (MDS) transformation to reduce the dimensionality of the original representations. More specifically, we used MDS on the dissimilarity matrix containing the Euclidean distances between the average activity of two task conditions normalized by the variance along the direction that goes from one condition to the other (see Methods). Each point in the MDS plots is the average firing rate of each task condition in a $300ms$ time window during the cue presentation (Figure 4). For each monkey, we highlighted the different dichotomies (groups of conditions) by drawing lines between the conditions that are in the same group. In particular, shape and the current response are in an abstract format in Monkey 1, while rule and current response are abstract in Monkey 2. For both monkeys, the current response is in an abstract format, while none of the two has the previous response in an abstract format.

Behavioral differences between monkeys reflect differences in the geometry

The differences in the representational geometry are so striking that they induced us to reanalyze the behavior to look for more subtle individual differences. We analyzed 65 and 77 sessions for Monkey 1 and Monkey 2, respectively. As already mentioned, we did not find any significant difference in the overall behavioral performance between the two monkeys (chi-square test, p-value=0.93; Figure 5A left). However, a significant difference emerged in the average reaction times (Mann-Whitney U test, p-value=10^{-15}; Figure 5A right) when the conditions were grouped as suggested by the differences in the representational geometry. Indeed, the neural analysis revealed that the shape is in an abstract format for Monkey 1 and the rule is abstract for Monkey 2. We computed the average behavioral performance for each condition separately, and then we grouped the correct trials according to shape (rectangle and square) and rule (stay and shift). There is not a significant difference in the behavioral performance between different shapes (chi-square test: p-value=0.78 in Monkey 1, Figure 5B left; p-value=0.06 in Monkey 2, Figure 5D left) and rules (chi-square test: p-value=0.11 in Monkey1, Figure 5B right; p-value=0.22 in Monkey 2, Figure 5D right) in both monkeys. Nevertheless, a significant difference in reaction times emerged across conditions in each monkey. In particular, Monkey 1, with the shape in an abstract format, has an average reaction time that significantly changes with the shape of the visual cue (Mann-Whitney U test: p-value = 0.002; Figure 5C, left) regardless of the rule (Mann-Whitney U test: p-value = 0.05; Figure 5C, right). On the opposite, Monkey 2, with the rule in an abstract format, shows an average reaction time that significantly changes with the rule (Mann-Whitney U test: p-value = 10^{-10}; Figure 5E right) regardless of the shape (Mann-Whitney U test: p-value = 0.28; Figure 5E left).

The differences in reaction times are significant and they nicely reflect the geometry, but they are relatively small. So we decided to further investigate the behavior to see whether these differences could be predicted by looking at the recent series of events and monkey responses. In particular, we fitted a multi-linear regression model to predict the reaction time on a trial by trial basis using three factors: the previous response, the shape of the visual stimulus, and the rule. We also considered all the interaction terms (see supplementary Figure 1). We found that the rule factor has a stronger weight in predicting reaction times in Monkey 2 than in Monkey 1 (Mann-Whitney U test: p-value=10^{-34}; Figure 5F). Viceversa, the shape is a stronger factor in predicting the reaction time of Monkey 1 (Mann-Whitney U test: p-value=10^{-34}; Figure 5F). Supplementary Figure 1 shows that the strongest factor in predicting the reaction time is the interaction of the previous response and the rule in both monkeys, because the combination of these two factors is essential to choose the correct response.

We also asked whether there is a relation between reaction time and neural results for other dichotomies. We focused on the four uncorrelated dichotomies that correspond to task relevant variables (shape, rule, current and previous response). In Figure 6 we plotted the

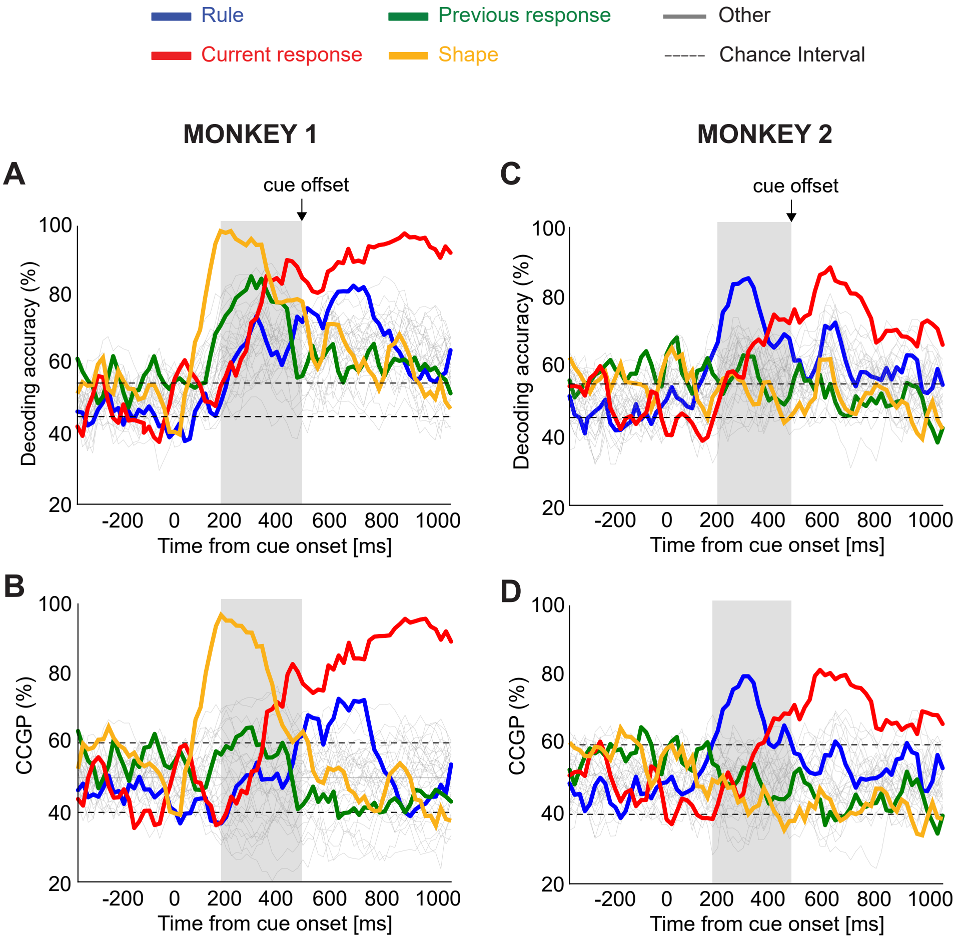


Fig. 2: **Decoding accuracy and CCGP as a function of time**. Time is aligned to the cue onset (time 0) and the presentation lasts 500ms (until the time of the cue offset indicated by the vertical black arrow). The horizontal dashed lines are ±2 standard deviations of 100 cross validations distribution obtained from null models. The grey vertical shade indicates the time bin starting at 200ms after cue onset until cue offset, in which we found a maximal difference between the neural representations of the two monkeys. **A)** Decoding accuracy of all the possible 35 dichotomies (i.e. all variables that correspond to grouping the conditions into two equal size groups) in Monkey 1. During the cue presentation, most of the dichotomies can be decoded, in particular all the main task variables indicated with different colors. The shape of the visual stimulus (orange) can be decoded with the highest accuracy, followed by the previous response (green), the current response (red), and the rule (blue). **B)** CCGP of the 35 dichotomies in Monkey 1. During the cue presentation, shape (orange) is in an abstract format with the highest CCGP, followed by the current response (red). Rule (blue) is not abstract during the cue presentation, but it becomes significantly different from chance after the cue offset. The previous response (green) is not in an abstract format. **C)** Decoding accuracy for Monkey 2. During the cue presentation, the rule (blue) and the current response (red) can be decoded. **D)** CCGP for Monkey 2. Differently from Monkey 1, rule (blue) is in an abstract format with the highest CCGP during cue presentation, followed by the current response (red). Shape (orange) and previous response (green) are not in an abstract format.

