Neural Basis of Rule-Dependent Flexible Mapping of Features to Categories in Prefrontal Cortex

A dissertation submitted to the faculty of the University of Minnesota
by

Min-Yoon Park

in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Thesis advisor: Matthew V. Chafee, Ph.D.

March 2020

Acknowledgements

I praise Ebenezer God who has been with me, has led me, and has helped me to this point. I praise Immanuel God who is with me, leads me, and helps me at this moment. I praise Jehovah-Jere God who will be with me, will lead me, and will help me always. I praise God for giving me the best.

I am grateful to my thesis advisor and mentor Matt for supporting my project, driving my intellectual development, and guiding me to become an independent researcher. I am thankful for the guidance and educational support by Apostolos, Ben, Paul, and Dave. I am also thankful for all the support and help from Dean, Dale, Adele, and the VA staff. I appreciate the prayer and care from my family, especially my parents in South Korea and my husband Ian. Finally, Monkey 026 and Monkey 037 were my two little heroes in my thesis project.

Abstract

A considerable body of research has documented the existence of neural signals that encode categories in primate prefrontal cortex. Comparatively little is known regarding how these neural representations are derived from sensory inputs, or more specifically, how neural signals that encode features are converted into neural signals encoding categories. Understanding that transform at a circuit level would shed needed light into the computational origin of abstract neural signals in prefrontal cortex. Here we analyze neural signals that encode features and categories in prefrontal cortex of monkeys performing a task that requires them to flexibly map one form of neural signal to the other. At a behavioral level, we show that rules influence which features of visual stimuli are sampled to compute the category of the stimulus. At the neural level, we show that the neural representation of features is relatively rule-independent, however the functional linkage between feature and category signals is re-routed as a function of which rule is in force. These results suggest that prefrontal circuits carrying out the feature-to-category transform can be dynamically reconfigured as a consequence of cognitive rules.

Table of Contents

List of Figures

1

Introduction

Everyday mental life involves a series of categorizations. Is this milk edible or spoiled? Is the water hot or cold? Is outside dangerous or safe? Classifying and dividing objects, conditions, and environments into meaningful groups are fundamental processes leading us to take actions and make decisions. Moreover, different features of the same object can be flexibly analyzed to assign the object to different categories based on prior knowledge, internal goals, or rules. Therefore, intelligent, goal-directed categorization often involves selecting which features are relevant to category membership and which are irrelevant.

The Stroop task (Stroop, 1936) is an example that gives us insight into how the brain flexibly divides incoming sensory information into relevant and irrelevant feature dimensions based on rules. In the classic Stroop task, words of colors are presented, and the task rule is to name the color of letters. The color of the printed ink is the relevant

sensory information while the written word is irrelevant. Congruent trials refer to trials in which the color of the printed ink matches the color indicated by the written word. Conflict trials refer to those in which the stimuli do not match. Importantly, earlier studies showed that the reaction time when naming the color of the ink was significantly faster during congruent trials than during control conditions (colored letters without meaning). Alternatively, the reaction time was significantly longer during conflict trials than during control conditions (Stroop, 1936; Cohen et al., 1990).

Two major conclusions can be drawn from the results regarding reaction time during the Stroop task. First, the brain does not simply filter out the irrelevant information but integrates the irrelevant information with the relevant information when it is judged to enhance performance (as in the congruent condition). This implies that the brain generates an internal strategy to maximize its performance. Second, the irrelevant information may interfere in the processing of the relevant information during the conflict condition (conflict condition).

Prefrontal cortex and cognitive control

The brain needs high-level cognitive control to manage this process, dividing, filtering and combining the massive and varied sources of sensory information in order to make cognitive decisions and select actions. The main brain area responsible for cognitive control must connect with other brain areas related with receiving sensory inputs and motor control. Prefrontal cortex (PFC) is a network hub of neocortical interconnections that receives and sends projections from nearly all sensory systems,

motor systems, and many subcortical structures (Miller and Cohen, 2001), with most of the connections being reciprocal (Ilinsky et al., 1985; Cavada and Goldman-Rakic, 1989a, 1989b; Croxson et al., 2005). Because of its massive connections with other functional domains, the PFC has been considered as a critical brain area for cognitive control and has been studied in both humans and non-human primates (Miller and Cohen, 2001; Goldman-Rakic, 2011; Goodwin et al., 2012; Crowe et al., 2013; Mante et al., 2013; Seo et al., 2014; Donahue and Lee, 2015; Blackman et al., 2016; DeNicola et al., 2020).

These studies have led to the current understanding that PFC has an essential role in cognitive control (goal-directed thought and behavior) (Miller and Cohen, 2001; Luria, 2012; Stuss and Knight, 2013; Fuster, 2015). The most famous example of a PFC lesion study was the Phineas Gage case of 1868. In this case, Mr. Gage suffered an accident, during which an iron bar passed through his skull and damaged his left PFC (Harlow, 1868). His personality and ability to organize his behavior changed dramatically as a result, indicating cognitive control deficits. Despite this early evidence of PFC's importance in cognitive control, several initial lesion studies in humans subsequently raised doubt regarding this role after they reported post-lesion preservation of mental capacity and behavior (Hebb, 1939; Petrie, 1952). However, more recent studies have confirmed that PFC controls a wide range of cognitive functions. Lesions of the dorsolateral PFC (Brodmann areas 9 and 46) in humans and non-human primates cause deficits in working memory (Passingham, 1985; Tsuchida and Fellows, 2009; Barbey et al., 2013), attention (Parker et al., 1998; Voytek et al., 2010), motivation (Ferrier, 1886; Duncan et al., 2008), rule-dependent learning and rule-switching (Milner, 1963; Shallice

and Burgess, 1991; Buckley et al., 2009; Moore et al., 2009), as well as planning and problem solving (Simon, 1975; Shallice, 1982; Burgess, 2000).

Imaging studies also have provided further evidence that PFC is a key brain area in cognitive control. PFC is functionally activated in association with working memory (D'Esposito et al., 1999, 2000), response selection (Rowe et al., 2000; Rowe and Passingham, 2001) and rule switching (Nakahara et al., 2002). Furthermore, brain imaging in humans showed that PFC is organized as a cascade of executive processes from premotor to anterior PFC regions that control behavior according to stimuli, the present perceptual context, and the temporal episode in which stimuli occur (Koechlin et al., 2003).

Neurophysiological studies in non-human primate have demonstrated that single prefrontal neurons reflect cognitive control. PFC neurons not only encode task states (Asaad et al., 2000) and abstract categories (Christoff et al., 2009; Crowe et al., 2013; Blackman et al., 2016), but also encode flexible cognitive strategies (Seo et al., 2014) and adjust categories of stimuli and actions based on changing rules (Goodwin et al., 2012; Crowe et al., 2013; Mante et al., 2013). Top-down signals from PFC to other brain areas play a role in representing, maintaining, ignoring and suppressing relevant and irrelevant information and actions (Miller and Cohen, 2001; Tanji and Hoshi, 2008; Mansouri et al., 2009; Merchant et al., 2011; Crowe et al., 2013; Mante et al., 2013; Donahue and Lee, 2015) providing insight into how relevant and irrelevant information are integrated in PFC (Miller and Cohen, 2001; Mante et al., 2013; Donahue and Lee, 2015).

Prefrontal cortex and categorization

Categorization refers to the process of dividing information into meaningful groups that share common features. To categorize stimuli or information, a number of processes are required, including acquiring category knowledge through experience, selecting feature dimensions relevant to category membership, applying internal boundaries on those feature dimensions, and deciding which side of those boundaries individual stimuli fall on to determine category membership. Not only that, but also all the cognitive processes should be flexible depending on the given environment, goal, rule, or circumstance. Because high-level cognitive processes are required, PFC has been considered as a critical brain area for categorization. Many of previous studies have shown that there are category signals in PFC (Freedman et al., 2001) and the signals in PFC are dynamically adjusted by category learning (Antzoulatos and Miller, 2011), categorization rule-learning (Sleezer et al., 2016) or flexible categorization processes (Goodwin et al., 2012; Crowe et al., 2013). However, how feature representations are flexibly transformed into category representations in PFC is still unknown.

Categorization has been studied widely in non-human primate PFC (Freedman et al., 2001, 2003; Antzoulatos and Miller, 2011; Merchant et al., 2011; Goodwin et al., 2012; Crowe et al., 2013; Mante et al., 2013; Roy et al., 2014). The first experiment to relate categorization to single neuron activity in primate PFC involved monkeys classifying a visual shape as either a dog or cat (Freedman et al., 2001). The critical feature of the experiment was that visual stimuli were continuously morphed between the two categories (that is they were morphed along a sensory continuum between dog and cat). These authors found that individual PFC neurons exhibited dichotomous category

signals, preferring exemplars from either the dog or cat categories. These signals were not tuned by the degree of morphing within the preferred category (that is they responded equally to all exemplars in that category). However, in this task, there were notable limitations. First, it was not clear whether the category signals recorded in PFC were abstract category signals or visual feature signals related to the shapes of the objects because the features of stimuli themselves defined the category, and the relationship between visual features and category membership was not varied for individual neurons. Hence, category signals were confounded with feature signals. Second, the feature-to-category transformation was also not clearly delineated, in part because of the above confound between feature and category signals at the neural level.

Another study investigating the neural basis of category representation in primate PFC utilized a prototype distortion task (Antzoulatos and Miller, 2011). In the task, distorted dot patterns from an exemplar dot pattern were given in the sample period, and monkeys reported their category memberships through a saccade response (with two different saccade directions indicating the two potential categories to which the dot pattern could belong). In the paper, strong PFC activity encoding the category of the stimulus as read out by the direction of the upcoming saccadic report was observed. However, because the relationship between the category of the stimulus and the direction of the motor response used to report category membership was never varied, a similar problem exists with this study as with the above dog and cat categorization task (Freedman et al., 2001), and that is that it is not possible to differentiate whether neural activity encoded the perceptual category of the stimulus or the direction of the associated motor response. Even though the authors asserted that the signals reflected novel abstract

categories, their task design may have resulted in neural signals that encode saccade directions but not purely abstract categories.

One of the shared weaknesses of simple categorization tasks of this type is that the neural computation transforming feature-into-category signals is hard to detect because the inflexible relationship between stimuli and categories confounds category signals with sensory or motor signals at the neural level. To overcome that weakness and to investigate the process of feature-to-category transformation, a rule-dependent categorization task was employed while neural activity in PFC was recorded in monkeys (Mante et al., 2013). In the task, stimuli were patches of moving dots that varied along two feature dimensions - one was the direction and coherence of dot motion and the other one was the colors of the dots. In the task, if the rule was direction, the dominant direction (left or right) was the relevant feature dimension upon which to categorize the stimuli. If the rule was color, the dominant color (red or green) was the relevant feature dimension. A visual cue was provided (the color of the fixation target) indicating the rule in force during the task.

In the beginning of this chapter, the question of how the brain utilizes relevant and irrelevant feature information to make categorical decisions was mentioned, and this study (Mante et al., 2013) investigated the relevant neural mechanisms in PFC. Based on their neurophysiological results, the authors proposed a model of how PFC decides the relevant action differently in response to the same visual stimuli. The authors found that irrelevant information was not filtered out before it reached PFC – rather PFC neurons encoded both the direction and color of the moving dot stimuli regardless of the rule in force. Based on a combination of neurophysiological and modeling results, the authors

proposed that PFC circuits implemented the rule switch required by their task by switching which population of feature neurons was integrated to drive population activity patterns along a line attractor through the rate space encoding the two possible directions of the saccadic response. Therefore, the selection was accomplished at the level of population dynamics.

This was an interesting result, however, there are limitations in this study. First, whereas the category of stimuli had to be flexibly determined between different contexts (rules), under any individual rule the relationship between the features of the stimulus and the category to which it belonged remained fixed. For example, if the dot pattern stimulus was predominantly red color and moving to the left, the same stimulus could never belong to the green or right categories. The task just required switching between two feature-based categories for each stimulus depending on the rule, rather than decorrelating feature and category variables by varying the relationship between features and categories. Since categories were strongly related with visual features, either motion or color, the category signal did not necessarily reflect the category of the stimulus as a sensory-abstracted concept. Moreover, category signals may have been confounded with the features of the saccade targets that monkeys had to choose between, because once the category of the stimulus was determined, so was the features of the saccade target that monkeys would select to report the category. This means that the suggested model of PFC function could be more about feature-to-action than feature-to-category transformations. Finally, in this task a visual cue was provided to indicate which rule (direction or color) was relevant. By cueing the current rule externally, flexible and dynamic processes in PFC related to rule selection could be restricted, and the recorded

data might not actually reflect the process of how PFC spontaneously categorizes objects in the real world.

One way to decorrelate the features of visual stimuli from the categories to which they belong is to base category membership on the relationship of one stimulus with respect to another, reference stimulus. Changing the reference changes the categorical relationship between stimuli without changing the features of the categorized stimulus. An example of relational categorization task of this type is the dynamic spatial categorization (DYSC) task (Goodwin et al., 2012). In this match-to-category task, monkeys categorized a dot in a visual display based on its spatial relationship to a line serving as a category boundary. In a horizontal orientation, the boundary instructed an above/below rule, and monkeys reported whether the dot was above or below the boundary, whereas in a vertical orientation, the boundary instructed a left/right rule, and monkeys reported whether the dot was to the left or right of the boundary. In this way, a single stimulus could be allocated to different categories (for example 'left' and 'above') depending on the orientation of the category boundary. , The authors found that single neurons in PFC exhibited rule-dependent category signals that encoded the category of a stimulus as a joint function of its position and the rule applied. . Even though the task itself was a relational categorization task, there are still limitations to be discussed. First, the same stimulus could be categorized into two different categories but only between rules – within each rule the relationship between features and categories was fixed. For example, one dot could be categorized as left or above depending on the rule but never into right or below. Only 8 stimulus locations were used, and the small number of stimuli could make it possible for monkeys to solve the task using a look-up table, meaning that

monkeys could remember responses associated with the full set of boundary and dot stimulus combinations. Finally, the rule was visually instructed like the rule-dependent absolute categorization task mentioned above (Mante et al., 2013).

Comparing the neurophysiological data recorded during the performance of different tasks, we have to consider that task design does not simply affect the interpretation of neural data but is likely to also influence the underlying neural dynamics as well. For example, in a task in which monkeys categorized the direction of a moving dot stimulus, it was reported that the parietal category signals lead prefrontal category signals (Fitzgerald et al., 2012). However, in the relational categorization tasks, which has flexible category boundaries, it was reported that the PFC category signals led parietal category signals (Merchant et al., 2011; Crowe et al., 2013). All three tasks evaluated monkey categorization, but the neural dynamics, i.e. either bottom-up or top-down control, varied depending on task demands.

Another important factor in task design affecting internal neural dynamics is how the rule is instructed. In categorization tasks, rules can be either externally guided by a cue or internally determined by trial-and-error feedback. In human studies, it has been reported that irrelevant features do not affect behavioral error rate when the rule was externally indicated in a visual stimuli classification task (Archer, 1954). However, irrelevant features influenced performance when the rule was not externally indicated in the same task (Hodge, 1959). Since the subjects in the second task condition reported the relevant feature (rule) in the task, the subject knew what the relevant information was, hence, the effect of the irrelevant feature was not a result of confusion. However, still performance was affected by the irrelevant features of stimuli. This difference implies

that internal selection of rules enhances internal computations searching for potentially relevant sensory information while externally cued rules restrict these potential dynamics. Previous studies in non-human primates indicate that prefrontal internal dynamics vary in relation to implicit rule-learning and switching in PFC, parietal cortex, orbitofrontal cortex, and striatum (Goodwin et al., 2012; Crowe et al., 2013; Brincat and Miller, 2016; Sleezer et al., 2016).