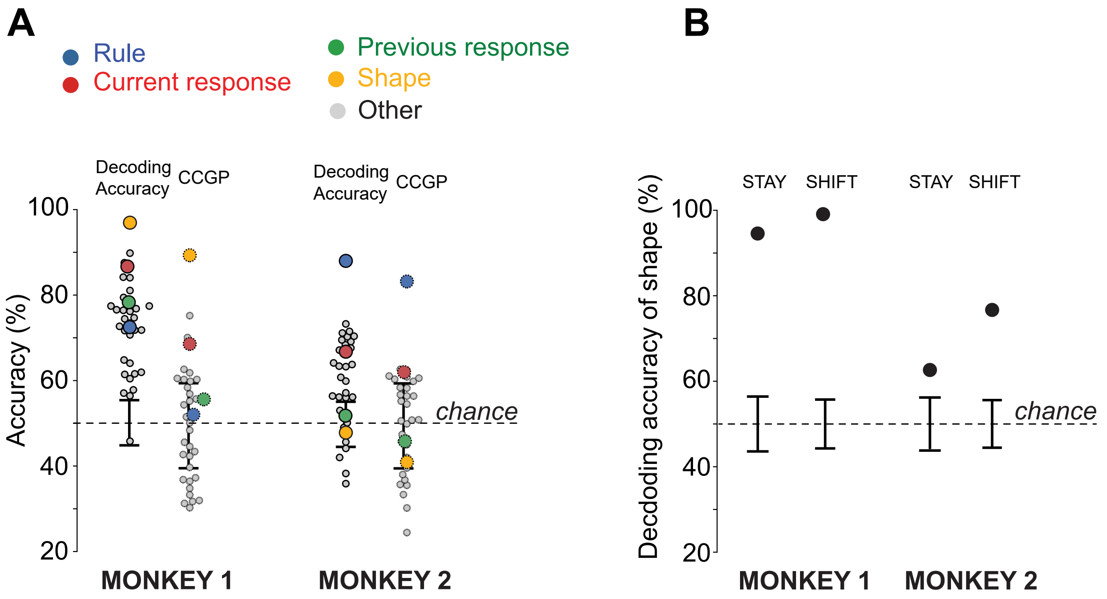


Fig. 3: **Summarizing the features of the representational geometry: decoding accuracy and CCGP during last 300ms of cue presentation.** **A)** Decoding accuracy (continuous-edge circles) and CCGP (dashed-edge circles) for each of the 35 dichotomies in a 300ms time bin during cue presentation in Monkey 1 (left) and Monkey 2 (right). Each circle is a different dichotomy. The four main dichotomies corresponding to task variables are highlighted with different colors. All the other dichotomies are in grey. Black error bars are the ±2 standard deviations around the chance level obtained from null models. **B)** On the left, the decoding accuracy of shape in Monkey 1 in the stay (left), and shift (right) rule. The linear decoder was trained after projecting the neural activity of each pseudo trial in a lower dimensional space using a Multi Dimensional Scaling transformation. In Monkey 1, shape can be decoded in both rule conditions. On the right, the same plot as on the left, but for Monkey 2. As for Monkey 1, shape can be decoded in both rule conditions, though the performance is lower. In Monkey 1, the variance explained by the MDS is 62% and 65%, for stay and shift, respectively. In Monkey 2, it is 62% and 54%, for stay and shift, respectively.

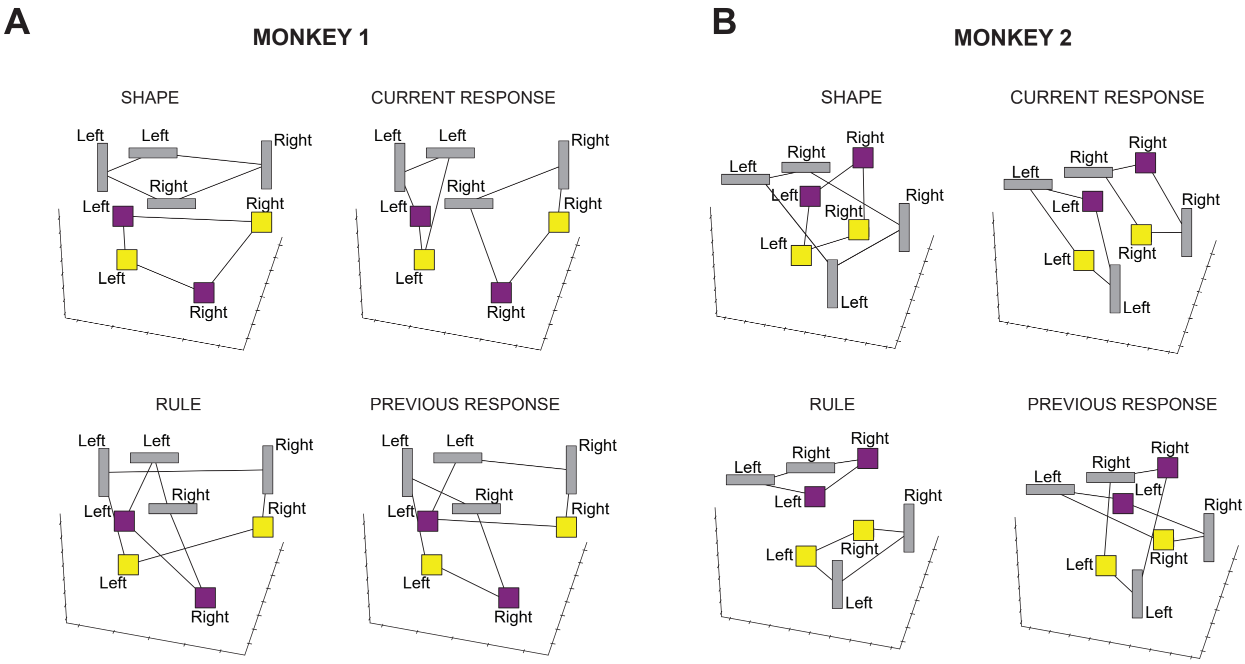


Fig. 4: **3-D Multi-Dimensional Scaling (MDS) plots**. Each point represents the average firing rate in the 300ms time window during the cue presentation for one of the eight task conditions. The visual cue is shown along with the current response. The connecting black lines highlight the dichotomy indicated in the label above each plot. Shape and rule are the two dichotomies that mostly characterize the difference in representational geometry between the two monkeys, while the current response is in an abstract format for both monkeys. The previous response is not abstract in neither of the monkeys. **A**) 3-D MDS plots in Monkey 1 for shape (top-left), current response (top-right), rule (bottom-left), and previous response (bottom-right). **B**) 3-D MDS plots in Monkey 2 for shape (top-left), current response (top-right), rule (bottom-left), and previous response (bottom-right).

differences in reaction time between the two values of each variable as a function of the decoding accuracy and CCGP during the last $200ms$ of the stimulus presentation (notice that this time interval is different from that analyzed in the previous figures). Although in this time interval the monkeys have still to initiate their motor response, it is likely that they already made a decision. We chose this interval because it is the one that most clearly shows the relation between the neural geometry (the decoding accuracy and CCGP for four dichotomies) and reaction time. We found that in both monkeys, there is a trend: as the decoding accuracy and CCGP increase, the difference in reaction time also increases (see Figure 6). Interestingly, for both monkeys the dichotomy with the largest CCGP/decoding accuracy corresponds to the variable encoding the current response. This is reflecting a bias in the reaction time for the left and right responses (Monkey 1: $(300 \pm 1)ms$ and $(328 \pm 1)ms$ for right and left, respectively; Monkey 2: $(297 \pm 1)ms$ and $(325 \pm 1)ms$ for right and left, respectively; (mean ± SEM)). Interestingly, the bias is the same for the two monkeys. The other dichotomies, that in this interval are less strongly encoded, are ranked as in previous intervals, and reveal again the main difference between the geometries of the representations recorded in the two animals. Indeed, shape (orange circle) and rule (blue circle) variables are flipped in the rank in the two monkeys. This observation confirms what we already noticed in the previous analyses, but it also shows that there is a hierarchy of dichotomies that all seem to affect some aspect of the behavior.

The representational geometry in an artificial neural network at different learning stages

To better understand the origin of the differences in representational geometry between the two monkeys, we trained a two layers feed-forward neural network to perform the visually cued rule-based task, and we analyzed the activations of the units in the second hidden layer during training. We know from previous studies that these simple feed-forward networks can easily generate abstract representations [6, 27]. The input to the network is the visual cue encoded by two one-hot vectors of three units each, and the previous response encoded by one one-hot vector of two units (Figure 7A). This input is passed through two hidden layers with 100 Rectified Linear units each, and the output of the network is the current response encoded by one one-hot vector of two units. We applied the same analysis framework used to study the neural data to the activations of the units of the second hidden layer. After ∼50 training epochs, the network performed the task with 100% accuracy (black curve, Figure 7B). At this training stage, all the main task variables can be decoded with high accuracy (Supplementary Figure 2). It is now interesting to focus on the changes of the representational geometry of the main task variables revealed by the CCGP analysis (colored curves, Figure 7B). We selected two training periods: Period 1 is defined as the set of epochs where the training performance is between 90% and 100%; Period 2 is the range of epochs from 70 to 100 where the training performance is constantly at 100% (vertical grey bars in Figure 7B). We observed that in Period 1, during the early phase of the high performance period, all dichotomies can be decoded but only shape is in an abstract format with the highest CCGP followed by the previous response (Figure 7C left). It is worth noticing that, albeit the rule and the current response can be decoded, they are not in an abstract format. Shape is abstract from the very beginning, indicating that it is abstract already in the input. This is due to the assumption that shape is abstract in the input because it is encoded by a population of highly specialized neurons. This is not an unreasonable assumption given that the monkeys most likely had already been familiar with numerous different shapes and that they created an abstract representation before the beginning of the recordings. CCGP and decoding accuracy increase because initially the weights are random, and the signal that can reach the second intermediate layer is relatively weak. It is only with learning that the signal increases, though initially the change in the geometry is modest, mostly due to the stretched distances between different conditions. In Period 2 all dichotomies can be decoded but now rule is in an abstract format along with previous and current response (Figure 7C right). Shape, instead, is decoded with high accuracy but it is not in an abstract format any longer. The representational geometry in Period 1 resembles the neural representation of Monkey 1 where shape is in an abstract format with the highest CCGP, while the representational geometry in Period 2 resembles the neural representation of Monkey 2 where rule is in an abstract format. The model also captured that the CCGP of shape goes from above chance to significantly below chance along training. The previous response, instead, is a variable that is represented in the model in a different way, since in both monkeys it is not in an abstract format. This is probably due to our simplifying assumption that previous response is a variable that is completely disentangled from the visual cue.

Discussion

Traditionally, studies on the primate brain focused on the features of the recordings that are conserved across monkeys. It is uncommon to report and discuss differences between monkeys and other animals often be-

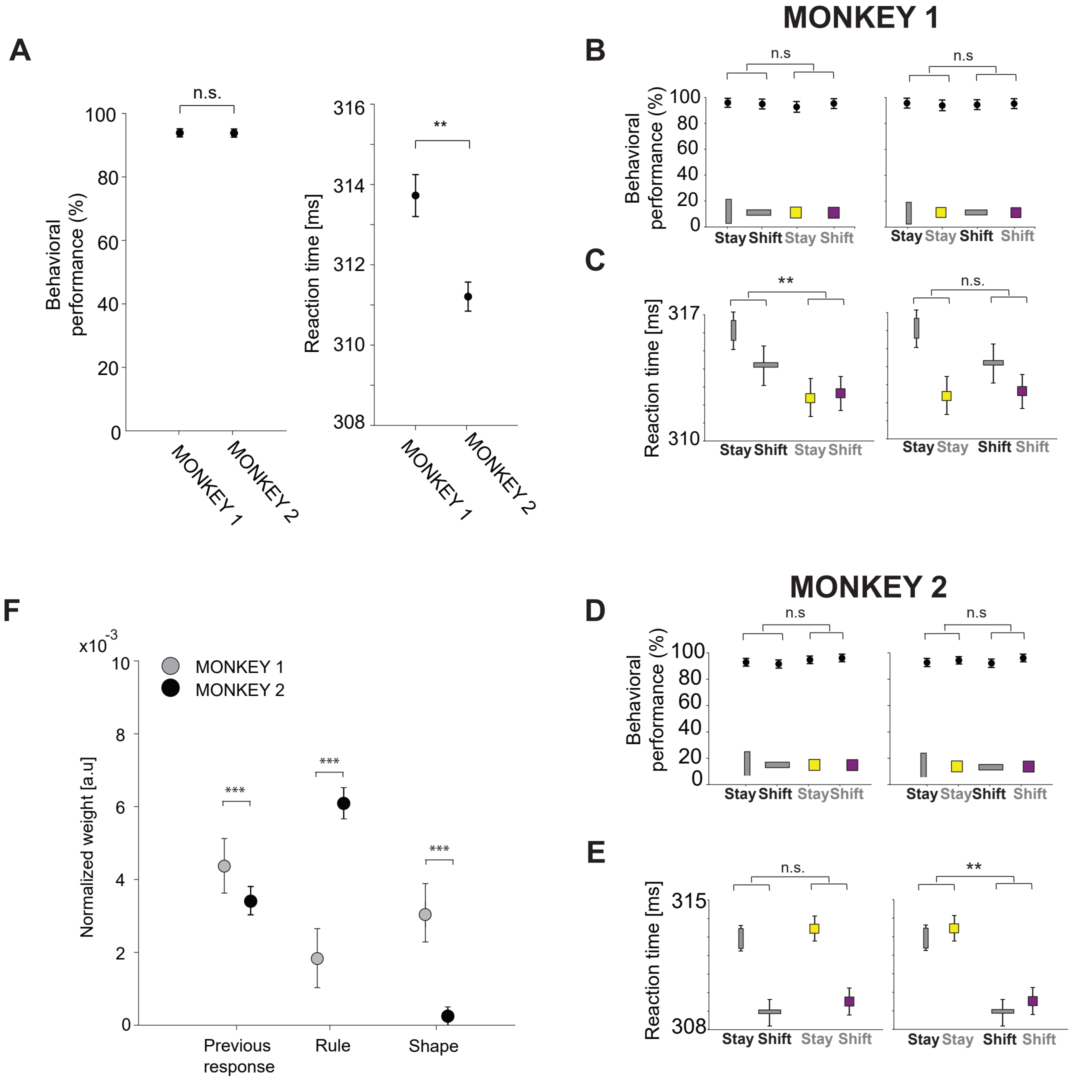


Fig. 5: **Behavioral performance, reaction times of the monkeys and a multi-linear regression behavioral model reflecting differences in the geometries.** **A)** Left: average behavioral performance across sessions for Monkey 1 and Monkey 2. Both monkeys performed the task with high accuracy. The error bars indicate the confidence interval at 95% of confidence level. Right: average reaction time across sessions for Monkey 1 and Monkey 2. A significant difference emerged in the average reaction time between the two monkeys. The error bars are the standard error of the mean. n.s. not significant: chi-square test, p-value>0.05. **: Mann-Whitney U test, p-value<0.01. **B)** Mean behavioral performance across sessions for Monkey 1 computed separately for each rule and shape. The x-axis indicates rule, and y-axis is the mean performance averaged across sessions. The visual cue of each condition is indicated at the bottom of the plot. On the left(right) the visual cue order reflects shape(rule). n.s. not significant: chi-square test, p-value>0.05. **C)** Mean reaction time across sessions for Monkey 1. As in B), the x-axis indicates rule, and y-axis is the reaction time averaged across sessions. The error bar are the standard error of the mean. n.s. not significant: Mann-Whitney U test, p-value>0.05; **: Mann-Whitney U test, p-value<0.01. **D)** The same as in B) but for Monkey 2. **E)** The same as in C) but for Monkey 2. **F)**Weights of the three independent factors predicting the reaction time of single trial in a multi-linear regression model. Weights are normalized to the maximum weight that is the previous response and rule interaction term in both monkeys (see Supplementary Figure 1). The error bars are the 2 standard deviations of weights across of 100 models. The variance explained (r-squared) by the models is 12% and 18% for Monkey 1 and Monkey 2, respectively. ***: Mann-Whitney U test, p<0.001.