To characterize the neural mechanisms in PFC that are involved in flexible computation of categories based on relationships between stimuli (rather than their features), and to elucidate the cognitive strategies involved, we trained two monkeys to perform a rule-selection categorization (RSC) task while recording in PFC. The RSC task is a rule-dependent categorization task, which requires computing the relationship between two visual stimuli flexibly depending on alternative categorization rules. In the task, monkeys categorized the relationship of a target stimulus to a reference stimulus along two feature dimensions, either SIZE or SPACE (position). Monkeys reported their categorical decisions by making a GO or NOGO decision to execute or withhold a saccadic response. The task included two conditions: (1) 'incongruent' trials in which application of the two rules to the target and reference stimuli yielded different responses (GO vs NOGO), in which case the response was rule dependent, and (2) 'congruent' trials, in which application of the two rules to the target and reference stimulus yielded the same response, in which case the response was rule independent. We fully randomized the stimulus positions, sizes and relationships to produce a very large number of stimulus combinations in order to discourage learning a look-up table. In addition, the categorization rule was not explicitly instructed in the task. Therefore, to solve the task

monkeys could not use the simple features of the stimuli but had to flexibly compute the relationships and infer the correct rule in each block using trial-and-error feedback. Furthermore, because the congruent and incongruent trials were introduced in random order in the task, monkeys could adapt their cognitive strategies depending on whether categorization along the two feature dimensions provided conflicting or reinforcing responses.

2

Rule Selection Categorization Task and Behavior

2.1 Introduction

How does the brain decide what information to use and how to integrate it? Utilizing the information available in different environments flexibly to make the best or the most efficient choice is an important process for human decision-making. In order to maximize rewards or benefit from decisions, the brain must dissociate relevant from irrelevant information among the massive incoming stream of sensory input and then decide how to assemble it. Criteria dividing incoming information into relevant and irrelevant is often based on an internal judgement instructed by goals or rules, and therefore, it is flexible rather than absolute.

The Stroop task provides insight into how the brain filters and integrates rule-relevant or rule-irrelevant visual information to respond correctly based on cognitive

rules (Stroop, 1936; Cohen et al., 1990). In the classic Stroop task, the reaction time to name the font color of written words was significantly shorter in a congruent condition when the color indicated by the word and the color of the font were the same (e.g. 'red' written with red letters), relative to a control condition in which the word did not represent a color. However, the reaction time to name the font color was significantly longer in a conflict condition when the color indicated by the word and the color of the font did not match (e.g. 'green' written with red letters) (Stroop, 1936; Cohen et al., 1990).

The above results indicate that the brain does not simply filter out the irrelevant information based on the current rule or cognitive set, but rather can integrate irrelevant information with rule-relevant information when it is determined to enhance performance. This implies that the brain generates internal strategies to maximize its performance. Further, irrelevant information appears to interfere with processing rule-relevant information in the conflict condition.

In the classic Stroop task, it was also interesting that the reaction time to read the word were not different between the control, congruent, and conflict conditions, and were significantly faster than the reaction time to name the font color (Stroop, 1936; Cohen et al., 1990). In the previous studies, this effect was named 'task conflict' and suggests that the pathway handling word information (language) is trained and, therefore, fundamentally stronger than the pathway handling color information in the brain (Cohen et al., 1990; MacLeod and MacDonald, 2000; Monsell et al., 2001; Monsell, 2003).

Accordingly, the classic Stroop task possesses two limitations in its ability to detect how rule-relevant and rule-irrelevant information strictly affect behavior or decision-making. First, because the rule and the task features (word and color) were prespecified and fixed (name the font color), the process by which relevant and irrelevant feature information interact or are combined to influence behavioral performance is not clear. Second, since the rule was fixed within the task, the brain was already restricted regarding the information it should use or to filter out. The Stroop task may decrease the dynamic use of information in the brain limiting our ability to investigate how information is flexibly utilized according to changing cognitive rules.

To evaluate how incoming information is flexibly classified as relevant or irrelevant and how this information is integrated or filtered through internal strategies to make a decision in the brain, we designed a rule selection categorization (RSC) task that required high-level cognitive processes in order to identify which incoming information was essential to ongoing cognitive operations using internal boundaries, judgements, or standards. In the RSC task, flexible computation, filtration, and integration were required to maximize behavioral performance. In addition, the task enabled us to analyze autonomous internal rule selection by trial-and-error feedback.

2.2 Methods

The RSC Task

Task design

Monkeys performed a task in which a reference stimulus first appeared, followed by a target stimulus. Both reference and target were circles of various sizes and positions in a visual display. Monkeys were required to categorize the target stimulus based on its relationship to a reference stimulus according to two different rules. One rule was the SPACE rule and the other rule was the SIZE rule (below). Within each block of trials, the rule switched once either from the SPACE rule to the SIZE rule, or from the SIZE rule to the SPACE rule. No explicit cue was provided indicating that the rule had switched or what current rule to apply to determine the category of the target (Fig. 2.5.1 A). Therefore, monkeys had to learn which rule was current using trial-and-error feedback. Under the SPACE rule the target could be categorized into a left or right category based on whether the target was located to the left or right of the reference stimulus, and under the SIZE rule the target could be categorized into a larger or smaller category based on whether the target was larger or smaller than the reference stimulus.

Trials began with a gaze fixation on a red central fixation target (Fig. 2.5.1). After 500 ms a reference stimulus (yellow circle) was presented for 500 - 700 ms. Next a target stimulus (yellow circle) was presented along with the reference stimulus for an additional 700 ms. Monkeys had to categorize the target based on relationship with the reference according to the current rule and once the central fixation target changed its color from red to green, monkeys had to report the category membership by making a GO or NOGO

saccade decision (if monkeys did not make saccade in 500ms to the peripheral target, the decision was considered as NOGO). If the target was categorized as 'left' or 'larger' according to the current rule, the required action was GO which was making a saccade from the central fixation target to the peripheral target stimulus. If the target was categorized as 'right' or 'smaller', the required action was NOGO which was maintaining gaze fixation on the central target until the end of the trial. After monkeys maintained gaze fixation either on the central target or the peripheral target for 500 - 1000 ms, a feedback ring was presented for 350 ms centered on the current eye position which was the location of the peripheral target for correct GO trials and the location of the central fixation target for correct NOGO trials. If the decided action matched with the assigned action, the color of the feedback ring was orange (indicating a correct trial) and a drop of juice (~0.1 mL) was given as a reward at the end of the trial. If the decided action did not match with the assigned action, the color of the feedback ring was blue and no reward was given (Fig. 2.5.1 B, C).

Each trial consisted of one reference and one target stimulus (Fig. 2.5.1 B, C) and new trial stacks were built for every behavioral session (corresponding to the recording of a single neural ensemble, Chapter 3). The trial stack building algorithm proceeded as follows. Step1: equal numbers of targets were newly assigned to each of 8 sectors surrounding the fixation target at random locations within the selected sectors and with random sizes everyday (Fig. 2.5.2 A). Step2: four reference stimuli were assigned to each target. The sizes and locations of the reference stimuli were randomly selected with the constraints that the target stimulus would be assigned to right/larger, left/smaller, left/larger, and right/smaller categories when related to each of the four reference stimuli

(Fig. 2.5.2 B). The center of the target and the reference stimuli presented at eccentricity 2.16°- 7.31° (Fig. 2.5.2 A). Step3: We replicated the stimulus set generated above by inverting the reference-target relationship between the two stimuli, doubling the number of trials. That is, for each pair of circular stimuli generated by the algorithm (stimulus A and B), we presented one trial with stimulus A preceding stimulus B (in which case A was the reference B the target), and one trial with stimulus B preceding stimulus A (in which case B was the reference and A target (Fig. 2.5.2 C). Step4: Exactly the same trial stack generated by the above was then presented once in a random order under the SPACE rule and again in random order under the SIZE rule in each block of trials. In each new block, a new trial stack was generated and presented under the two rules. Because of this algorithm the portions of GO and NOGO, left and right, and larger and smaller category trials were all equally 50% (Fig. 2.5.2 D).

In the RSC task, the design algorithm was designed to decorrelate categories from visual features as well as saccade direction and amplitude (Fig. 2.5.3). First, categories were visual-feature independent because the same target having the same color, size and location on the screen could be categorized into four different categories depending on the rules and the relationships with the corresponding reference stimuli (Fig. 2.5.3 A). Second, the categories were visual-pattern independent because each same visual pattern consisting of a single reference-target stimulus combination was assigned to four different categories (larger, smaller, left, right) with equal probability as a result of the inversion of the order of presentation (and hence reference/target status) of each stimulus pair, and the duplication of the complete set of reference/target pairs between the two rules (Fig. 2.5.3 B). Lastly, categories were decorrelated from saccade direction and

amplitude because saccade direction varied randomly with respect to category (targets were distributed equally among the 8 sectors), and on GO trials, each saccade was associated with left and larger categories with equal probability (e.g., in Fig. 2.5.3 C, note the three trials illustrate different saccade directions and amplitudes that all report assignment of the target to the 'larger' category).

Response congruency

The RSC task defines two relational category dimensions. The SPACE category dimension consists of left and right categories, and the SIZE category dimension consists of larger and smaller categories. In the task, the combinations of the two category dimensions create four two-dimensional categories considering the two categorical relationships defined by each pair of stimuli as a compound; left/larger, right/larger, left/smaller, and right/smaller (Fig. 2.5.6 A). Because each category has and assigned response to report that category under a given rule (e.g., GO response for the left category and NOGO for the right category under the SPACE rule), the two-dimensional categories fall into two different types of compound categories; congruent and incongruent. Congruent categories require the same response regardless of what categorization rule is applied and include the left/larger category, which requires the GO response under both the SPACE or SIZE rule, and the right/smaller category, which requires the NOGO response under both rules. Incongruent categories in contrast require conflict resolution since application of one rule to the compound category requires the GO response and the other rule the NOGO response. The incongruent categories include the left/smaller

category, which requires the GO response under the SPACE rule but the NOGO response under the SIZE rule, and the right/larger category, which requires the NOGO response under the SPACE rule but the GO response under the SIZE rule.

Data Analysis

Sliding-window learning curve and logistic regression analysis of switch behavior

We aligned trials within blocks to the trial on which the rule switched and computed the mean proportion of correct trials within a 20-trial sliding window to generate learning curves to capture improvements in performance (Fig. 2.5.5 A, C). Bin 1 (after the rule switch) including the switch trial and the preceding 19 trials. Bin 20 included the switch trial and the subsequent 19 trials. Therefore, bin1 to 20 included the switch trial. We applied logistic regression analysis to the behavioral data to model the log odds of making a GO response as a weighted function of the GO status of the trial under the rule before the switch (Pre: SPACE or SIZE), and the rule after the switch (Post: SIZE or SPACE) by fitting the following model (Fig. 2.5.5 B, D):

$$logit(P_{GO}) = \beta_0 + \beta_1 Pre_{GO/NOGO} + \beta_2 Post_{GO/NOGO}$$

Where $logit(P_{GO})$ refers to the log odds of a GO response by each monkey, '$Pre_{GO/NOGO}$' is the dummy coded correct GO/NOGO decision determined by application of the rule before the switch to each reference-target pair, and '$Post_{GO/NOGO}$' is the dummy coded correct GO/NOGO decision determined by application of the rule after the switch to each reference target pair.

Psychometric curve

I defined the perceptual difficulty of the left-right discrimination under the SPACE rule as the angle formed between the vertical axis and the line between the target and reference stimuli. Smaller angles were associated more vertically arranged reference-target pairs for which the left-right relationship was more difficult to determine (Fig. 2.5.7 A). Similarly, we defined the perceptual difficulty of the larger-smaller discrimination under the SIZE rule as the percent difference between the radii of the reference and target stimuli (Fig. 2.5.7 B). We plotted the mean percent correct performance on the last 100 trials of each rule block (when performance had reached an asymptotic level) as a function of the level of perceptual difficulty along the rule-relevant and rule-irrelevant feature dimensions on incongruent (Fig. 2.5.7 C-F) and congruent trials (Fig. 2.5.8 A-D). We analyzed differences in proportion correct across levels of difficulty within each psychometric curve using the Kruskal-Wallis test followed by post-hoc pairwise Tukey-Kramer tests.

Modeling strategy selection according to trial congruency on feature dimensions

I applied logistic regression analysis to performance in the last 100 trials of each block to evaluate how the log odds of a correct response depended on the perceptual difficulty of the category judgement along the relevant and irrelevant feature dimensions by fitting the following model:

$$logit(P_{cor}) = \beta_0 + \beta_1 Relevant + \beta_2 Irrelevant$$

Where $logit(P_{cor})$ refers to the log odds of correct response, '*Relevant*' is the difficulty of the relevant feature dimension which is the SPACE feature dimension under the SPACE rule and the SIZE feature dimension under the SIZE rule, and '*Irrelevant*' is the difficulty of the irrelevant feature dimension which is the SPACE feature dimension under the SIZE rule and the SIZE feature dimension under the SPACE rule. The model was applied on incongruent and congruent trials individually to compare the difference of the use of target features according to trial congruence. Once the models were fitted, each model (predicted percent correct performance) for incongruent and congruent trials was plotted with three axes which are proportion correct, relevant feature dimension, and irrelevant feature dimension in 3D space (Figs. 2.5.7 G, 2.5.8 E).

2.3 Results

The Rule Selection Categorization Task

The rule selection categorization (RSC) task is a rule-dependent categorization task that requires a flexible categorization of the same visual stimuli. One of the major differences between the RSC task and previous flexible rule-dependent categorization tasks (Goodwin et al., 2012; Mante et al., 2013; Roy et al., 2014) is that the RSC task required autonomous rule selection by trial-and-error feedback, while other studies externally instructed what categorization rule to apply. Therefore, the RSC task is unique by forcing monkeys to be flexible in deciding what information is relevant and irrelevant. Moreover, the task allows monkeys to filter or integrate information dynamically through their own internal strategies to make the best decision.

The RSC task consisted of two rules (SPACE and SIZE rule; Fig. 2.5.1 A). Each rule defined two relational categories (left or right category under the SPACE rule and larger or smaller category under the SIZE rule). Each category had an assigned action for reporting its category membership (left-GO, right-NOGO, larger-GO, and smaller-NOGO). Monkeys performed the task, which required them to categorize a target based on a relationship with a given reference stimulus according to the two different rules (Fig. 2.5.1 B, C). In each recording session, we switched the rule a single time—either from the SPACE rule to the SIZE rule or the SIZE rule to the SPACE rule. This was done without explicit cues indicating the rule-switching point or the current rule to apply. Therefore, monkeys had to learn which rule was current by trial-and-error feedback.

Trials began with gaze fixation on a red central target (Fig. 2.5.1 B, C; fixation). Two yellow circles were then presented in sequence. The initial stimulus was the reference and the subsequent stimulus was the target to be categorized according to the current rule and relationships with the given reference (Fig. 2.5.1 B, C; reference, target). Once the central fixation target changed its color from red to green, monkeys had to report the category membership through one of two required actions, either GO or NOGO (Fig. 2.5.1 B, C; decision). If the required action for the category membership was GO, the correct choice for monkeys was to make a saccade from the central fixation target to the peripheral saccade target (Fig. 2.5.1 B; decision). If the required action was NOGO, the correct choice was to maintain gaze fixation on the central target until the end of the trial (Fig. 2.5.1 B; decision). After gaze fixation on either the central fixation or peripheral saccade target for 500 - 1000 ms, a feedback ring was presented for 350 ms on the current eye position, which was the location of the peripheral target for a GO decision

or the location of the central fixation target for a NOGO decision (Fig. 2.5.1 B, C; feedback). If the decided action matched the assigned action, the color of the feedback ring was orange and a drop of juice (~0.1 mL) was given as a reward at the end of the trial. If the decided action did not match with the assigned action, the color of the feedback ring was blue, and no reward was given (Fig. 2.5.1 B, C; feedback).

Each trial consisted of one reference and one target stimulus (Fig. 2.5.1 B, C), and new trial stacks were built for every recording session. In a trial stack, equal numbers of targets of random sizes were newly assigned to random locations within each of eight sectors surrounding the fixation target (Fig. 2.5.2 A), and four reference stimuli of different sizes were assigned at random locations surrounding each target location such that the relationship of the target to the reference placed the target into the right/larger, left/smaller, left/larger, and right/smaller categories (Fig. 2.5.2 B). The center of the target and the reference stimuli presented at eccentricity 2.16°- 7.31° (Fig. 2.5.2 A). To prevent monkeys from categorizing reference-target pairs based on the visual pattern formed by the two stimuli, all trials were presented again with the target and reference appearing in reversed order. In this scenario, the final visual pattern was preserved, but the relationship of the target and reference were inverted (Figs. 2.5.1 B, C, 2.5.2 C). Once we built a trial stack with the above algorithm, the exact same trial stack was presented in a random order under the SPACE rule and SIZE rule in a trial block (Figs. 2.5.1 B, C, 2.5.2 D). Because of this algorithm, the portions of GO and NOGO, left and right, and larger and smaller were all equally 50% (Fig. 2.5.2 D). This design decorrelated category and rule from the features of the visual stimuli.