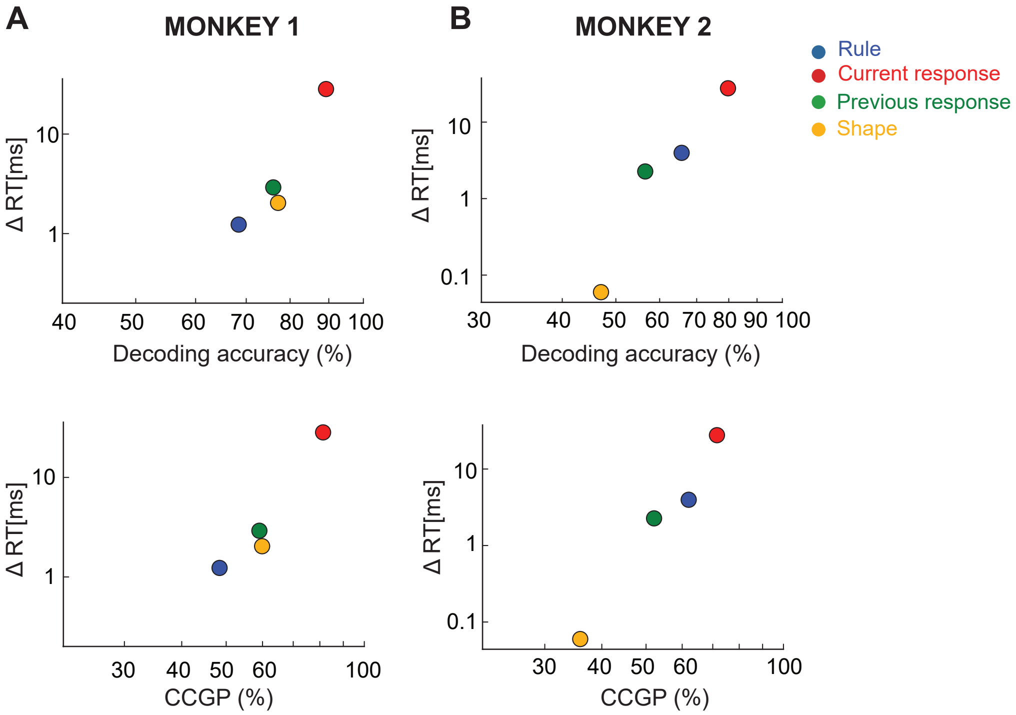


Fig. 6: **Reaction times versus neural results for four uncorrelated dichotomies that correspond to task relevant variables. A)** Top: Decoding accuracy versus difference in reaction times (Δ RT) for each of the four uncorrelated dichotomies in Monkey 1. The decoding accuracy is computed in the last $200ms$ before the cue offset, notice that this is not the same time interval as the one studied in Figure 3A. As the decoding accuracy increases, the difference in reaction time also increases. Bottom: CCGP versus difference in reaction time in Monkey 1. **B)** Top: decoding accuracy versus difference in reaction times for Monkey 2. There is a trend between the two variables, and the main difference with Monkey 1 comes from shape (orange circle) and rule (blue circle) which are now flipped in the rank. Bottom: CCGP versus difference in reaction times for Monkey 2.

cause it is difficult to study and interpret them. Here we showed that it is possible to find clear differences between the representational geometry of two monkeys, and that they correlate with subtle but significant differences in the behavior. One of the advantages of our approach, based on the analysis of the representational geometry, is that it allowed us to study systematically many different interpretable aspects of the geometry of the representation that potentially cause different behaviors. To characterize the representational geometry, we considered the decoding accuracy and the cross-condition generalization performance for every possible dichotomy of the experimental conditions. The number of dichotomies grows rapidly with the number of conditions, almost exponentially for balanced dichotomies (using Stirling approximation $\sim 2^C/\sqrt{2\pi C}$ where C is the number of conditions). Even though some of the dichotomies are correlated (the full characterization of the geometry requires only $\sim C^2$ numbers), we can still examine systematically a large number of potentially different behaviors. Moreover, the dichotomies are interpretable and, often, they correspond to some of the task key variables. This is the case in our analysis, in which the dichotomies with the largest decoding accuracy and CCGP during the cue presentation correspond to shape for one monkey and rule for the other. These dichotomies suggested a way to compute the reaction time for different groups of conditions, and revealed significant differences in the behavior, which were not detectable from the initial analysis of the bare performance.

The analysis of the geometry revealed that there is an interesting "structure" in the arrangement of the points that represent different conditions in the firing rate space: for one monkey shape is an important abstract variable (a more "visual" monkey) and for the other it is the rule (a more "cognitive" monkey). This

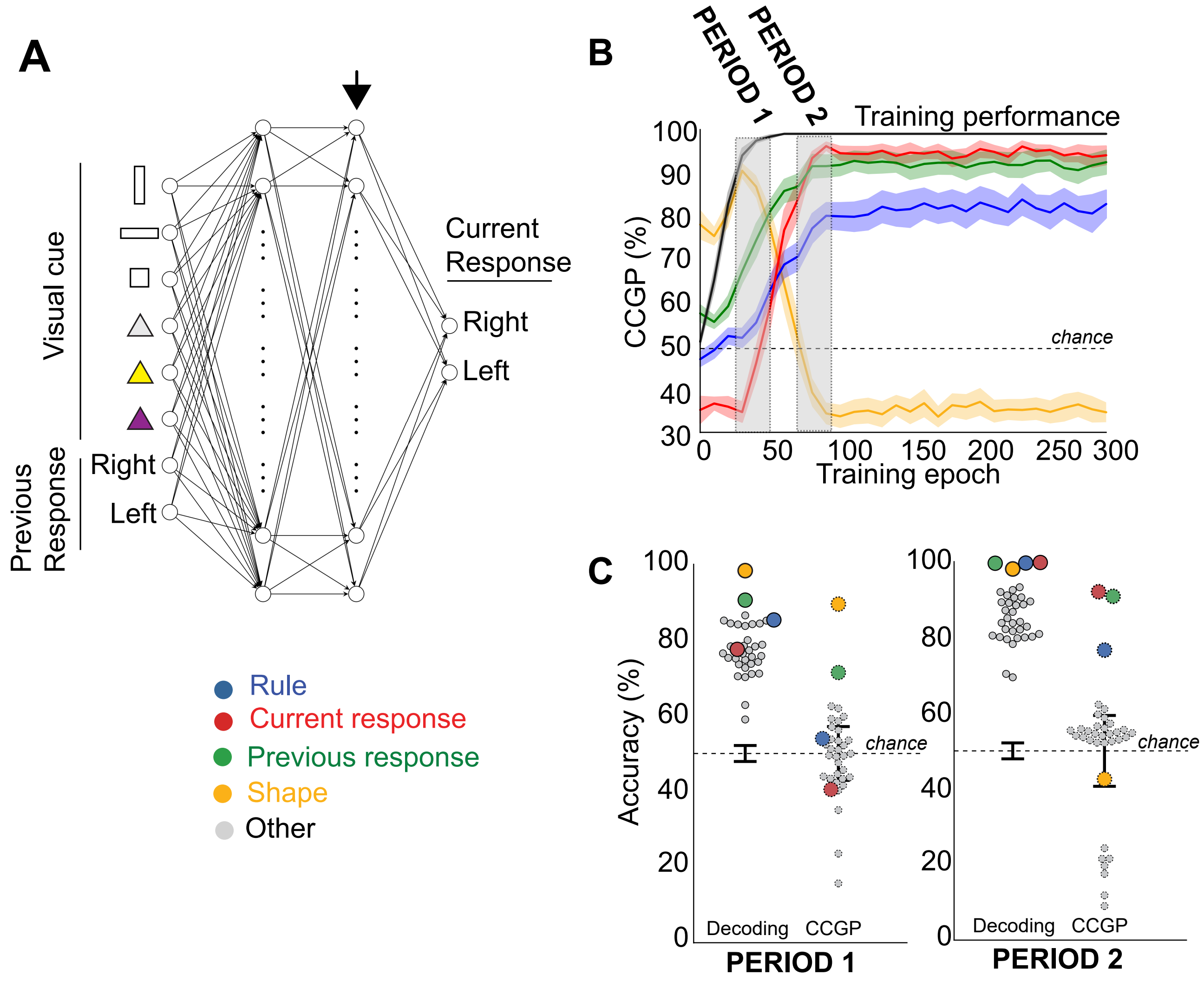


Fig. 7: **Artificial neural network trained to perform the visually cued rule-based task.** **A**) Architecture of the two layers feed-forward neural network. The input is passed through two hidden layers with 100 Rectified Linear units each. Six input units encode the visual cue (vertical rectangle, horizontal rectangle or square combined with color, grey, yellow, or purple, indicated by the triangles). Other two input units encode the previous response (right or left). The output of the network is the current response (right or left) encoded by two units. The vertical black arrow indicates the hidden layer in which we analyzed the units' activations. **B**) CCGP along the training epochs for the four main task variables averaged across 10 models. The shaded area around the mean signal is the standard error of the mean across 10 models. The grey vertical bars indicate the training epochs of interest, labeled as Period 1 and Period 2. **C**) Left: beeswarm plot with the decoding accuracy and CCGP in Period 1 for all the 35 dichotomies. Right: beeswarm plot with the decoding accuracy and CCGP in Period 2 for all the 35 dichotomies. Each circle is a dichotomy. The error bar are ±2 standard deviation around the chance level obtained from random models.