Monkeys autonomously select categorization rules

Regarding behavioral performance, the proportion correct was stable and high before the rule switch was applied. After the switch, performance dropped and showed a learning curve in both monkeys (Fig. 2.5.5 A, C). Monkey 026 performed better than monkey 037 before the rule switch and was quicker to learn the new rule after the switch (Fig. 2.5.5 A, C). This behavioral performance pattern matches with patterns of regression coefficients relating the log-odds of a GO responses to the GO status of the target-reference pair under the rule in force before the switch ($\mathrm{Pre}_{\mathrm{Go/NOGO}}$, orange; Fig. 2.5.5 B, D) and the GO status of the target-reference pair under the rule in force after the rule switch ($\mathrm{Post}_{\mathrm{Go/NOGO}}$, green; Fig. 2.5.5 B, D). In the last 100 trials before the rule switch, the regression coefficient of $\mathrm{Pre}_{\mathrm{Go/NOGO}}$ (β_1) increased as monkeys performance under the first rule in the block and the regression coefficient of $\mathrm{Post}_{\mathrm{Go/NOGO}}$ (β_2) decreased in both monkeys. However, after the rule switch, β_1 decreased but β_2 increased and the trend reversed with a crossing point at around trial 20 for monkey 026 and around trial 80 for monkey 037 (Fig. 2.5.5 B, D).

The RSC task has two relational category dimensions. The space category dimension consists of left and right categories, and the size category dimension consists of larger and smaller categories. In the task, the two category dimensions combine to create four two-dimensional categories: left/larger, right/larger, left/smaller, and right/smaller (Fig. 2.5.6 A). Because each reference-target pair defines to categories according to the two rules, and because each category has an assigned response for reporting its category membership under a give rule (e.g., GO response for the left

category and NOGO for the right category under the SPACE rule), the conjunction of the two categories defined by each reference-target pair fall into two different types of compound categories—congruent and incongruent depending on the whether the two categories defined by each reference target pair instruct the same (congruent) or different (incongruent) responses (Fig. 2.5.6 A). Congruent compound categories require the same responses regardless of the rules, including the left/larger category, which always requires the GO response under the both rules, and the right/smaller category, which always requires the NOGO response. Incongruent categories instruct conflicting responses depending on which categorization rule is in force (SIZE or SPACE), and include the left/smaller category, which requires the GO response under the SPACE rule but the NOGO response under the SIZE rule, and the right/larger category, which requires the NOGO response under the SPACE rule but the GO response under the SIZE rule.

There are two behavioral hypotheses as to how monkeys categorize the target based on this congruence. The first hypothesis is that monkeys may ignore congruence and pay attention only to the current, rule-relevant category dimension. In this case, monkeys would categorize targets based simply on the one-dimensional category distinction defined by the current rule (such as the left or right category under the SPACE rule or the larger or smaller category under the SIZE rule). The second hypothesis is that monkeys may be aware of congruency and pay attention to the both rule relevant and irrelevant category dimensions depending on whether they instruct conflicting responses or not. In this case, monkeys may categorize the target based on its two-dimensional category (such as the left/larger or the right/smaller categories). To test these hypotheses,

we compared behavioral performances on congruent trials and incongruent trials (Fig. 2.5.6 B, C). We found that performance on the incongruent trials started at around chance (~0.5) after the rule switch and then gradually increased. In contrast, performance on congruent trials remained at a high level (~1.0 for monkey 026 and ~0.8 for monkey 037) through the rule switch and afterwards and in both monkeys (Fig. 2.5.6 B, C). These behavioral results indicate that both monkeys categorized congruent and incongruent categories differently to solve the task. We combined the behavioral data of the two monkeys for further analyses because it showed the same basic pattern in relation to performance on congruent and incongruent trials (Figs. 2.5.5 A-D, 2.5.6 B, C).

One possibility to explain the improvement in performance on congruent relative to incongruent trials in figure 2.5.6 B and C is that on congruent trials, monkeys leveraged information from both relevant and irrelevant feature dimensions to make their categorical decision, since categorizing reference-target pairs along the two dimensions led to the same motor response. In that case, monkeys would categorize stimuli at the two-dimensional level as compound categories (e.g. 'larger-left) to maximize their rewards during the task. This would mean that monkeys paid attention to not only the current rule-relevant category (e.g., the left or the right space categories under the SPACE rule), but also the current rule-irrelevant category (e.g., the larger or the smaller size categories under the SPACE rule). To test how monkeys differentially processes rule-relevant and rule-irrelevant information on congruent and incongruent trials, we separated the behavioral data based on trial congruency and analyzed in the extent to which the two different feature dimensions influenced choice. For the behavioral analyses, we focused on the last 100 trials of each block. In the last 100 trials, the

performance on the incongruent trials were consistently over 0.8 in both monkeys (Fig. 2.5.6 B, C), which implies that monkeys knew what relational rule they had to apply to categorize the given target.

Categorization rules influence behavioral feature integration

In the RSC task, there are two relational feature dimensions (space and size relationships), along with different levels of difficulty in discriminating the relationship of the reference and the target along the two dimensions (Fig. 2.5.7 A, B). When the angle of the reference and the target is close to vertical, the discrimination on the space relationship, i.e. whether the target is to the left or right of the reference, is harder than when the angle is close to horizontal (Fig. 2.5.7 A). When the size ratio of the target and reference is closer to 1.0, it becomes harder to discriminate the size relationship, i.e. whether the target is bigger than the reference or smaller (Fig. 2.5.7 B). When the SPACE rule is the current rule that monkeys must apply to solve the task, the rule-relevant feature dimension is the space feature dimension and the rule-irrelevant feature dimension is the size feature dimension. Alternatively, when the SIZE rule is the current rule, these assignments are switched.

Psychometric curves are plotted as functions of relational feature dimensions in figure 2.5.7 C-F for incongruent trials and figure 2.5.8 A-D for congruent trials. The x-axis for the space relationship curves is the relational angle (°) of the target and the reference from vertical. The x-axis for the size relationship curves is the absolute value of the size difference (%) between the stimuli. If monkeys used a certain feature dimension

to compute the relationship between the two stimuli, the behavioral performance (proportion correct) on the harder trials should be lower than the performance on the easier trials in relation to that feature dimension. On incongruent trials, when monkeys were likely to optimize their performance by processing one feature dimension and ignoring the other, the relationship between performance and perceptual difficulty was seen only for the rule-relevant dimension. Under the SPACE rule, on incongruent trials, performance fell off markedly as the perceptual difficulty of the left-right categorical discrimination became more difficult (Fig. 2.5.7 D, F, SPACE relationship; Kruskal-Wallis test, $\chi^2 = 101.2$, df = 8, $p < 0.001$). Similarly, under the SIZE rule, on incongruent trials, performance was worse on trials when the larger-smaller categorical discrimination was more difficult, although the decrement in performance was not as steep as it was under the SPACE rule (SIZE relationship; $\chi^2 = 61.7$, df = 5, $p < 0.001$). Decreasing performance with increasing perceptual difficulty was not evident on incongruent trials along the rule-irrelevant feature dimension (performance as a function size relationship under the SPACE rule and the space relationship under the SIZE rule), implying that monkeys did not pay attention to the rule-irrelevant feature dimension to make their categorical decisions (Fig. 2.5.7 C, E, SPACE relationship; Kruskal-Wallis test, $\chi^2 = 13.8$, df =8, $p = 0.09$, SIZE relationship; $\chi^2 = 7.4$, df = 5, $p = 0.19$).

On congruent trials, when the two feature dimensions synergistically informed categorical judgements regarding the required GO or NOGO response, performance functions in relation to perceptual difficulty were shifted upward overall in relation to both the relevant and irrelevant feature dimensions (Fig. 2.5.8 A-D, compare Fig. 2.5.7 C-F), suggesting that monkeys leveraged both sources of information to make categorical

decisions. For example, performance in relation to the rule-relevant feature dimensions was better on the hard trials of congruent trials than on the hard trials of incongruent trials (Figs. 2.5.7-8. Kruskal-Wallis test, df = 1, $p < 0.001$). For example, the monkeys' performance in classifying the target into the left or right category under the SPACE rule was about chance (~0.5) on the incongruent trials that were hard, such as a 10° angle-relationship between the target and the reference (Fig. 2.5.7 D), whereas, their ability to perform the same classification on the congruent trials of the same difficulty was notably higher at ~0.8 (Fig. 2.5.8 B). This means that monkeys could discriminate the relationships of the stimuli better on the congruent trials than on the incongruent trials, suggesting that they utilized more information than just the rule-relevant feature dimension to solve the task on these trials.

To evaluate how monkeys utilized feature information differently depending on congruency, we fit the same logistic regression model to performance data on the congruent and incongruent trials individually and plotted the models predicting performance in relation to both relevant in irrelevant feature dimensions in 3D space (Figs. 2.5.7 G, 2.5.8 E). The model we applied was $logit(P_{cor}) = \beta_0 + \beta_1 Relevant + \beta_2 Irrelevant$, where $logit(P_{cor})$ refers to the log odds of correct response, '*Relevant'* is the perceptual difficulty along the rule-relevant feature dimension, and '*Irrelevant'* is the perceptual difficulty along the rule-irrelevant feature dimension. The 3D logistic model fits (colored surface) of performance on the incongruent trials indicate that monkeys used only the relevant feature dimension to solve the task (Fig. 2.5.7 G, $\beta_1 = 0.2752$, $p < 0.05$; $\beta_2 = 0.0016$, $p = 0.92$). However, the 3D logistic model fits of performance on the congruent trials indicate that monkeys used both the relevant and the irrelevant feature

dimensions to solve the task (Fig. 2.5.8 E, β_1 = 0.3368, $p < 0.05$; β_2 = 0.1385, $p < 0.05$). Altogether the psychometric curves and the 3D model plots show that monkeys used different strategies on the incongruent and the congruent trials. The data also demonstrate that on congruent trials (trials in which the assigned responses were the same regardless of the rule), monkeys additionally utilized information from the irrelevant feature dimension to maximize the accuracy of their categorical decisions.

2.4 Discussion

The amount of incoming sensory information to the brain such as taste, smell, vision, hearing, or touch is tremendous in every moment even though we are not aware of how vast the amount of information is. Because the brain cannot digest every single piece of sensory information that it has access to, the information must be selectively processed under higher-level cognitive control based on what we want, what the surrounding circumstances are, or what our previous experience was. To make the best choice, the brain handles incoming information flexibly, decides how to use the refined information dynamically, and then takes action. However, we still have incomplete understanding of how information is selectively processed, in particular as needed to advance ongoing cognitive operations, and how the processed information is combined to influence our actions. To answer these questions with behavioral evidence, in this chapter, we evaluate how monkeys flexibly classify incoming information as relevant or irrelevant by application of implicit rules that monkeys spontaneously developed to dynamically

integrate information based on its varying relevance to the cognitive process of categorization.

The primary findings we report here are the following. First, monkeys can learn abstract categorization rules and to selectively apply these rules autonomously using trial-and-error feedback without explicit indication of the current rule provided by an external cue. Second, monkeys can judge what information is relevant or irrelevant to the cognitive operation of categorization under the current rule, and can develop and apply their own strategies of how to integrate or filter out relevant and irrelevant information flexibly depending on different trial conditions.

Previous studies have shown that top-down control of selective attention focuses information processing on goal-relevant information, with improved performance being critically dependent on the ability to ignore or filter out irrelevant information (Ploner et al., 2001; Zanto and Gazzaley, 2009; van Moorselaar and Slagter, 2020). In our study, we find instead that monkeys use rule-irrelevant information flexibly when that information can enhance decision making, even if contradictory to the current cognitive rule. This difference in how irrelevant feature information influences choice could reflect differences in task design. In the prior human studies, an external cue specified the task rule to apply that determined what information was relevant to behavioral choice. It could be that external cueing restricted how the brain sampled sensory input. In the RSC task, the rule was internally determined and fewer external constraints imposed. This may have enabled the brain to adapt information sampling strategies to maximize performance.

The RSC task is to my knowledge the first rule-dependent relational categorization task developed without explicit rule indication, as well as the first to explore autonomous rule selection under conditions when feature and category task variables were clearly separated by basing categories on relationships between stimuli, as well as the first to relate categorical performance to perceptual difficulty along relevant and irrelevant feature dimensions. These task conditions allowed monkeys to generate dynamic strategies and allowed us to evaluate how sensory information was processed to influence cognition in a strategy-dependent manner. With our task design, monkeys could not solve the task by using a look-up table or by using simple visual features or patterns. Therefore, monkeys were required to compute the relationships between paired stimuli in order to solve the task. The relationships of stimuli were not derived passively by visual input, but rather monkeys had to compute the relationships actively and in a goal-oriented manner to make categorical decisions. In order to make the best and most economical choice to maximize reward with minimal effort, monkeys were given a choice to decide which relationships to calculate. Therefore, it is interesting that monkeys still computed the rule-irrelevant feature dimensions and utilized this information to make categorical decisions on congruent trials even when they knew which rule was current (Fig. 2.5.8 E). We can conclude that monkeys knew the current rule, particularly in the last 100 trials, because the psychometric curves on incongruent trials show that monkeys selectively processed the relevant feature dimension according to the current rule (Fig. 2.5.7 D, F) and the logistic regression model of GO/NOG response probability clearly switched in a rule-dependent manner (Fig. 2.5.5 B, D).

Division of incoming sensory information along relevant and irrelevant feature dimensions has been a topic for learning theory and concept-formation studies (Archer, 1954; Hodge, 1959; Rabbitt, 1964; Razik, 1971). In prior concept-formation studies, it was reported that irrelevant feature did not affect the behavioral error rate when classifying visual stimuli in the case that irrelevant-feature dimensions were never used to classify the stimuli and the rule was instructed. However irrelevant features did influence performance if irrelevant feature dimensions were sometimes used to classify stimuli under alternative rules, and in the case that the rule was not instructed (Archer, 1954; Hodge, 1959; Rabbitt, 1964; Razik, 1971), a set of conditions directly analogous to those operating in the RSC task. In this second group of studies, subjects were required and were able to report which dimension was relevant to classify the visual stimuli, which means that the subjects knew what dimension to pay attention to, but the irrelevant feature dimension still affected their performance.

In this chapter's introduction, I mentioned a potential limitation in the Stroop test, in which the rule is explicitly instructed, so that the brain can automatically restrict what information to use or to filter out, therefore limiting the requirement to dynamically select and use information flexibly according to changing cognitive strategies. Also, I mentioned that given this limitation, I assume that our RSC task is a good model to investigate how the brain adaptively interrogates sensory input to derive specific items of information needed to meet changing cognitive demands in order to maximize behavioral performance.

Not only does the behavioral data provide support for the hypothesis that monkeys performing the RSC task flexibly analyze sensory input to meet the information

demands of ongoing cognitive processes, (Figs. 2.5.6 B, C, 2.5.7, 2.5.8), but human studies provide support that humans exhibit similarly flexible sensory processing when cognitive demands change (Archer, 1954; Hodge, 1959; Rabbitt, 1964; Razik, 1971). As in these previous studies, monkeys' behavioral performance in the RSC task was affected by the irrelevant feature dimension under conditions in which the rule was not explicitly indicated. The task design and behavioral results were similar between the RSC task and the prior human studies except that the RSC task involved relational categorization and monkeys did not provide a behavioral response indicating what feature dimensions were chosen as relevant in order to make categorical decisions. The interesting result from the human studies is that the subjects knew what feature dimension was relevant, but their performance was still affected by the irrelevant feature dimension. Based on similarities with the human studies, it seems possible that monkeys may have cognitively identified feature dimensions as irrelevant to the current rule but been influenced by them nonetheless

Further evidence that irrelevant information can influence cognitive processing is given by the classic Stroop test (Stroop, 1936; Cohen et al., 1990), demonstrating that reaction time is longer if the written word and font color conflict in comparison to when they agree, indicating that the brain does not simply filter out irrelevant information (provided by the written word in this case), but integrates irrelevant information to either enhance or disrupt performance. We did not measure the reaction time in the RSC task because the response of the animal was delayed until presentation of a 'GO' signal suggesting that differences in processing times would be obscured by the delay imposed before the animals were allowed to respond. However, logistic regression models (Figs.

2.5.7 G, 2.5.8 E) as well plots of performance over trials since the rule switch exhibited differences on congruent and incongruent trials (Fig. 2.5.6 B, C) that bore some similarity to the reaction time pattern observed in the Stroop test. The irrelevant features were not simply filtered out, but rather were integrated with the relevant features on the congruent trials to maximize behavioral performance (Fig. 2.5.8).

In this chapter, I introduced hypotheses as to how and when relevant and irrelevant feature information is defined and used flexibly to make the best decision in the RSC task. However, there are still a few points to address. On congruent trials, performance remained at a high right through the rule switch in both monkeys (Fig. 2.5.6 B, C). Because I only analyzed behavior in the last 100 trials for the psychometric curves and the logistic regression modeling, it is unclear how features were flexibly used on congruent trials during the learning period, which was the period when the rule was not clear for the monkeys. Also, we do not know whether use of irrelevant feature information was restricted to trials on which category discriminations along the relevant dimension were particularly difficult, or whether irrelevant feature information was always used and integrated with relevant feature information nonselectively on congruent trials. This question could be answered by stratifying trials according to perceptual difficulty along the relevant feature dimension, and testing the hypothesis that irrelevant feature information had a stronger influence on performance when the discrimination along the relevant feature dimension was more difficult.

We also do not know what the exact influences correct and incorrect feedback have in the learning period and whether this feedback modifies how features are mapped to categories as performance improves. We can readily assume that correct feedback will

give a hint that the current strategy is correct and reinforce the current strategy. Incorrect feedback will give a clue that the current strategy is wrong, and hence, the strategy should be modified. The learning curves and the response model (Figs. 2.5.5, 2.5.6) indicate that monkeys actually switched their internal representations of the relevant rule, but what makes monkeys switch their strategy? More specifically, on error trials, how did monkeys judge whether the applied rule was correct but their perception was wrong, or the applied rule was wrong but their perception was correct? Even in the last 100 trials, monkeys still made errors but the errors did not make them switch the current rule. Further investigation is needed to answer this with behavioral modeling. In the next chapter, we will address this topic by discussing the properties of neural data recorded in prefrontal cortex during task performance.

Lastly, we also need to clarify whether monkeys used the relevant and irrelevant features alternatively on the congruent trials, combined the two feature dimensions after they are individually computed, or generated new feature dimension by integrating the two feature dimensions. This question will be discussed further in the next chapter as well when we discuss the neural correlates in prefrontal cortex.

2.5 Figures

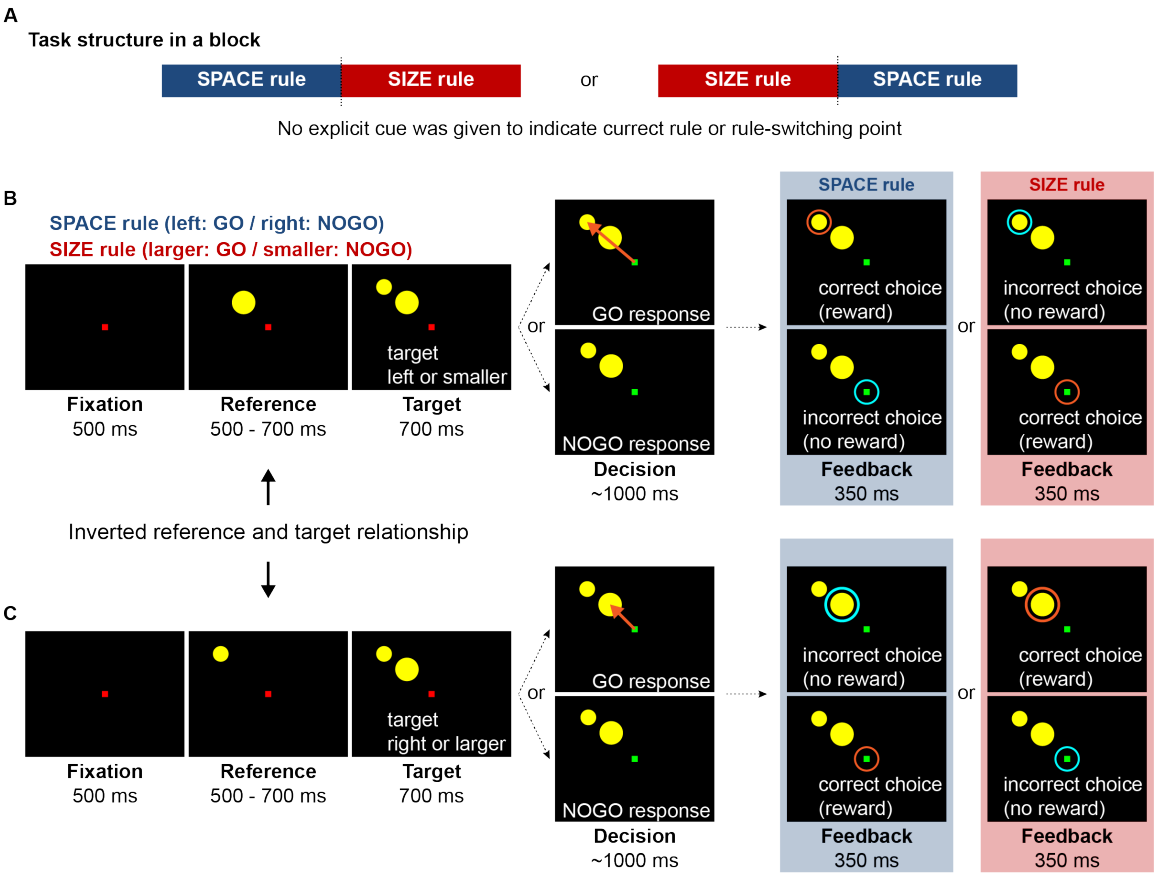

Figure 2.5.1 ***The rule selection categorization (RSC) task***

A. *Rule sequence. Two rules (SPACE rule, SIZE rule) were introduced sequentially in a block and the order of the rules was randomly selected in each block. (Blocks started with the SIZE rule switched to the SPACE rule part way through the block, or vice versa.) No explicit cue was given to indicate the rule-switching point or which rule to apply to categorize the targets. Rather, the subject learned which rule is current by trial-and-error feedback.* ***B.*** *Task design and trial sequence in the RSC task. There were two rules (SPACE rule, SIZE rule), two categories for each rule (SPACE rule: left / right, SIZE rule: larger / smaller), and two assigned behaviors (GO: make saccade to target, NOGO:*

fixate on a center dot). Under the SPACE rule the assigned behaviors were left-GO and right-NOGO and under the SIZE rule the assigned behaviors were larger-GO and smaller-NOGO. The order of events in every trial was as follows: (i) fixation: 500ms eye fixation on center dot was required to start a trial; (ii) reference: the reference circle was presented for 500-700ms on the screen; (iii) target: the target circle was presented with the reference for 700ms; (iv) decision: when the color of the center dot changed from red to green, the subject had to either fixate his eyes on the center dot (NOGO) for 1000ms or make saccade to the target (GO, if the subject did not make saccade in 500 ms, the decision was considered as NOGO) and fixate his eyes on the target for 850ms. The subject made a choice depending on its internal representation of the rule and judgement about the category membership of the target based on the relationship with the given reference; (v) feedback: the feedback ring was presented for 350ms whether the choice was correct or an error on the current eye position (correct: orange color / error: blue color). If the choice was correct, the subject was rewarded at the end of the trial; If the subject broke center fixation before the decision or broke center or target fixation after the decision, before the end of the feedback period, the trial was aborted. ***C.*** *Example of an inverted-order trial. Panel C trial has the same stimulus pair as panel B, but the order in which the stimuli are presented is reversed, inverting the target-reference relationship. For example, under the SPACE rule, the same visual pattern of reference and target stimuli can indicate either the left spatial relationship (panel B), or the right spatial relationship (panel C), depending on the order in which the two stimuli were presented. All trials were generated as inverted pairs of this type, decorrelating the category implied by the display from the visual pattern presented by reference and target circles.*

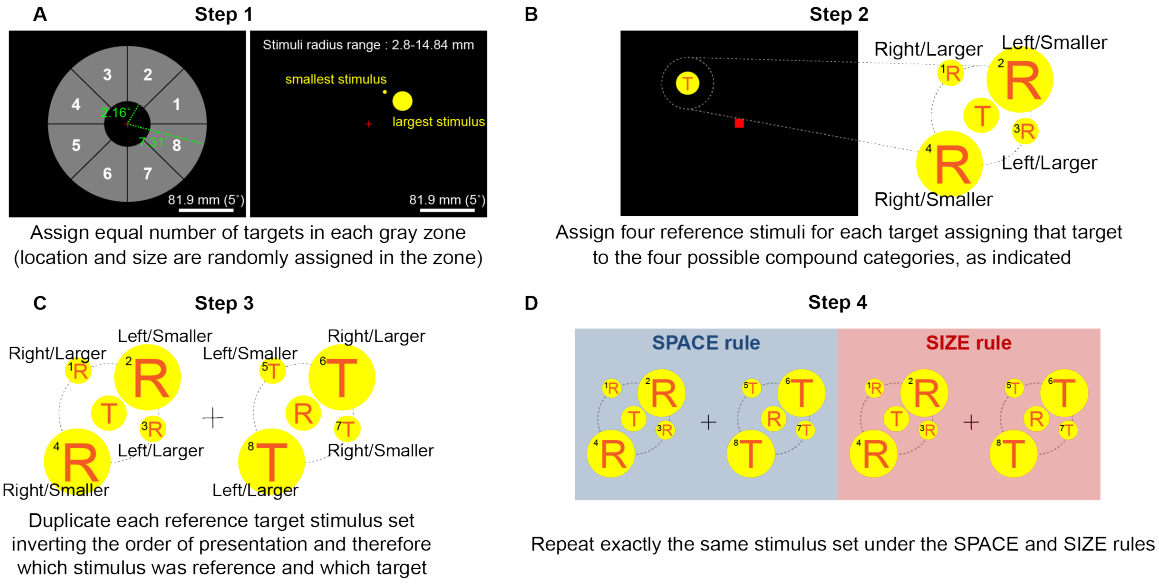

Figure 2.5.2 Trial stack design

A. *Step1: equal numbers of targets were newly assigned to each of eight sectors surrounding extending in eccentricity from 2.16°- 7.31° of visual angle from the fixation target. Targets were of randomized size and were positioned at random locations within the sectors. The scaled smallest and largest stimuli in the task were presented in the second panel.* ***B.*** *Step2: each target was assigned four reference stimuli such that the target, in relation to each reference stimulus, belonged to the four possible compound relational categories right/larger, left/smaller, left/larger, and right/smaller.* ***C.*** *Step3: Duplicate the each set of reference-target stimuli inverting the order of presentation, and therefore the reference-target relationship.* ***D.*** *Step4: Duplicate the stimulus set again and present once under the SPACE rule and once under the SIZE rule within each block. In the final trial stack, half of the trials were GO and half NOGO trials, half were left and half right trials, and half were larger and half smaller trials. Therefore, categories and responses were equiprobable in the trial stack.*

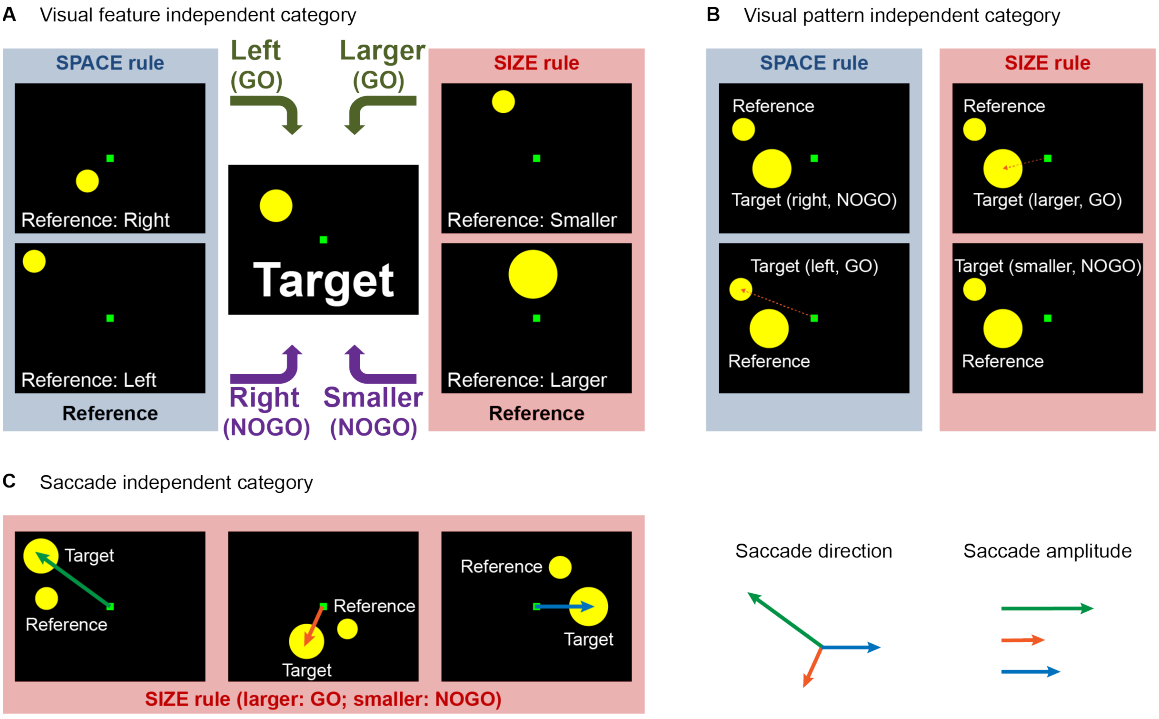

Figure 2.5.3 The categories in the RSC task are decorrelated from visual feature, pattern, saccade direction, and amplitude

A. *The categories are visual-feature independent because each target of a given color, size and location on the screen was assigned to four different compound relational categories with equal probability depending on the rule and which reference it was paired with.* ***B.*** *The categories are visual-pattern independent because each stimulus combination was assigned to four different categories by inverting the order in which the stimuli were presented, and hence the reference-target relationship, as well as by presenting each stimulus pair under the SPACE and SIZE rules.* ***C.*** *The categories were decorrelated from the motor response because each target location and size was associated with GO and NOGO responses with equal probability, and each target location and size was associated with each of the four compound relational categories with equal probability. This decorrelated saccade direction and amplitude from*

category. In the example shown, three saccades of different directions and amplitudes all assign the target to the 'larger' category under the SPACE rule.