essentially means that for the first monkey the points corresponding to different conditions in the firing rate space are grouped according to shape if one projects the activity on the coding direction of shape (notice that it is only in this subspace that the points cluster, as in the original space the points are still distinct and allow for the encoding of other variables). Analogously, the points are grouped according to the rule in the other more "cognitive" monkey. Both geometries, and even one in which the points are at random locations in the firing rate space (e.g. when the animal is basically using a lookup table strategy) allow for high performance. This is probably why we cannot see significant differences in the overall performance of the two monkeys. However, these geometries have different computational properties that would be revealed only in novel tasks that involve generalization or learning of new rules. For example, the more "cognitive" monkey for which rule is in an abstract format, would probably learn rapidly a novel task in which the rules are the same but the visual cues change. The new visual cues could be "linked" to the pre-existent groups that represent in an abstract format the two possible rules. The other more "visual" monkey is probably in a different learning stage, and the grouping, which is useless for performing the task or for generalizing to new similar tasks, is mostly dictated by the representations of the sensory inputs, as in the early stages of learning of the simulated network.

Indeed, simulations of a simple artificial neural network trained to perform the task used in the experiment, reveal that the two monkeys could be in a different learning stage. In the simulations, the representation of the rule and shape changes as learning progressed. In particular, when the network is in an early phase of training, shape is represented in an abstract format, while the rule is not, although the performance is already high. Later, the performance only slightly increases, but the geometry changes more dramatically, with shape that is no longer abstract while rule becomes an abstract variable, reflecting what is probably a significant change in the cognitive strategy. Similarly, Tsuda et al. [28] recently showed that the different strategies of monkeys and humans in solving a working memory task (monkeys seem to apply a recency-based strategy while humans a target selective strategy [29, 30]) could correspond to two different learning stages of a simple recurrent neural network.

In our experiment, the explanation of the model is in line with the history of the monkeys' training, for which we do not have data, but only some notes: Monkey 2, whose strategy is more "cognitive" and would correspond to a later learning stage of the simulated network, went through a longer training period than Monkey 1 because of its tendency not to switch between rules, persisting with the same response across trials.

Although the model suggests that the differences are due to the training duration, it is also possible that the monkeys would have adopted different strategies even at the same learning stage. We know from machine learning studies on curriculum learning that artificial neural networks can solve the same task in different ways depending on the order of presentation of the samples and more generally, on the details of the learning process [31, 32]. Differences in strategies have been described in experimental studies, in particular in the information representations of the reward [33], in the strategies adopted by two monkeys to solve the same task [34], and in some abstraction tests [35]. A recent study, using a more complex task as the well know pac-man game, has even shown that different strategies can be flexibly switched based on different task demands [36].

In our study we did not test whether abstract representations could lead to the generalization to new stimuli. Introducing a generalization test would have allowed, for example, to test whether the abstract format of the rule, in the second monkey, generated a faster generalization to a new set of rule cues than in the first monkey. Future studies on abstraction should be planned to test whether the task variables encoded in an abstract form, as opposed to those that are not, would facilitate the generalization of the rules to new items or conditions. The ability of generalization has been reported by several studies on macaques [37, 38, 39, 35]. For example, Falcone et al. [39] have shown that monkeys can transfer the nonmatch-to-goal rule from the object domain to the spatial domain in a single session, and Sampson et el. [35] have shown that abstraction can allow generalizing to new conditions, such as new foods, of the rule to choose the worst between two options.

Moving to chronic recordings surely offers the opportunity to follow in time, by recording before, during, and after a task is fully learned, the formation of neural representational geometries already during the training phases. Planned behavioral generalization tests to new task conditions are critical to test the correlation between the geometry of the representation of a given variable and the animal performance in generalization tasks. These future studies will probably highlight even more individual differences, and will allow us to define more precisely what a strategy is, how it is represented in the brain, and to predict and test behavioral consequences in a number of novel situations.

Funding

The work was supported by the the Simons Foundation, Neuronex (NSF 1707398), the Gatsby Charitable Foundation (GAT3708), and the Swartz Foundation. A.G. was supported by the Sapienza University of Rome

(H2020:PH1181642DB714F6).

Author Contributions

A.G. and S.T. conceived and designed the experiments and collected the data. V.F. and F.S. analyzed the data under the supervision of S.F. The data was interpreted by V.F., F.S., S.T., A.G, and S.F., who also wrote the article.

Acknowledgments

We are grateful to L.F. Abbott for his comments on the manuscript and for insightful dicussions. A.G. and S.T. thank Steve Wise, and Andrew Mitz for their numerous contributions. We thank M. Alleman, A. Battista, R. Nogueira, J. Johnston, J. Minxha, L. Posani, K. Rajan, and M. Rigotti for many valuable and knowledgeable discussions.

Competing interests

Authors have no competing interests to report.

References

[1] V. Aguillon-Rodriguez, D. Angelaki, H. Bayer, N. Bonacchi, F. Cazettes, G. Chapuis, A.K. Churchland, Y. Dan, E. Dewitt, M. Faulkner, et al. The international brain laboratory: standardized and reproducible decision-making in mice. *eLife*, 10, 2021.

[2] M. Rigotti, O. Barak, M.R. Warden, X.J. Wang, N.D. Daw, E.K. Miller, and S. Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013.

[3] S. Fusi, E.K. Miller, and M. Rigotti. Why neurons mix: high dimensionality for higher cognition. *Current opinion in neurobiology*, 37:66–74, 2016.

[4] M. Kaufman, M. Benna, M. Rigotti, F. Stefanini, S. Fusi, and A. Churchland. The implications of categorical and category-free mixed selectivity on representational geometries. *Current opinion in neurobiology*, page in press, 2022.

[5] F. Stefanini, L. Kushnir, J.C. Jimenez, J.H. Jennings, N.I. Woods, G.D. Stuber, M.A. Kheirbek, R. Hen, and S. Fusi. A distributed neural code in the dentate gyrus and in ca1. *Neuron*, 107(4):703–716.e4, 2020.

[6] S. Bernardi, M.K. Benna, M. Rigotti, J. Munuera, S. Fusi, and C.D. Salzman. The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4):954–967.e21, 2020.

[7] S.Y. Chung and L.F. Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology*, 70:137–144, 2021.

[8] J.V Haxby, J.S. Guntupalli, A.C Connolly, Y.O Halchenko, B.R. Conroy, M.I. Gobbini, M. Hanke, and P.J. Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

[9] M. Jazayeri and S. Ostojic. Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Current Opinion in Neurobiology*, 70:113–120, 2021.

[10] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[11] I. Higgins, S. Racanière, and D. Rezende. Symmetry-based representations for artificial and biological general intelligence. *Frontiers in Computational Neuroscience*, 16, 2022.

[12] R. Nogueira, C.C. Rodgers, R.M. Bruno, and S. Fusi. The geometry of cortical representations of touch in rodents. *bioRxiv*, 2021.

[13] L. Boyle, L. Posani, S. Irfan, S.A. Siegelbaum, and S. Fusi. The geometry of hippocampal ca2 representations enables abstract coding of social familiarity and identity. *bioRxiv*, 2022.

[14] J. Minxha, R. Adolphs, S. Fusi, A.N. Mamelak, and U. Rutishauser. Flexible recruitment of memory-based choice representations by the human medial frontal cortex. *Science*, 368(6498):eaba3313, 2020.