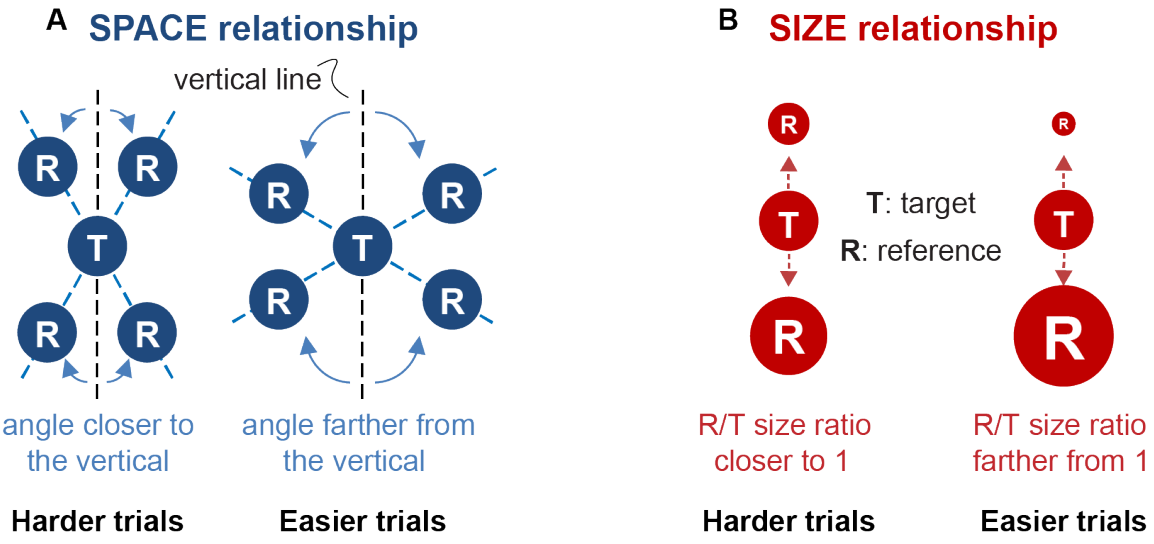

Figure 2.5.4 Space and size feature dimensions

In the RSC task, there are two relational feature dimensions (space and size). Reference-target pairs vary in perceptual difficulty along each dimension. ***A.*** *The space relationship. When the angle formed between the reference and the target stimulus is closer to vertical, the left-right spatial relationship discrimination is more perceptually challenging than when the angle is farther from vertical (closer to horizontal).* ***B.*** *The size relationship. When the size ratio between the target and the reference radii is closer to 1.0, the larger-smaller relationship discrimination is more perceptually challenging than when the ratio is farther from 1.*

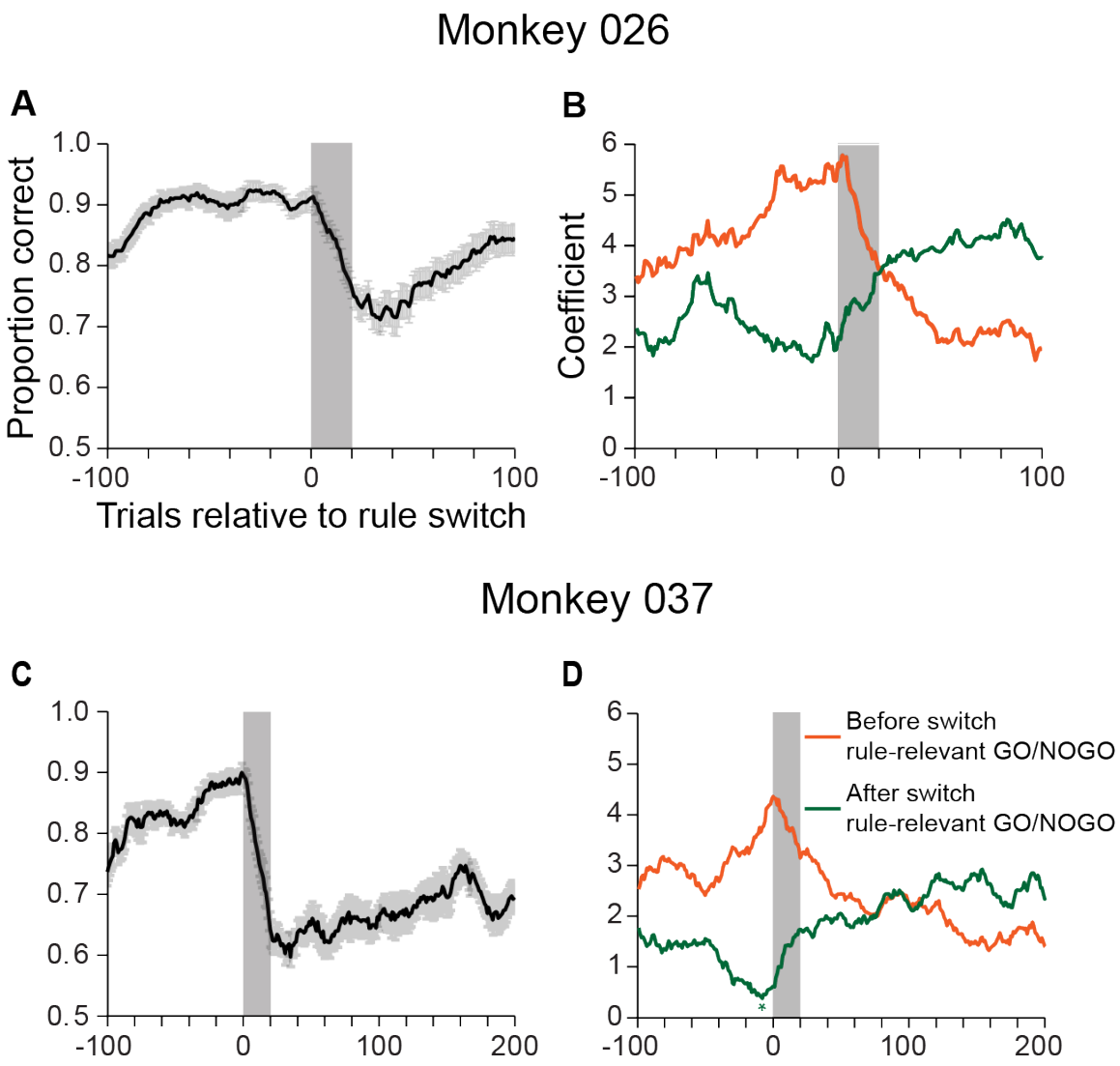

Figure 2.5.5 Learning curves and behavioral models

A, C. *Learning curves for monkey 026 (A) and monkey 037 (C). Mean proportion of correct displayed as a function of trial number within a sliding window of 20 trials. Gray shading indicates SEM. Points within the vertical gray band indicate mean performance over a 20-trial window including the switch trial.* ***B, D.*** *Mean logistic regression coefficients within a 20-trial sliding window for monkey 026 (B) and monkey 037 (D). We fit the following logistic model:*

$$\boldsymbol{logit}\,(\boldsymbol{p_{GO}}) = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 Pre_{GO/NOGO}} + \boldsymbol{\beta_2 Post_{GO/NOGO}}$$

The model relates the log odds of a GO response to the GO status of the trial according to application of the rule before the switch to each reference-target pair ($\boldsymbol{\beta_1}$*, orange) and the GO status of the trial according to application of the rule after the switch (*$\boldsymbol{\beta_2}$*, green).*

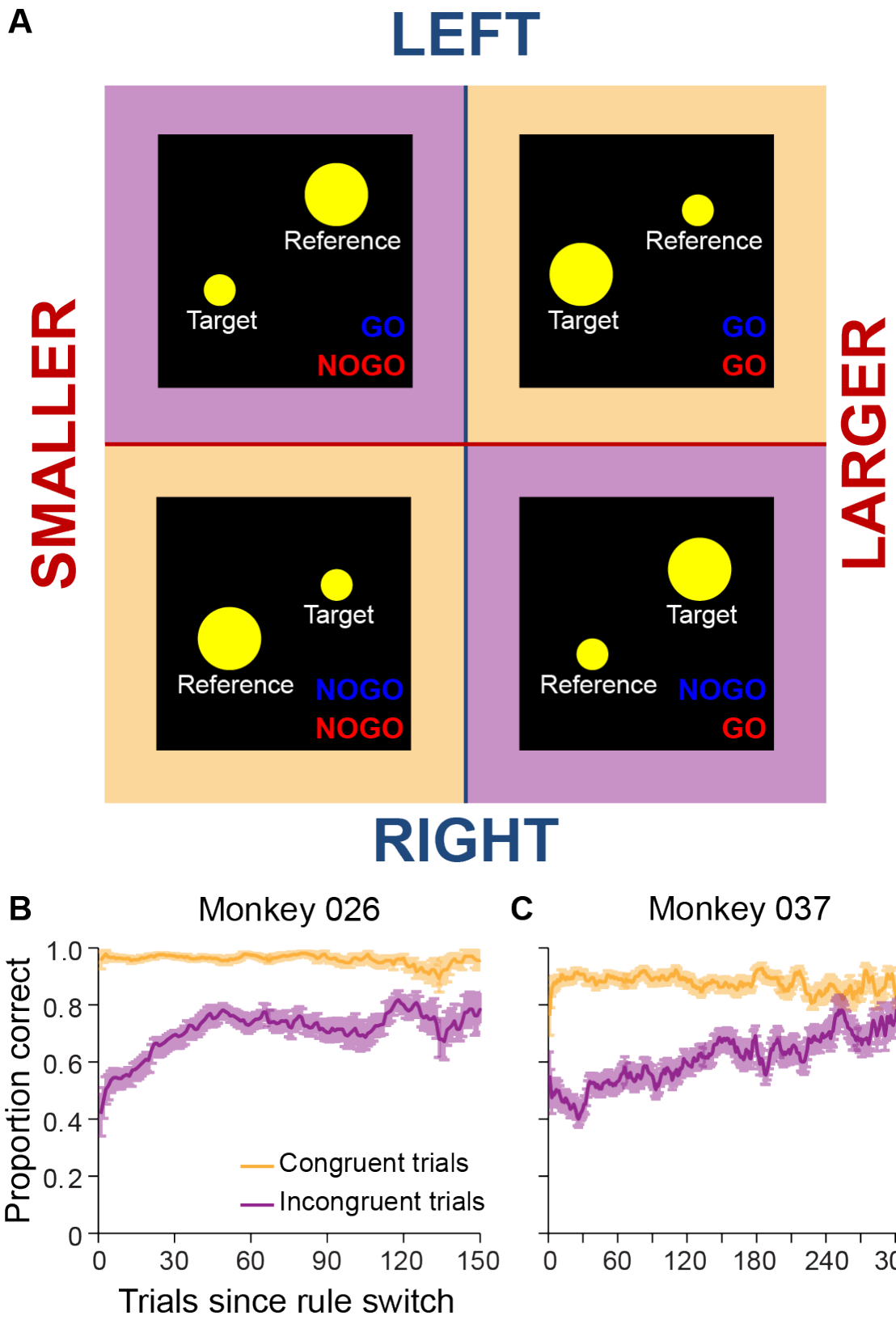
A
LEFT
SMALLER
LARGER
RIGHT
Reference
Target
GO
NOGO
Reference
Target
GO
GO
Target
Reference
NOGO
NOGO
Target
Reference
NOGO
GO
B
Monkey 026
C
Monkey 037
Proportion correct
1.0
0.8
0.6
0.4
0.2
0
Congruent trials
Incongruent trials
0
30
60
90
120
150
Trials since rule switch
0
60
120
180
240
300

Figure 2.5.6 Response congruency table and separated learning curves

A. *Response congruency. The RSC task defined two relational category dimensions. The space category dimension consists of left and right categories, and the size category dimension consists of larger and smaller categories. In the task, the combinations of the two category dimensions create four two-dimensional compound categories; left/larger, right/larger, left/smaller, and right/smaller. Because each category has a response assigned to it in order to report membership of the reference-target pair in that category under a given rule (e.g., GO response for the left category and NOGO for the right category under the SPACE rule), the two-dimensional categories fall into two different classes with respect to response congruency. Congruent categories (orange boxes) require the same responses regardless of which rule (SIZE or SPACE) is applied to the reference-target pair and include the left/larger category, which requires the GO response under the both rules, and the right/smaller category, which requires the NOGO response under both rules. Incongruent categories (purple boxes) require different responses depending on which rule is applied to categorize the reference-target pair, and include the left/smaller category, which requires the GO response under the SPACE rule but the NOGO response under the SIZE rule, and the right/larger category, which requires the NOGO response under the SPACE rule but the GO response under the SIZE rule.* ***B, C.*** *Learning curves plot proportion correct trials as a function of trials since the rule switch, separated by response congruency for monkey 026 (B) and monkey 037 (C). Orange indicates performance on congruent trials, purple on incongruent trials (shading indicates SEM).*

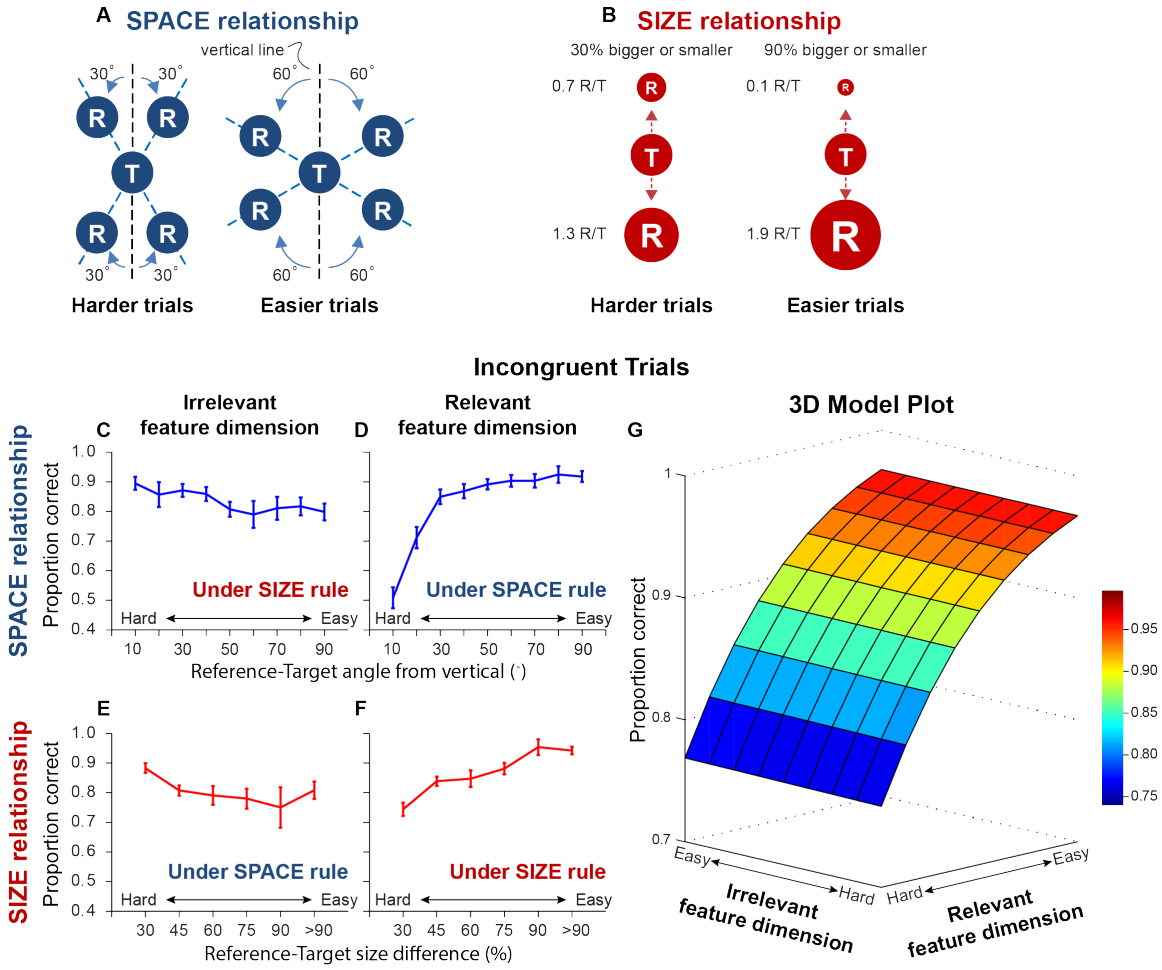

Figure 2.5.7 Relation of performance to perceptual difficulty of categorical discriminations on incongruent trials

A. *The space feature dimension. The perceptual difficulty of the left-right relational category distinction is harder when the reference-target angle is closer to the vertical.* ***B.*** *The size feather dimension. The perceptual difficulty of the larger-smaller relational category distinction is harder when the reference and target stimuli are closer to the same size, and the ratio between their radii closer to 1.* ***C, D.*** *Proportion correct performance as a function of the perceptual difficulty of the left-right category discrimination. Perceptual difficulty is expressed as the angle (in degrees) formed by the*

line between the reference and target stimuli and the vertical axis. Psychometric data are plotted separately for trials when the spatial feature dimension was irrelevant (C; under the SIZE rule) and relevant (D; under the SPACE rule) to behavioral choice. ***E, F.*** *Proportion correct performance as a function of the perceptual difficulty of the larger-smaller category discrimination. Perceptual difficulty is expressed as the difference in size between the target and reference stimuli (expressed as a percentage difference of the reference target radius). Psychometric data are plotted separately for trials when the size feature dimension was irrelevant (E; under the SPACE rule) and relevant (F; under the SIZE rule) to behavioral choice.* ***G.*** *3D surface representing the performance (proportion correct trials) predicted by a logistic regression model fit to the behavioral data on incongruent trials. Logistic regression analysis was applied to evaluate whether the log-odds of a correct response on the last 100 trials under given rule (z-axis of 3D plot) depended on the perceptual difficulty of categorical discriminations along the relevant and irrelevant feature dimensions as defined by the current rule (x- and y-axes of the 3D plot). We fit the parameters of the following logistic model:* $logit(P_{cor}) = \beta_0 + \beta_1 Relevant + \beta_2 Irrelevant$*, where Relevant and Irrelevant indicate perceptual difficulty of the category discrimination along the rule-relevant and rule-irrelevant feature dimensions respectively. The logistic regression analysis indicated that on incongruent trials, performance depended on the relevant but not the irrelevant feature dimension* ($\beta_1 = 0.2752$, $p < 0.05$; $\beta_2 = 0.0016$, $p = 0.92$).

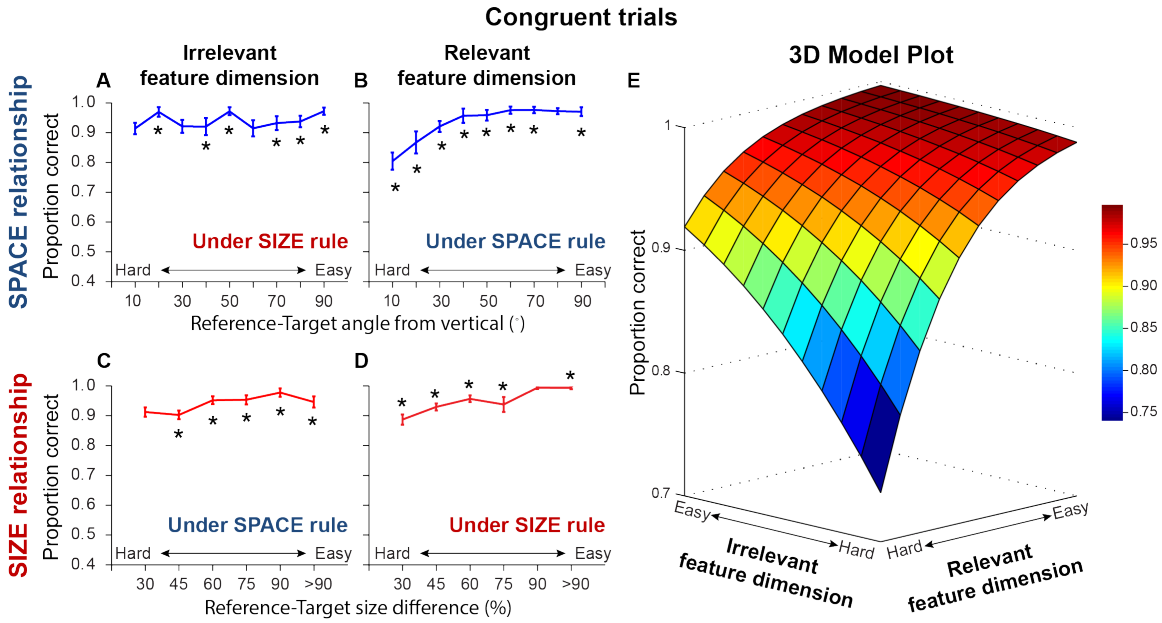

Figure 2.5.8 Relation of performance to perceptual difficulty of categorical discriminations on congruent trials

Conventions as in Fig. 2.5.7 above. Asterisks indicated that performance on congruent trials is significantly different from performance on incongruent trials at points along x-axis. The logistic regression analysis indicated that on congruent trials, performance depended on both the relevant and irrelevant feature dimensions ($\beta_1 = 0.3368$, $p < 0.05$; $\beta_2 = 0.1385$, $p < 0.05$).

3

Neural Correlates of Strategy-Dependent Feature to Category Transformation in Primate Prefrontal Cortex

3.1 Introduction

Categorization is a process by which the brain divides stimuli into meaningful groups that share common attributes, and it is fundamentally important to human cognition. Once we categorize an object, we can retrieve knowledge about the category to predict the behavior of the object and therefore direct our behavior in relation to it. Language includes a rich set of symbols to convey categorical knowledge (nouns and verbs for example are symbols for categories of objects and events). In spite of its importance, we still have incomplete understanding of how neural systems in the primate brain categorize sensory input. At its most fundamental and mechanistic level, categorization must involve a mapping from neural signals that encode the features of

stimuli to neural signals that encode the set of potential categories to which the stimulus could be assigned based on its features. Since any single stimulus can belong to a multitude of potential categories, flexible categorization implies a choice about which features will be analyzed and where boundaries will be placed along sensory continua to compute category membership. This places categorization under executive control. How we categorize an object reflects what our goals are, and what rules we apply to compute category membership, and this can change rapidly as environmental conditions change.

Approaching categorization from this perspective highlights the computation mapping stimulus features to stimulus categories as the core computation, the essential step. Although prior studies have investigated the neural basis of flexible and rule-based categorization in primate PFC (Goodwin et al., 2012; Crowe et al., 2013; Mante et al., 2013; Roy et al., 2014), we still have incomplete understanding of the nature or neural mechanisms of this transform. First, to effectively address how features map to categories, it is necessary to cleanly dissociate neural signals encoding features and categories. Since the features of a stimulus determine its category membership, the two are difficult to dissociate. Prior studies have shown that PFC neurons exhibit dichotomous, step-like responses to stimuli that vary along a feature continuum, such as shape (Freedman et al., 2001, 2003; Wutz et al., 2018) or direction of motion (Mante et al., 2013; Sarma et al., 2016). However, since categories correspond to ranges of feature values, feature tuning may still contribute to category selectivity, and step-like rather than continuous tuning to stimulus features is difficult to demonstrate on a cell by cell basis. Another approach to isolate category signals is to investigate how imposition of orthogonal category boundaries on the same stimulus set modifies the category signals

the stimuli evoke (Goodwin et al., 2012; Mante et al., 2013; Roy et al., 2014). These data demonstrate that the mapping between stimulus features and categories is flexible on a rapid time scale, but they do not entirely dissociate feature and category signals nor directly address how the one is mapped to the other. For example, in some prior studies, the category of a stimulus corresponded to the direction of the motor response that reported category membership (Antzoulatos and Miller, 2011). In others, the category of a stimulus corresponded to ranges of feature values relating to retinal location (Goodwin et al., 2012) direction of visual motion (Mante et al., 2013; Sarma et al., 2016), color (Mante et al., 2013) or shape (Freedman et al., 2001; Wutz et al., 2018). Even in cases where feature-to-category relationships changed according to a rule (showing that this relationship was flexible), within a given rule, feature and category signals were incompletely dissociated.

One way to address this question is to investigate the specific domain of relational categories. Categorical relationships, such as larger than, or brighter than, that exist between two objects depend on the direction of the comparison, which object is being categorized, and which is serving as a reference. Therefore by inverting the relationship between categorized and reference objects, the categorical relationship between the objects can be inverted without changing the features of the objects themselves. This facilitates dissociating neural signals that encode features and categories.

Here we investigate the neural representation of categorical relationships between stimuli in primate prefrontal cortex under conditions in which one of two different relationships could be relevant to behavioral choice. In one condition, monkeys computed the categorical size relations ‘larger than’ and ‘smaller than’. In the other, monkeys

computed the categorical spatial relations ‘right’ and ‘left’ (with respect to a reference object). This enabled us to dissociate neural signals coding these categorical relationships from other signals encoding the features of individual stimuli. The categorization rule (size or space) changed in blocks of trials, and monkeys determined which categorization rule to apply based on trial-and-error feedback. Here we show that the mapping of feature to category signals is modulated by cognitive rules. This suggests that executive control operates at a circuit level to modify the computational properties of prefrontal circuits.

3.2 Methods

Neural Recording

The target recording region was the principal sulcus in the dorsolateral prefrontal cortex (dlPFC, Brodmann area 9 and 46) in the left hemisphere. Prior to surgery and recording 3T structural MRI and CT images of monkeys were obtained to localize the target region. We employed the Circerone software package to generate rotatable 3D MRI volumes from the 2D image stacks for each animal (Miocinovic et al., 2007). Cicerone superimposes electrode trajectories on the 3D MRI reconstructions based on the stereotaxic coordinates and the angle of implanted recording chambers. We used this feature to target electrode trajectories to prefrontal cortical targets within the dorsal and ventral banks of the principal sulcus (Brodmann area 46) as well as the medially adjacent convexity of the dorsolateral prefrontal cortex (Brodmann area 9). To confirm recording locations at the end of recordings we inserted a stainless-steel probe into the center of the cluster of recording sites in prefrontal cortex and obtained a post-recording CT scan. Co-

registering the CT scan with the visible image of the electrode trajectory relative to the recording chamber and MRI structural image stacks in Cicerone allowed us to visualize representative electrode trajectories relative to the sulcal anatomy of prefrontal cortex. We confirmed electrode recording locations within Brodmann's areas 9 and 46 using this approach.

To prepare monkeys for acute neural recordings, they underwent an aseptic surgical procedure under isoflurane gas anesthesia (1-2%). We made a craniotomy at stereotaxic location selected using Cicerone over the dorsolateral prefrontal cortex in the left cerebral hemisphere in both monkeys. Five titanium points to support a stainless-steel halo for stabilizing head position were attached to the skull with titanium screws. We positioned a 19 mm i.d. PEEK plastic recording chamber at the proper angle over the craniotomy, and the posts, recording chamber, and screws were covered by a layer of surgical bone cement. Monkeys were administered injectable analgesia (Buprenex, 0.05mg/kg) twice a day for several days postoperatively.

Neurophysiological recordings were obtained using 16-channel or 32-channel silicon vector arrays ('Edge', 16 electrodes in a linear array, 100 µm spacing; 'Poly2', 2 staggered 16 electrode linear arrays, 50 µm spacing; NeuroNexus Technologies Inc.; Vector Array; Ann Arbor, MI). Electrode arrays were attached to the end of a stainless-steel rod (400 µm O.D.) and advanced through a stainless-steel guide tube (600 µm O.D.) penetrating the dura using a motorized Microdrive (NAN Instruments; NAN C Drives; Nazareth Illit, Israel). We advanced the electrode array until spiking activity was evident on the majority of electrodes, and recorded the activity of that ensemble of neurons for one block of RSC trials including trials performed under both the SIZE and SPACE rules

(in randomized order). We then advanced the electrode to isolate the activity of a new neural ensemble and recorded activity for an additional block of RSC trials. Neural data was recorded via a 64-channel digital signal processing data acquisition system (Tucker Davis Technologies Inc., Alachua, FL). Amplified and filtered electrode signals were digitized at a frequency of 24kHz and saved to disk with time synchronized eye position signals as well as coded time stamps indicating when specific stimuli were displayed, and when specific (target or nontarget) saccade responses were made. Kilosort Suite (open-source software package, Pachitariu et al., 2016) was used for off-line sorting of the action potentials of individual neurons based on a principle components analysis of recorded signals.

Data Analysis

Identification of neurons encoding stimulus features and categories

We applied analysis of covariance (ANCOVA) to identify neurons in which firing rate varied in relation to the absolute features of individual reference or target stimuli, as well as the relative difference between the features of reference and target stimuli. We entered firing rates in the reference and target periods on the last 100 trials in each rule block as the dependent variables and used firing rate in the fixation period as the covariate to account for trial-by-trial fluctuation in baseline firing rate. To evaluate the influence of the absolute features of individual stimuli, we entered the size and angle of the reference or target stimulus as the factor in two-way ANCOVA To evaluate the influence of relative features of reference-target pairs, we entered the relative angle

formed by the reference-target pair relative to the vertical axis, and the relative size expressed as a percentage of the reference target radius as the factors in a two-way ANCOVA. Finally, we evaluated the influence of the perceptual difficulty of the size or position discrimination, which was the absolute value of the relational parameters, capturing how near (or far) the pair of reference and target stimuli were to equality either in size or position. The range of values along the relational and difficulty dimensions were divided into 5-10 levels and the level entered as the factor in the ANCOVA.

Once subsets of neurons were identified in which firing rate varied significantly ($p < 0.05$) in relation to one of the absolute or relative feature variables, the firing rate of each neuron was aligned to its preferred feature level and then the response functions of individual neurons averaged over all cells in the population to produce the population average 3D surfaces showing modulation in average firing rate (vertical axis) as a function of size and space feature variables (Figs. 3.5.1-4 A, B). We also identified neurons in which firing rate related significantly to the interaction between size and space, and generated 3D surfaces showing averaged responses aligned to the preferred size-space combination for each neuron (Figs. 3.5.1-4 C).

In order to identify neurons in which firing rate varied in relation to the relational categories of individual target, we applied ANCOVA. We entered firing rates in the target periods on the last 100 trials in each rule block as the dependent variables and used firing rate in the fixation period as the covariate to account for trial-by-trial fluctuation in baseline firing rate. To evaluate the influence of the relational categories and of individual stimuli, we entered the relational categories of the target stimulus as the factor in one-way ANCOVA. The entered category values were either two space categories (left

or right) or two size categories (larger or smaller) for one-dimensional category identification or four compound categories (left/larger, left/smaller, right/larger, or right/smaller) for two-dimensional category identification. Once subsets of neurons encoding each category were identified spike density functions (SDFs) were constructed to visualize the population neural activity patterns (see below description for the details).

Decoding analysis

I applied time-resolved pattern classification (Klecka et al., 1980; Johnson et al., 2002; Crowe et al., 2010) to quantify and to compare the strength and timing of population signals encoding stimulus features, and stimulus category using the 'classify' MATLAB function (Figs. 3.5.9-11). Neurons included in the populations used for feature, category and response decoding were selected by ANCOVA (as described above). I performed the classification using leave-one-out cross validation. Specifically, I iteratively selected successive single 'test' trials to perform the decoding (until all trials were used) and treated all remaining trials as 'training' trials used to train the decoder at each step. Training the decoder consisted of defining the mean and covariance of population activity patterns that were observed in the subsets of trials corresponding to each level of the decoded variable in the training data. The decoding analysis compared the population activity pattern on test trials to the mean population activity patterns observed for each level of the decoded variable in the training data, and returned the most likely value for the decoded variable given the pattern of population activity on the test trial, for example 'left' when decoding spatial category, or 'larger' when decoding size

category. The decoding analysis also returned the posterior probability associated with the correct result based on the stimuli shown on that trial (e.g. the posterior probability of 'left' in the case that the spatial category of the target relative to the reference on that trial was in fact left). The proportion of correctly decoded trials and the mean posterior probability of the correct value provided measures for the strength of the population representation of the corresponding variable. I generated confusion matrices to indicate the proportion of trials that the decoder returned the correct value for the decoded variable (along the diagonal) or returned erroneous values (off-diagonal).

To capture fluctuations in population signals over time, I performed time-resolved decoding by applying the classifier to patterns of population activity measured within a sliding window passed through a sequence of single neuron firing rates measured in 50 ms bins. The sliding window was three bins wide. At each time-step, I constructed a population activity vector by taking the mean firing rates of each neuron in the three 50 ms bins within the window and concatenating the firing rates into a single vector. I then performed the decoding analysis over all trials at that time point, and computed the mean posterior probability associated with the correct value of the decoded variable. Sliding the window through the trial produced a time course of posterior probabilities associated with variation in the strength of the population signal encoding the variable of interest. To test the significance of differences in the posterior probability time series across task conditions, I performed a false-discovery rate (FDR) corrected analysis to maintain the overall Type 1 error rate at 0.05 (Fujisawa et al., 2008).