[15] Y. Xie, P. Hu, J. Li, J. Chen, W. Song, X.J. Wang, T. Yang, S. Dehaene, S. Tang, B. Min, and L. Wang. Geometry of sequence working memory in macaque prefrontal cortex. *Science*, 375(6581):632–639, 2022.

[16] M.F. Panichello and T.J. Buschman. Shared mechanisms underlie the control of working memory and attention. *Nature*, 592(7855):601–605, 2021.

[17] G. Okazawa, C.E. Hatch, A. Mancoo, C.K. Machens, and R. Kiani. Representational geometry of perceptual decisions in the monkey parietal cortex. *Cell*, 184(14):3748–3761.e18, 2021.

[18] E.H. Nieh, M. Schottdorf, N.W. Freeman, R.J. Low, S. Lewallen, S.A. Koay, L. Pinto, J.L. Gauthier, C.D. Brody, and D.W. Tank. Geometry of abstract learned knowledge in the hippocampus. *Nature*, 595(7865):80–84, 2021.

[19] C. Stringer, M. Pachitariu, N. Steinmetz, M. Carandini, and K.D. Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571:361–365, 2019.

[20] I. Higgins, L. Chang, V. Langston, D. Hassabis, C. Summerfield, D. Tsao, and M. Botvinick. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, 12(1):1–14, 2021.

[21] L. She, M.K. Benna, Y. Shi, S. Fusi, and D.Y. Tsao. The neural code for face memory. *bioRxiv*, 2021.

[22] H. Sheahan, F. Luyckx, S. Nelli, C. Teupe, and C. Summerfield. Neural state space alignment for magnitude generalization in humans and recurrent networks. *Neuron*, 109(7):1214–1226.e8, 2021.

[23] J.W. Krakauer, A.A. Ghazanfar, A. Gomez-Marin, M.A. MacIver, and D. Poeppel. Neuroscience needs behavior: Correcting a reductionist bias. *Neuron*, 93(3):480–490, 2017.

[24] S. Tsujimoto, A. Genovesio, and S.P. Wise. Comparison of strategy signals in the dorsolateral and orbital prefrontal cortex. *The Journal of neuroscience*, 31(12):4583–92, 2011.

[25] N. Kriegeskorte, M. Mur, and P.A. Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4, 2008.

[26] P.C. Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India.*, 2:49–55, 1936.

[27] W.J. Johnston and S. Fusi. Abstract representations emerge naturally in neural networks trained to perform multiple tasks. *bioRxiv*, 2021.

[28] B. Tsuda, B.J. Richmond, and T.J. Sejnowski. Exploring strategy differences between humans and monkeys with recurrent neural networks. *Manuscript in preparation.*

[29] J.H. Jr Wittig and B.J. Richmond. Monkeys rely on recency of stimulus repetition when solving short-term memory tasks. *Learning and memory*, 21:325–333, 2014.

[30] J.H. Jr Wittig, B. Morgan, E. Masseau, and B.J. Richmond. Humans and monkeys use different strategies to solve the same short-term memory tasks. *Learning and memory*, 23:644–647, 2016.

[31] P. Soviany, R. Tudor Ionescu, P. Rota, and N. Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130:1526–1565, 2022.

[32] D. Kepple, R. Engelken, and K Rajan. Curriculum learning as a tool to uncover learning principles in the brain. *International Conference on Learning Representations*, 2022.

[33] P. Enel, J.D. Wallis, and E.L. Rich. Stable and dynamic representations of value in the prefrontal cortex. *eLife*, 2020.

[34] T. Tsujimoto, H. Shimazu, Y. Isomura, and K. Sasaki. Theta oscillations in primate prefrontal and anterior cingulate cortices in forewarned reaction time tasks. *Journal of Neurophysiology*, 103(2):827–843, 2010.

[35] W.W.L. Sampson, S.A. Khan, E.J. Nisenbaum, and J.D. Kralik. Abstraction promotes creative problem-solving in rhesus monkeys. *Cognition*, 176:53–64, 2018.

[36] Q. Yang, Z. Lin, W. Zhang, J. Li, X. Chen, J. Zhang, and T. Yang. Monkey plays pac-man with compositional strategies and hierarchical decision-making. *eLife*, 11:e74500, 2022.

[37] E. Procyk, Peter Ford D., C. Amiez, and J.P. Joseph. The effects of sequence structure and reward schedule on serial reaction time learning in the monkey. *Cognitive Brain Research*, 9(3):239–248, 2000.

[38] A.A. Wright and J.S. Katz. Mechanisms of same/different concept learning in primates and avians. *Behavioural Processes*, 72(3):234–254, 2006.

[39] R. Falcone, S. Bevacqua, E. Cerasti, E. Brunamonti, M. Cervelloni, and A. Genovesio. Transfer of the nonmatch-to-goal rule in monkeys across cognitive domains. *Plos one*, 8(12):e84100, 2013.

[40] S. Tsujimoto, A. Genovesio, and S.P. Wise. Neuronal Activity during a Cued Strategy Task: Comparison of Dorsolateral, Orbital, and Polar Prefrontal Cortex. *Journal of Neuroscience*, 32(32):11017–11031, aug 2012.

[41] V. Fascianelli, S. Tsujimoto, E. Marcos, and A. Genovesio. Autocorrelation Structure in the Macaque Dorsolateral, But not Orbital or Polar, Prefrontal Cortex Predicts Response-Coding

Strength in a Visually Cued Strategy Task. *Cerebral Cortex*, 29(1):230–241, jan 2019.

[42] V. Fascianelli, L. Ferrucci, S. Tsujimoto, and A. Genovesio. Neural correlates of strategy switching in the macaque orbital prefrontal cortex. *Journal of Neuroscience*, 40:3024–3034, 2020.

[43] A. Genovesio, P.J. Brasted, A.R. Mitz, and S.P. Wise. Prefrontal cortex activity related to abstract response strategies. *Neuron*, 47(2):307–20, jul 2005.

[44] A. Genovesio, S. Tsujimoto, and S. Wise. Encoding problem-solving strategies in prefrontal cortex: Activity during strategic errors. *The European journal of neuroscience*, 27:984–90, 03 2008.

[45] T. Bussey, S. Wise, and E. Murray. ”the role of ventral and orbital prefrontal cortex in conditional visuomotor learning and strategy use in rhesus monkeys (macaca mulatto)”: Correction to bussey et al (2001). *Behavioral Neuroscience*, 115:1317–1317, 12 2001.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[47] S. Fusi, M. Annunziato, D. Badoni, A. Salamon, and D.J Amit. Spike-driven synaptic plasticity: theory, simulation, vlsi implementation. *Neural Comput*, 12:2227–58, 2000.

[48] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[49] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Materials and Methods

Subjects

All the details about the experiment are reported in the original article [24]. Here we give only a brief description of these details.

Two male rhesus monkeys (Macaca mulatta, 10–11kg in weight) were trained to perform a visually cued rule-based task. All experimental procedures were in agreement with the Guide for the Care and Use of Laboratory Animals and were approved by the National Institute of Mental Health Animal Care and Use Committee.

Each monkey, while performing the task, sat in a primate chair, with the head fixed in front of a video monitor 32cm away. An infrared oculometer (Arrington Research, Inc., Scottsdale, AZ) recorded the eye positions.

Data collection and histology

Up to 16 platinum iridium electrodes (0.5–1.5$M\Omega$ at 1kHz) were inserted into the cortex with a multielectrode drive (Thomas Recording) to record single-cell activity from dorsolateral prefrontal cortex (Figure 1C). The recording chambers (18mm inner diameter) were positioned and angled according to magnetic resonance images (MRI). The single-cell potentials were isolated off-line (Off Line Sorter, Plexon), based on multiple criteria, including principal component analysis, the minimal interspike intervals, and close visual inspection of the entire waveforms for each cell. Eye position was recorded with an infrared oculometer (Arrington Research). The recording sites were localized by histological analysis and MRI (see Tsujimoto et al. [24] for more information).