Spike density functions

Spike density functions (SDFs) were constructed to visualize the population neural activity patterns of neurons encoding categories (Figs. 3.5.5-6), rule (Fig. 3.5.7), congruency (Fig. 3.5.8 A), and response (Fig 3.5.8 B). ANCOVA was applied to identify the groups of neurons. Then single neuronal SDFs were created from spike trains of neural activity from each trial by convolution with a Gaussian kernel (kernel width: 40 ms) using the 'ksdensity' function in MATLAB. The SDFs were averaged over the subsets of neurons and the population SDFs were illustrated.

Rule neuron classification

Four types of rule neurons were defined by time trends in their firing rates (Fig. 3.5.7). I defined Type I rule neurons as those in which the baseline firing rate differed before and after the rule switch without showing additional modulation in firing rate in relation to events within the RSC trial. These neurons appeared to encode rule tonically. (Fig. 3.5.7 A). I defined type II through type IV rule neurons as those neurons in which firing rate exhibited phasic changes time locked to task events that in addition exhibited further modulation in firing rate as a function of the rule. Type II rule neurons exhibited significant increase in firing rate in the reference period (Fig. 3.5.7 B), type III in the target period (Fig. 3.5.7 C) and type IV in the decision period (Fig. 3.5.7 D), and these task-driven neural responses were larger or smaller depending on which rule was in force (Fig. 3.5.7 B-D). To identify rule neurons, I applied an ANCOVA in which the factor was the rule, the covariate was firing rate in the fixation period, and the dependent

variable was firing rate in on the last 100 trials of each rule block in the fixation (type I) reference (type II), target (type III), and the decision (type IV) periods. Once rule neurons were identified and classified, the mean SDF on preferred rule trials was normalized to the mean SDF on non-preferred rule trials (Fig. 3.5.7 E-L). Additionally, to see how rule signals changed when the rule switched, I plotted mean population firing rate of all rule neurons on preferred rule trials (within the trial period corresponding to the type of rule neuron) as a function of trial number relative to the rule switch (Fig. 3.5.7 M, N).

Regression analysis of feature encoding as a function of congruency

I applied linear regression analysis to trial-by-trial firing rates on the last 100 trials of each rule block to evaluate whether activity in a given task epoch depended on the rule-relevant and the rule-irrelevant feature dimensions by fitting the following model (Fig. 3.5.12):

$$R = \beta_0 + \beta_1 Relevant + \beta_2 Irrelevant$$

Where *R* refers to the firing rate, '*Relevant*' is the rule-relevant feature dimension (the SPACE feature dimension under the SPACE rule and the SIZE feature dimension under the SIZE rule), and '*Irrelevant*' is the rule-irrelevant feature dimension (the SPACE feature dimension under the SIZE rule and vice versa). The model was fit to firing rates on incongruent and congruent trials individually to evaluate the neural encoding of target features according to trial congruence.

To evaluate the strength and significance of the relationship between neuronal activity and the interaction of the relevant and the irrelevant feature dimensions, the following model was applied (Fig. 3.5.12).

$$R = \beta_0 + \beta_1 Relevant + \beta_2 Irrelevant + \beta_3 Relevant \cdot Irrelevant$$

Sliding-window regression analysis

I applied regression analysis to firing rates on the last 100 trials of each trial block measured within a sliding-window (100 ms window, 20 ms steps) first, to evaluate the significance of the relationship between neuronal activity and task factors (feature or category), and second, to determine if there was a significant difference in the timing of signals or in the proportion of explainable variance (PEV) across the task factors (Fig. 3.5.13). We used ANCOVA to identify pure target feature neurons (modulating activity in relation to the features but not the category of the target stimulus), and pure target category neurons (exhibiting the converse pattern of activity), and then I fit the following linear models to the activity of the corresponding neural populations:

$$R = \beta_0 + \beta_1 Feature + \varepsilon \qquad or \qquad R = \beta_0 + \beta_1 Category + \varepsilon$$

Where *R* refers to the firing rate, '*Feature'* and '*Category'* were dummy coded categorical variables, and ε is the error (residual).

Using the results of the regression analysis, we computed proportion of explainable variance (PEV) after (Olejnik and Algina, 2003; Brincat and Miller, 2016):

$$\omega^2 = \frac{SS_{between\ groups} - df\ X\ MSE}{SS_{total} + MSE}$$

and represented the activity of each neuron throughout the trial as a time series of PEV values quantifying the proportion of variance in neural activity over trials attributable to target features or categories at each time point. We then restricted populations of target feature and target category neurons to equal numbers of neurons, ranked neurons in each population according to the time to peak PEV, and compared the timing and strength of neuronal recruitment in each population. I applied the Kolmogorov–Smirnov test the cumulative distributions of the time to peak PEV to determine if there was a significant difference in the timing of target feature and category signals. Also, I applied a permutation test (1000 iterations) to determine if there was a significant difference in the mean PEV across the groups of target feature and category neurons by randomly shuffling neurons between groups and recomputing the average PEV time course in each group. Original differences in feature and category signals were considered significant if they exceeded the 95^{th} percentile of the bootstrap distribution ($p < 0.05$).

Signal transmission analysis

I applied signal transmission analysis (Crowe et al., 2013) to activity patterns in simultaneously recorded groups of neurons encoding either the feature or the category of the target stimuli to see whether information encoded by these groups of neurons was correlated over time in such a way as to imply a functional linkage between the groups of neurons. (Fig. 3.5.14). This analysis first applies pattern classification to convert patterns of activity in 50 ms time bins in nonoverlapping groups of simultaneously recorded neurons into a time series of posterior probabilities capturing fluctuations in the

strength of neural representation of specific task variables, and the determines whether the posterior probabilities are correlated at different lags within a sliding window, after taking autocorrelation in the time series into account. Detection of significant correlation in coded information implies that the two groups of neurons communicate (are functionally coupled). To carry out the transmission analysis, first I applied ANCOVA to identify nonoverlapping subsets of simultaneously recorded neurons that either encoded the features but not category of target stimuli, or encoded the category but not the features of target stimuli. (rule-relevant and rule-irrelevant separately to see whether there was rule effect in the functional linkages). Second, I applied time-resolved ensemble decoding (see above) to firing rates of feature and category neurons measured in a sequence of 50 ms time bins to produce posterior probability time series reflecting simultaneous fluctuations in the strength of neural signals encoding features or categories within trial. Third, I fit ARIMA models of order to the resulting posterior probability time series. The ARIMA procedure made the time series stationary (by differencing them twice), and then fit the coefficients of a linear model that predicted each value of the times series as the weighted sum of the 10 preceding values (lags 1-10) in order to capture the autocorrelation structure of the time series. Additionally, the linear prediction of each value was improved by including the weighted sum of errors in predicting the preceding two values. Hence the order of the models was [10, 2, 2]. We used the residuals from the ARIMA fits in subsequent stages of the transmission analysis. This isolated variation in the posterior probability time series, related to coded information, that could not be explained by their own history, and therefore would more accurately

reflect influence of extrinsic input (e.g. from another group of simultaneously recorded neurons).

At the fourth and final step in the transmission analysis, I regressed the category residual posterior probability time series onto the feature posterior probability time series at a lag of 50 ms, or the reverse, within a sliding window of 400ms advanced in 50 ms time steps. This produced transmission functions consisting of a time series of F-statistics measuring fluctuation in the strength of functional coupling between groups of neurons encoding features and categories. To determine whether differences in transmission functions were significant, I compared differences in the original data to a permutation distribution of differences computed after randomly shuffling post-ARIMA posterior probability time series between feature and category groups and repeating the transmission analysis. Original differences were considered significant if they exceeded the 95th percentile of the permutation distribution.

3.3 Results

The rule selection categorization (RSC) task we developed is a rule-dependent categorization task that requires flexible categorization of the same set of visual stimuli. One of the major differences between the RSC task and previous flexible rule-dependent categorization tasks (Goodwin et al., 2012; Mante et al., 2013; Roy et al., 2014) is that the RSC task requires autonomous rule selection using trial-and-error feedback, while other tasks often provided an external cue to indicate what categorization rule to apply. Therefore, first, the RSC task requires that monkeys internally generate a

neural representation of the categorization rule based on outcomes rather than directly extract the rule from visual input. Second, the task requires that the subject evaluate and select what rule to apply based on feedback that in some cases is ambiguous (since failures could reflect either selecting the wrong rule, incorrectly perceiving the relationship between stimuli, or selecting the wrong response). Finally, since the RSC task defined two categorization rules based on different feature dimensions of the stimuli, rules were likely to influence how features were mapped to categories. Thus, the task allows us to not only investigate the mapping of features to categories at a neural level, but how this mapping is modified by executive control.

PFC encodes absolute and relational features of stimuli

I identified PFC neurons that encoded the absolute size and position of reference and target stimuli, as well as the feature relationship of the target with respect to the reference stimulus by ANCOVA.. I then constructed 3D population average tuning surfaces in which the z-axis represented population average firing rate, and the x and y-axes represented the position and size dimensions for absolute feature encoding, or the target-reference angle and the target-reference size difference for relative feature encoding. I then aligned the activity of each neuron to its preferred feature value along the feature dimension that it encoded, and averaged the activity at each combinations of size and position in the feature surface over all the neurons in the population.

In PFC, we found subsets of neurons encoding the absolute features of the reference stimulus, (Fig. 3.5.1), the absolute features of the target stimulus, (Fig. 3.5.2) as

well as the relational features of the target with respect to reference (Fig. 3.5.3, 3.5.4). Overall, average population activity exhibited a similar pattern. Population average firing rate was modulated by variation in stimulus features along the encoded feature dimension, while firing rate did not systemically vary along the nonencoded feature dimension (Figs. 3.5.1-4 A, B). The large majority of PFC neurons encoding stimulus features (~90% of all feature-selective neurons) as main effects in the ANCOVA encoded size or space but not both dimensions. However, a substantial fraction of prefrontal neurons was influenced by the interaction between size and space (Figs. 3.5.1-4 C). Generally, stimulus features could be accurately decoded from the population activity of prefrontal feature selective neurons, although angle decoding was somewhat more robust than size decoding (Figs. 3.5.1-4 D, E).

PFC neurons encode one- and two-dimensional relational categories

The RSC task required computing the relational category of the target with respect to the reference. To visualize neural signals involved in representing relational categories, I identified neurons encoding the position (left/right) or size (larger/smaller) relational category of the target with respect to the reference using a one-way ANCOVA applied to firing rates in the target period with space and size categories as factors. To evaluate the influence of the rule on category coding, I performed separate ANCOVA on subsets of trials performed under the SPACE and SIZE rules. This analysis identified neurons that encoded rule-relevant categories, including neurons that preferred the left and right categories under the SPACE rule, (Fig. 3.5.5 A), as well neurons that preferred

the larger and smaller categories under the SIZE rule (Fig. 3.5.5. C). However, the analysis also revealed rule-irrelevant neurons with inverted rule-category preferences, including neurons that preferred left and right categories under the irrelevant SIZE rule (Fig. 3.5.5 B), as well as larger and smaller categories under the irrelevant SPACE rule (Fig. 3.5.5 D). Thus, the neural representation of categories in prefrontal cortex was not restricted to the categories that were relevant under the current rule. Rather all permutations of rule and category preference were evident. This population of neurons was selected on the basis of being influenced by category membership along one and not the other feature dimension in the ANCOVA, hence their activity exhibited modulation only in relation to a single feature dimension. (For example, left and right preferring neurons exhibited little modulation in activity in relation to the size category, Fig. 3.5.5A). Category signals in this population maintained their category preference on congruent and incongruent trial subsets and appeared little influenced by response congruency (Fig. 3.5.5, 2nd and 4th columns).

Since behavioral data in figures 2.5.6 B, C and 2.5.8 as well as the neurophysiological data in figure 3.5.1-4 C implied that in some cases monkeys used combinations of size and feature dimensions to solve the task, we investigated whether individual PFC neurons encoded two-dimensional, compound categories defined by the intersection between the size and space category dimensions. I found within PFC that subsets of neurons preferring each of the four possible compound categories existed, including neurons that preferred the left/larger, right /smaller, left/smaller, or right/larger categories (Fig. 3.5.6).

PFC neurons encode the rule

In the RSC task, monkeys were required to identify the current rule (SPACE or SIZE) using trial-and-error feedback without being given an external cue. By task design, rule signals were abstract, in the sense that they were dissociated from stimulus size, position, as well as the direction of the required motor response. We found that during RSC performance neurons encoding the rule in force were a dominant neural representation in PFC, present in the activity of more than 1,300 neurons in our sample (Fig. 3.5.7). We characterized four types of rule neurons in the PFC based on the time in the trial that they exhibited rule-selective activity (Fig. 3.5.7 A-D). Type I rule neurons encoded rules as a tonic, state variable that changed across the two halves of the block in relation to whether the SIZE or SPACE rule was in force, and according to the rule preference of each neuron (Fig. 3.5.7 A, E, I). I found separate populations of Type I rule neurons that preferred the SPACE rule (Fig. 3.5.7 E) and the SIZE rule (Fig. 3.5.7 I; population firing rate expressed as the ratio of activity on preferred to nonpreferred rule trials). These neurons did not exhibit additional modulation in firing rate time-locked to task events. In individual Type II, III and IV rule neurons, rule signals were combined with modulations in firing rate that were time-locked to task events (Fig. 3.5.7 B-D). Type II displayed a significant change in firing rate during the reference period (Fig. 3.5.7 B), type III during the target period (Fig. 3.5.7 C), and type IV during the decision period (Fig. 3.5.7 D). In each of these populations, rule-selective signals were combined with event-driven responses (Fig. 3.5.7 E-L). I found separate populations of Type II-IV rule neurons that exhibited stronger event-driven modulations in firing rate under the SPACE rule (Fig. 3.5.7 F-H) and under the SIZE rule (Fig. 3.5.7 J-L).

A critical question is how the neural representation of the rule in PFC related to the autonomous selection of which rule to apply in making behavioral decisions in the RSC task. One way to approach this question is to examine the time course of rule representation in PFC in relation to the rule-switch in each block of RSC trials. To address this question, I averaged the firing rate over the time course for that trial (to produce a single mean firing rate per neuron for that trial) on each preferred rule trial for each Type I-IV neuron, and then averaged over all neurons to obtain a population average firing rate for that trial. Plots of the population average firing rate of rule-selective neurons provides a measure for the switch in rule representation in PFC at the neural level in relationship to the switch trial in the block of RSC trials (Fig. 3.5.7 M, N). Importantly, I found several aspects of rule signals at the level of population neural activity in PFC that matched aspects of rule-switch behavior as captured by my behavioral model (Fig. 2.5.5 B, D). First, both the neural representation of the rule as reflected by population neural activity (Fig. 3.5.7 M, N; orange), and the influence of the rule on choice as reflected by the logistic regression coefficient in the behavioral model (Fig. 2.5.5 B, D; orange), reflected predominantly the pre-switch rule early in the trial block (e.g. whichever rules was imposed at the beginning of the trial block, SIZE or SPACE). Second, both the neural representation of the rule as reflected by population neural activity (Fig. 3.5.7 M, N; green), and the influence of the rule on choice at the behavioral level (Fig. 2.5.5 B, D; green), reflected predominantly the post-switch rule late in the trial block. These reversals of trends eventually led to a crossing point in the time courses measuring the influence of pre- and post-switch rules both on neural activity (Fig. 3.5.7 M, N) and behavioral choice (Fig. 2.5.5 B, D). Lastly, the behavioral crossing

points (Fig. 2.5.5 B, D) and the neural crossing points (Fig. 3.5.7 M, N) both occurred earlier in monkey 026 than in monkey 037.

PFC neurons encode response and response congruence

If monkeys use different strategies to solve the RSC task on congruent and incongruent trials and if PFC indeed integrates relevant and irrelevant information on congruent trials only, then information as to whether the categorical relationships of the target to the reference stimuli instruct congruent or incongruent responses should be encoded in neural activity after the target is displayed but before the decision as to how to respond is made. In fact, I found PFC neurons that encoded response congruence as a task variable and that this signal emerged following target onset (Fig. 3.5.8 A). Separate neural populations existed which preferred congruent (Fig. 3.5.8 A; left) and incongruent (right) trials.