The behavioral task

A sequence of the task events of the visually cued rule-based task is shown in Figure 1A [24, 40, 41, 42]. For clarity, previous works' authors referred to this task as the visually cued strategy task. The stay and the shift rules were designed as strategies because they represented a simplification of the repeat-stay and change-shift strategies used in previous neurophysiological studies [43, 44]. These two strategies were identified by Bussey et al. [45] studying the behavior of monkeys during the learning of visuomotor associations. The monkeys in their study adopted spontaneously the strategies to facilitate learning. As opposed to the previous studies of this task, here we refer to "strategy" as a possible way adopted by the monkey to solve the task, and to "rule" what is instructed to the monkey to perform the task. In each trial, the monkey was required to make a saccade towards one of the two spatial targets, according to a shift or stay rule cued by a visual instruction (Figure 1B). The appearance of a fixation point (a 0.6° white circle) located at the center of the video screen, with 2 peripheral targets (2.0° white square frames) placed 11.6° to the left and right of the fixation point, represented the beginning of a trial. The monkey had to maintain fixation on the central spot for 1.5s; after that, a cue period of 0.5s followed. During the cue period, a visual cue appeared at the fixation point. In each trial, one visual cue was chosen pseudorandomly from a set of four visual cues: a vertical (light gray) or horizontal (light gray) rectangle with the same dimensions (1.0°× 4.9°) and brightness, or a yellow or purple square with the same size (2.0°× 2.0°) (Figure 1B). Each visual cue instructed either the stay or shift rule. The stay rule, instructed by the vertical rectangle or the yellow square, cued the monkey to choose the same target chosen in the previous trial (as shown in the two consecutive trials' example in Figure 1A). Conversely, the horizontal rectangle or the purple square instructed the shift rule, which required the monkey to choose the target not chosen in the previous trial. The end of one trial and the beginning of the next one were separated by an intertrial interval of 1s. The first trial required a random choice of the target since no previous response could be integrated with the information on the current rule. Moreover, in the first trial, the monkey was always rewarded. The monkey had to maintain the fixation on the central point during the whole fixation period (1.5s) and the cue period (0.5s) as well as during a subsequent delay period of 1.0, 1.25, or 1.5 s, pseudorandomly selected. The fixation window was a ±3° square area centered on the fixation point. Both monkeys maintained fixation accurately and rarely made a saccade within the fixation window [24, 40]. Any fixation break during the fixation, cue, or delay periods led to abortion of the trial. The fixation point and the two peripheral targets were kept on the screen for the whole duration of the delay period. The disappearance of the fixation spot represented a go signal, instructing the monkey to choose one target by making a saccade to one of them. When the monkey fixated one of the targets, both squares became filled. The entry of the gaze into the response window was labeled as target acquisition. The monkey had to maintain the fixation on the target for 0.5s (pre-feedback period). Any fixation break during the pre-feedback period led to abortion of the trial. After the pre-feedback period, in the case of correct response, feedback was provided as a liquid reward (0.2ml drop of fluid) or, in case of incorrect response, as red squares over both targets. In the case of an error, the same cue was presented again in the following trial, called "correction trial". Correction trials were presented until the monkey responded correctly. Usually, after an error, there was not more than a correction trial [24, 40].

Neurons and trials sample selection, pseudo-simultaneous population trials, and task conditions definition

We analyzed the neural activity of each monkey separately, only in complete and correct trials, from $400ms$ before the cue onset until $500ms$ after the cue offset. Linear decoders were trained and tested on pseudo-simultaneous population trials (pseudo trials). We defined a pseudo trial as the combination of spike counts randomly sampled from every neuron in a specific time bin and task condition [2]. The task condition is one of the eight possible combinations of task variables listed in Figure 1D. We analyzed the activity of neurons recorded in at least five trials per task condition.

Pseudo trials were generated as follows: given one time bin t and task condition p, for every neuron we randomly picked a trial of task condition p, and we computed the spike count in the time bin t. The single pseudo trial γ, for condition p at time bin t, is then $\gamma^p(t) = (\gamma_1^p(t), \gamma_2^p(t), ..., \gamma_N^p(t))$, where N is the number of recorded neurons, and γ_i^p (i is the neuron identity, $i = 1, ..., N$) is the spike count. We repeated this procedure 100 times, ending up with 100 pseudo trials per task condition and time bin.

Since we did not know a priori which task variables are represented by the neural ensemble, and in order not to introduce any bias in the selection of the task variables to decode, we defined a dichotomy as each pairing of the task conditions in group of four, for a total of 35 dichotomies [6]. Each dichotomy is a variable that could be decoded. Four of the 35 dichotomies overlap with the task variables. All the other dichotomies cannot be explicitly interpreted in terms of any of the task variables, but rather as a combination of task variables which we referred to as other dichotomies. In particular, the four dichotomies which overlap with the task variables are the previous response, rule, current response, and the shape of the visual cue (Figure 1D). The latter identifies whether the visual cue was a rectangle or a square, that could also be interpreted as grey colored and non grey colored cue.

Decoding of the neural population activity

For each dichotomy, that is a binary variable, we trained a Support Vector Machine (SVM) classifier with a linear kernel [46] to classify the spike count into either of the two values of the dichotomy. In all the SVM classifiers we set a regularization term equal to 10^3. We tried a range of regularization terms from 1 up to 10^3, without any significant change in the final results. We decoded the neural activity in a $200ms$ time bin stepped by $20ms$ along time from $400ms$ before the cue onset until $500ms$ after the cue offset. The linear classifier was trained on pseudo trials built from randomly selected trials. In more details, for every neuron we selected the 80% of the trials as training set, and the remaining 20% as testing set to build the pseudo trials. We cross validated the linear decoder 100 times, by randomly choosing the 80% of the pseudo trials as training set, and the remaining 20% as testing set. We showed the final accuracy of the linear decoder as the ratio between the number of correct predictions to the total number of predictions on the testing set averaged across the cross validations. To evaluate the statistical significance of the neural signal, we built a null model by randomly shuffling the task condition labels among the pseudo trials. For each shuffle, we trained a linear decoder on the shuffled training set and we assessed its accuracy on the shuffled testing set. We repeated the shuffle procedure 100 times obtaining a null model distribution. We defined the chance interval as the interval between 2 standard deviations of the null model distribution around the chance level at 50%.

The data were extracted by custom MatLab functions (The MathWorks, Inc., Natick, MA, USA). All decoding analyses were performed by using scripts of the scikit-learn SVC package along with custom Python scripts [46].

Neural representation of variables in an abstract format and the Cross Condition Generalization Performance

After assessing which task variables are decoded, we asked in what format they are represented. In particular, we asked whether they are represented in an abstract format. A variable could be defined to be in an abstract format when a linear decoder trained to classify the value of the variable can generalize to new task conditions never used for training. To assess to what extent a variable is in an abstract format, we computed the Cross Condition Generalization Performance (CCGP), that is the performance of a linear decoder in generalizing to new task conditions not previously used for training [6]. The difference between the traditional cross-validated linear decoder and the cross condition generalization is in the data used for training and testing the classifier. In the traditional-fashioned decoding analyses, a decoder is trained on a sub-sample of trials randomly picked from each (experimental) condition, and tested on the held-out trials retained from each condition. At the end, the decoder is trained and tested on all the conditions, and the generalization is only across trials. The CCGP, instead, is computed by training a linear decoder only on a fraction of trials from a subset of conditions, and tested on trials belonging to new conditions not used for training. The generalization is now not only across trials, but also across conditions.

We assessed the CCGP for each of the 35 dichotomies as follows. Given a dichotomy, defined as a pairing of task conditions in group of four, we trained the decoder to classify the value of the dichotomy using trials from three task conditions from each side of the dichotomy, and tested it on the one held out condition from each side. Since each side of the dichotomy has four task conditions, there are 16 possible ways of choosing the training and testing condition set. For each choice of training and testing, we applied 10 cross validations, randomly choosing the 80% of training trials and 20% of testing trials. We reported the average performance across all the 16 possible choices of training and testing conditions and the 10 cross validations for each dichotomy. To assess the statistical significance of the CCGP, we built a null model where the geometrical structure in the data was destroyed, but keeping the variables still decodable [6]. To do that, we applied a discrete rotation to the noise clouds (the trials firing rate of each condition) by permuting the axes of the firing rate space, and randomly assigning neural activity to neurons. We repeated this procedure for each cluster separately. We generated 100 null models and for each of them we computed the CCGP for all dichotomies, as done on real data. We defined the chance interval for the CCGP measure as the interval between 2 standard deviations of the null model distribution around the chance level at 50%.

Multi Dimensional Scaling analysis

We used the Multi Dimensional Scaling (MDS) transformation to seek a low-dimensional representation of the data. We computed the metric MDS, where the dissimilarity matrix was built as follows. The neural activity was first averaged across pseudo trials within each task conditions, and then we constructed a $p_c \times p_c$ matrix (with p_c indicating the number of conditions) which stored the Euclidean distance between the average firing rate between each paired condition. In order to keep information regarding the noise cloud of each task condition, we normalized the Euclidean distance matrix by the squared root of the sum of the variance of each condition along the distance direction between the two clouds. For the analysis based on single pseudo trial (Figure 3B), the dissimilarity matrix was defined as a $p_t \times p_t$ matrix, with p_t indicating the total number of pseudo trials across all conditions. This dissimilarity matrix stored the Euclidean distance between the firing rate of each pair of pseudo trials, and it was normalized as described above.

Behavioral analyses

We computed the behavioral performance and reaction times of each monkey separately, combining all the sessions we considered for the neural analyses. We computed the reaction time (RT) only in complete and correct trials. The RT is defined as the time difference between the go signal and target acquisition in each trial. In order not to bias the results due to outliers, we removed those trials with RT larger than 3 standard deviations from the mean. Since the neural analyses revealed that the difference between the two monkeys comes from different representational geometry of the rule and the shape of the visual cue, we grouped trials per rule (stay-shift) and shape (rectangle-square), for a total of four conditions. We compared the distribution of RTs of trials with different rules and shapes, separately. To test whether the RTs distributions were significantly different, we run the Mann-Whitney U-test (p-value<0.05).

Moreover, for each of the previous four task conditions, we computed the average performance across the sessions. The error bar of the estimated average performance was assessed by applying the following formula [47]:

$$\sigma_{+/-} = \frac{Pn + \frac{k^2}{2} \pm k\left[P(1-P)n + \frac{k^2}{4}\right]^{\frac{1}{2}}}{n+k^2}, \qquad (1)$$

where n is the number of trials used to compute the performance across sessions, P is the average performance, and k is the confidence level in terms of standard deviation that we fixed equal to 2. To assess whether the performance was statistically different between different conditions, we applied the chi-squared test (p-value<0.05).

Multi-Linear Regression Model for behavior

To better investigate the behavioral differences between the two monkeys, we fitted a multi-linear regression model on a single trial basis. We included in the model only complete and correct trials, and we discarded those trials with reaction time larger than 3 standard deviations from the mean as done in the behavioral analysis. For each trial, we took three independent binary input factors to the model: rule (+1/-1), previous response (+1/-1), and shape (+1/-1). We also included all the interaction terms. The output of the model is the reaction time, and the multi-linear model is defined as follows:

$$\begin{aligned} \mathrm{RT} = &\omega_1 \times [\text{rule}] + \omega_2 \times [\text{previous}] + \omega_3 \times [\text{shape}] + \\ &\omega_4 \times [\text{rule} * \text{previous}] + \omega_5 \times [\text{previous} * \text{shape}] + \\ &\omega_6 \times [\text{rule} * \text{shape}] + \eta, \end{aligned} \qquad (2)$$

where $\omega_{1,\dots,6}$ are the weights of each factor, and η is a constant term. We fitted 100 models, each time randomly subsampling trials from each task conditions, in each monkey separately. The number of trials per task condition was set to the minimum number of trials across conditions. We fitted each model by using the ordinary least squares method [48]. For each factor, we compared the weights' distributions across models between the two monkeys using the Mann-Whitney U test (p-value$<$0.05).

Representational geometry in an artificial neural network trained to perform the visually cued rule-based task

We trained a two layers feed-forward neural network to perform the visually cued rule-based task. The inputs to the network are the features of the visual cue and previous response, because they are the information provided to the monkeys to perform the task. In more details, the features of the visual cue are defined by two one-hot vectors, with three units each encoding whether the visual cue is a vertical, or horizontal rectangle, or a square, and the color of the cue, i.e. yellow, purple,or gray. The previous response (right, left) is encoded by a one one-hot vector with two units. We added a gaussian noise to the input patterns ($\mu = 0$, $\sigma = 1$). For each training epoch we generated 4000 trials, 500 trials per task condition. The input is passed through two hidden layers of 100 Rectified Linear units each. The output of the network is the current response (right,left) encoded by a one-hot vector of two units.

To train the network, we used the PyTorch framework [49]. Each layer's weights were randomly initialized from uniform distribution $\mathcal{U}$(-k, +k), where $k = \frac{1}{\sqrt{nfeatures}}$, and $nfeature$ is the number of layer's units. We trained the network with back-propagation using the Adam algorithm as optimizer ("Adam" in PyTorch), with learning rate equal to 0.001. Training proceeded for 300 epochs of 500 mini-batches each. We used the Mean Squared Error as loss function ("MSELoss" in PyTorch), and rectified linear function ('ReLu' in PyTorch) as activation function for each unit. We analyzed the representational geometry of the activation of the units in the second hidden layer along the training epochs, using the same analytic tools we used to analyze the neural data.

Supplementary Figures

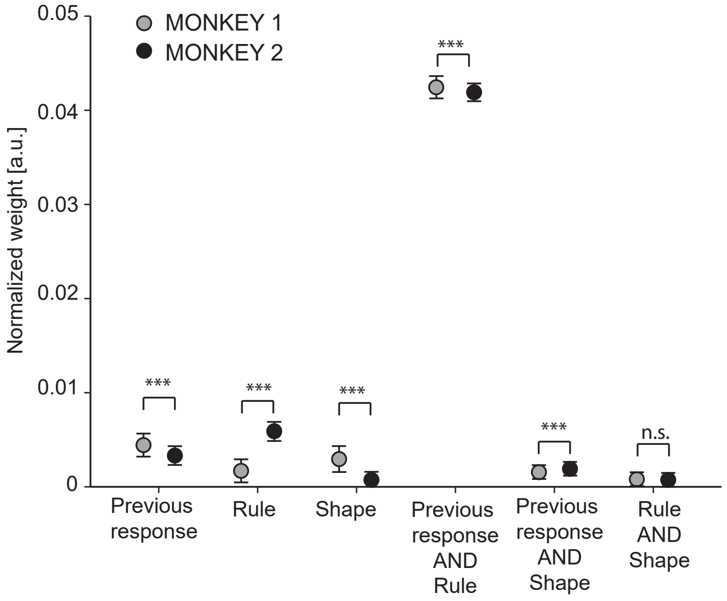


Fig. 1: **Multi-Linear regression analysis results**. Mean of the distribution of the weights of 100 multi-linear regression models. The reaction time is predicted on a single trial using three factors: previous response, rule, and shape along with the interaction terms. The interaction of previous response with rule has the strongest factor in both monkeys, since the combination of these two factors is essential to elaborate the correct response. The error bars are the 2 standard deviations of weights across of 100 models. n.s.: not significant; *** Mann-Whitney U test: p<0.001.

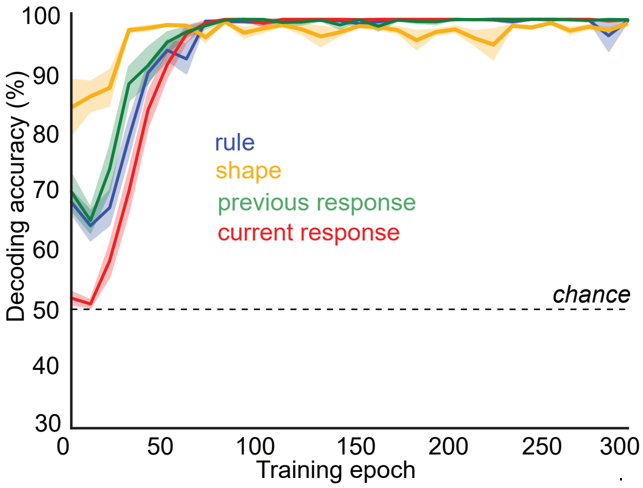


Fig. 2: **Decoding accuracy of the main four task variables along the training of an artificial neural network.** Decoding accuracy of task variables in the second hidden layer of a two layers feed-forward neural network during the training to perform the visually rule-based task. After ∼50 training epochs, all the variables are decoded with high accuracy until the end of the training epochs. The shaded area around the mean signal is the standard error of the mean across 10 models.