Both the definitions of the relational categories and response congruency were not a priori dependent on which rule was in force in the RSC task. However, by definition the required GO / NOGO response on incongruent trials was a function of which rule was in force. I performed a one-way ANCOVA on firing rates during the target period in last 100 trials under the both rules using GO/NOGO status as a factor, and interestingly, I could identify PFC neurons that encoded the GO (Fig. 3.5.8 B, left) and the NOGO (right) status of the target-reference pair in a rule-independent manner even before the decision period (indicated by the change in the color of the fixation target) started. This

suggests that there are subsets of neurons that simply encode the future responses in the PFC before the actual decision, regardless of the current rule (Fig. 3.5.8 B).

Feature, category, and response decoding

I conducted a time-resolved population decoding analysis to examine the strength and timing of neural representations of absolute features (Fig. 3.5.9), relational features, and relational categories (Figs. 3.5.10-11) in PFC. I performed these analyses on subsets of trials when the absolute features, relational features and relational categories were either relevant to behavioral choice as defined by the current rule (Figs. 3.5.9-11; orange), or irrelevant to behavioral choice as defined by the current rule (blue), in order to determine how these neural representations were influenced by the rule in force. In addition, I contrasted relational category and GO/NOGO encoding as a function of response congruency (Fig. 3.5.11 C-E), to determine how these neural representations were influenced by response conflict.

Interestingly, the rule in the RSC task did not seem to strongly influence feature representation in PFC. Decoding time courses for the absolute features (size or space) of the reference and target stimuli were similar when the feature was relevant to the categorical decision required by the rule and when it was irrelevant. Posterior probabilities of the absolute reference feature increased when the references were presented and decreased before the targets were presented comparably on trials in which the feature of the reference stimulus encoded by neural activity was relevant or irrelevant to the rule in force (Fig. 3.5.9 A). Posterior probabilities of the absolute (Fig. 3.5.9 B) and

relational (Fig. 3.5.10 A) target feature similarly increased when the targets were presented and decreased before the decision period started comparably on trials in which the feature of the target stimulus encoded by neural activity was relevant or irrelevant to the rule in force (the apparent prolongation of absolute target feature signals on rule relevant trials in Fig. 3.5.9 B was not significant). These data suggest that the switch in neural encoding or information processing in PFC that relates to the switch in computation required by the RSC task to decide whether to compute one categorical relationship or the other between reference and target stimuli occurs somewhere downstream of encoding the features of the reference and target stimuli themselves.

In the RSC task, at some point after the features of reference and target stimuli are represented by neural activity, the relationship between the two stimuli must be computed. Our data suggests this might be done in stages. When the target stimulus is presented, it defines a relational feature parameter defined with respect to the reference stimulus, namely a difference in size or position encoded as a continuous, scalar variable. Relational feature signals code the degree to which reference and target stimuli differ along either the size or space dimensions. At a subsequent stage, the brain must divide this continuous variable into dichotomous, left/right, larger/smaller categorical representations in order to compute the response required in the task. Some insight into the how PFC neurons transform relational feature signals into relational category signals may be provided by analysis of the activity of neurons that carry both signals. By comparing ANCOVA results, I was able to differentiate pure relational feature neurons encoding only relational feature information but not relational category information (Fig. 3.5.10 B) and feature-category combined neurons encoding both relational feature and

category information (Fig. 3.5.10 C). Decoding relational feature from the populations of relational feature neurons (all together or separated into subpopulations by the presence or absence of category signals),I found that these neural populations exhibited a transient, early signal after target onset encoding the relational feature of the target with respect to the reference stimulus, as indicated by a brief increase in the posterior probability associated with the correct relational feature value defined by the reference-target pair (Figs. 3.5.10 A; 3.5.10 B, C, left). This signal decayed before the response was made. I also found that these neural populations exhibited a delayed and lasting signal after target onset that encoded the relational category of the reference-target pair (Fig. 3.5.10 B, C, right), that was particularly prominent in the subpopulation of relational feature neurons that also exhibited a significant relation to relational category by the ANCOVA (Fig. 3.5.10 C, right). The sequencing of relational feature and relational category signals evident in the decoding is consistent with a sequential computation first of the feature relationship between the reference and target stimuli, with a subsequent dichotomization of this representation at a later stage when the signal is categorized.

Also, of particular interest to the stage of processing at which rule contingency is implemented, I found that whereas relational feature signals were weakly influenced by rule relevance (Fig. 3.5.10 A, B left, C left), relational category signals were strongly influenced by rule relevance (Fig. 3.5.10 B, right, C, right). This provides evidence that executive control is implemented at the computational stage where continuous relational signals are transformed into categorical relational signals in PFC. This may reflect the fact that in the RSC task, relational feature signals did not map directly onto response strategies, whereas relational category signals did. For example, a target may be to the

near left or far left, or be a little larger or lot larger than the reference stimulus, defining a relational feature signal that had no bearing on response selection, because these parametrically different relationships mapped to the same category. However, whether the target is larger or smaller than the reference, or to the left or right of the reference stimulus, has direct bearing on the response required, because these different categories mapped to different responses. The implication of this is that it is the behavioral readout that imposes the necessity of switching computations in PFC at the neural level, and it is at the stage where abstract signals are converted into action plans that this switch is implemented.

This result was confirmed when decoding relational category from the larger population including all relational category neurons (Fig. 3.5.11 A). On rule-relevant trials, when the category encoded by neurons bore directly on response selection, category decoding reached a high level and stayed there for quite some time leading up the response and even persisting after (Fig. 3.5.11 A, orange), whereas on rule-irrelevant trials, when the category encoded by neurons did not bear on response selection, the relational category signal increased transiently, but decayed back down to baseline levels leading up to the time that the response was made. When I separated these decoding results according to rule, I found that this population dynamic in relation to the rule relevance of category signals, with relational category signals persisting when rule-relevant, decaying when not, was replicated under the SIZE and SPACE rules, suggesting it was a rule-generalized aspect of PFC function (Fig. 3.5.11 A, B). Our behavioral data suggested that on congruent trials, monkey integrated both rule-relevant and irrelevant feature dimensions to augment their behavioral choices (Fig. 2.5.8). Our

neurophysiological data also suggested that a population of neurons existed in PFC that encoded compound, two-dimensional categories, defined by the intersection of size and space dimensions, which could contribute to feature integration on congruent trials when size and space relationships synergistically instructed the same response. I investigated whether response congruency modulated category representation in the population of one-dimensional category neurons (sensitive to category distinctions along size or space dimensions but not both), and two-dimensional category neurons (sensitive to category distinctions along both size and space dimensions). I found that whereas category signals were weakly modulated by response congruence in one-dimensional category neurons (Fig. 3.5.11 C), category signals were strongly modulated by response congruency in two-dimensional category neurons (Fig. 3.5.11 D). This demonstrates that in neurons coding both size and space dimensions, response congruency mattered. However, the direction of the effect was unanticipated, namely, category signals in two-dimensional neurons were stronger on incongruent trials than congruent trials (Fig. 3.5.11 D). The reason for the direction of this effect is not clear at this stage. One possibility is that differential (correct/error) feedback for credit assignment to two-dimensional category signals is only meaningful on incongruent trials, where an error could indicate not to utilize two-dimensional categories the next time that similar conditions (e.g. stimuli) arise. On congruent trials, differential (correct/error) feedback is not informative as to whether to use one one-dimensional or two-dimensional category representations.

More PFC neurons reflect the interaction between size and space feature dimensions on incongruent trials

The above results suggested that one-dimensional and two-dimensional category representations were differentially sensitive to response congruency. I found parallel evidence in a separate regression analysis. In this analysis, I regressed firing rates in the target period onto the rule relevant and irrelevant relational feature dimensions, as well as the interaction between these factors. (Relational feature dimensions are parametric differences between the size and position of target and reference stimuli.) I performed separate regression analyses using congruent and incongruent trial subsets, and counted the numbers of neurons that were significantly influenced by the relevant and irrelevant feature dimensions, as well as their interaction. I found that the number of cells in which firing rate related significantly to the main effects of the relevant or irrelevant feature dimensions did not significantly vary as a function of response congruence (Fig. 3.5.12; ΔRel (relevant), Chi-square test, $\chi^2 = 0.17$, df = 1; p = 0.68, ΔIrr (irrelevant), $\chi^2 = 0.55$, df = 1; p = 0.46). However, the number of cells in which firing rate related significantly to the interaction between the relevant and irrelevant feature dimensions significantly increased on congruent relative to incongruent trials (Fig. 3.5.12; ΔRel · ΔIre (interaction), Chi-square test, $\chi^2 = 12.31$, df = 1; $p < 0.05$). These data provide convergent single neuron evidence that feature integration as read out by the number of neurons sensitive to the interaction between size and space feature dimensions at the population level increased when the two feature dimensions instructed the same response on congruent trials

Feature signals are earlier and influence category signals

To compare the timing and strength of feature and category signals in single neurons at the population level in PFC, I applied a sliding-window regression analysis to quantify the strength of the relationship between firing rate and the absolute feature of the target stimulus as well as the relational category of the target stimulus as a function of time within the trial. I defined the strength of the relationship between single neuron firing rate and each predictor as the proportion of explained variance (PEV) attributable to the predictor. Application of this analysis to each neuron produced a time series of PEV values capturing when within the trial that neuron carried feature or category signals. Ranking neurons in the population according to the time to peak PEV produced heat maps (Fig. 3.5.13 A) that represented population recruitment curves, or the timing with which individual neurons were activated to carry feature and category signals across the population, as indicated by the diagonal bands of warmer color indicating a stronger relationship between neural activity and feature or category. Plotting the time to peak PEV recruitment curves on the same axes revealed that feature signals significantly preceded category signals (Fig. 3.5.13 B; Kolmogorov–Smirnov test, $p < 0.05$).

I compared the strengths of feature and the category signals as a function of time by plotting the population average PEV associated with each predictor as a function of time in the trial (Fig. 3.5.13 C) and applied a permutation test to evaluate their differences. In the comparison, the population average PEV attributable to target feature was significantly larger than the PEV attributable to target category during the target period, consistent with feature signals preceding category signals during the target period when the feature-to-category transformation was likely to take place (Fig. 3.5.13 C;

permutation test, $p < 0.05$). These results provide insight into the pattern of information flow between feature and the category neurons in the PFC.

I applied a signal transmission analysis (Crowe et al., 2013) to the information about target feature and category encoded by fluctuating activity patterns of simultaneously recorded target and feature neurons (Fig. 3.5.14). This enabled me to evaluate whether and when feature signals influenced category signals over time, implying functional coupling between feature and category neurons as well as the directional transmission of information between them. It also enabled me to determine whether and when the pattern of interaction between feature and category neurons varied as a function of whether the feature information encoded by neural activity was relevant or irrelevant to category computations as a function of the rule in force. This latter question addresses how executive control is implemented by dynamically changing patterns of communication between prefrontal neurons interacting in circuits. To implement the analysis measuring functional coupling between neuronal groups, first, I identified subsets of neurons within each simultaneously recorded PFC ensemble in which activity related to target feature and not category ('feature neurons'), or to target category and not feature ('category neurons'). Then, I applied pattern classification to patterns of activity in these neuronal subsets measured in a sequence of 50 ms time bins. This converted fluctuating patterns of activity in feature and category subsets of neurons to time series of posterior probabilities capturing fluctuation in the strength with which they encoded target feature and target category. At this stage I could ask the question whether fluctuations in feature and category information were correlated over time and if so at what lag. After accounting for autocorrelation in the probability time series (using

ARIMA modeling, Methods), I regressed the time series of category posterior probabilities onto the times series of feature posterior probabilities within a sliding window at a lag of one 50 ms time bin to measure the influence of feature on category signals (Fig. 3.5.14, red). This analysis determined whether variation in feature information predicted variation in category information one time bin later. To measure the reverse interaction, namely the influence of category on feature signals, I regressed feature signals onto category signals at a lag of one 50 ms time bin (Fig. 3.5.14, blue). This analysis revealed that when the feature dimension encoded by neural activity was relevant to behavioral choice under the current rule, feature signals drove category signals relatively early in the trial, about 500 ms after target onset, before the behavioral response (Fig. 3.5.14, red). The reverse interaction, in which category signals drove feature signals, emerged approximately 400 ms later in the trial, during and after the behavioral response (Fig. 3.5.14, blue). Both patterns of interaction depended on the simultaneity of the feature and category signals, because when the simultaneity of feature and category probability time series was broken by trial shuffling (otherwise keeping the times series intact), transmission between feature and category representations was significantly weaker (horizontal red and blue bars in Fig. 3.5.14 indicate periods when transmission in the original data exceeded the 95^{th} percentile of a bootstrap distribution generated by trial shuffling the feature and category time series). Of particular relevance to the question of the circuit basis of executive control in PFC, I found that both forms of neuronal communication (feature-to-category, and category-to-feature) were markedly attenuated when the feature information encoded by neural activity patterns in PFC was irrelevant to the behavioral choice based on the rule in force (Fig. 3.5.14). Interestingly,

simultaneous correlation in target and category signals (Fig. 3.5.14, green) was much weaker than lagged correlation between these signals (Fig. 3.5.14, blue and red), suggesting that the transmission analysis detected communication between these neural populations rather than common drive from an extrinsic source. These results provide evidence that the pattern of information flow between neural populations encoding features and categories is flexible and adaptive in prefrontal cortex, such that information flow between neurons is re-routed when the rule governing the transformation of feature representations into category representations changes from the SIZE to the SPACE rule or vice versa.

3.4 Discussion

Prior studies have shown that PFC is a key brain area for cognitive control specifically as it relates to the imposition of different rules that flexibly define which categories visual stimuli belong to (Freedman et al., 2001; Antzoulatos and Miller, 2011; Goodwin et al., 2012; Crowe et al., 2013; Roy et al., 2014). However, we do not fully understand the computational strategies that the brain employs to determine category membership based on the features of visual stimuli, that is we do not yet understand the feature-to-category transformation at the behavioral level. Additionally, we do not understand the neural circuit basis of computations that mediate the feature-to-category transformation at the neural level, specifically how the flow of information from feature to category neurons is re-routed in prefrontal circuits to implement flexible categorization. This last question has significant implications for our understanding of

cognitive control. Categorization as required by the RSC task provides one example of the brain selecting what analytical algorithm to apply to the sensory input, based on trial-and-error feedback. This is a special instance of decision making, where the brain is deciding between two competing cognitive operations, rather than selecting among competing stimuli or responses. We have incomplete understanding of how reinforcement signals and learning principles apply specifically to the selection of cognitive operations in prefrontal cortex.

To address these questions, we recorded neural activity in PFC while monkeys performed the Rule Selection Categorization (RSC) task. The RSC task is a rule-dependent categorization task that requires the flexible routing of stimulus feature to stimulus category representations. To perform this task successfully, monkeys had to select whether to compute the relationship (difference) between target and reference stimuli along the size dimension (cognitive operation 1) or space feature dimension (cognitive operation 2), therefore the task required that monkeys make a decision between competing cognitive processes computing relationships along different stimulus dimensions. They had to convert a continuous feature-difference single into a categorical feature-difference signal to determine whether the target was left/right or larger/smaller in relation to the reference stimulus and then report their categorical decision by making a GO/NOGO saccadic response. The decision about which feature dimension to analyze and which categorization rule (SIZE or SPACE) to apply to determine the response was driven by trial-and-error feedback (reinforcement) rather than by an external cue. Thus, monkeys had to compute the value of two competing cognitive processes by integrating reward history over trials in which the rules were internally applied.

Here I present neurophysiological data that provides insight into how prefrontal circuits transformed feature into category representations and how this transformation was dynamically modified to implement the computational flexibility required by the task. First, prefrontal neurons encoded the absolute features (size or space) of reference and target stimuli considered individually (Figs. 3.5.1-2). Second, prefrontal neurons computed relational features, namely the difference between target and reference stimuli along either size or space dimensions (Figs. 3.5.2, 3.5.10). Third, prefrontal neurons integrated size and space information to represent compound categories reflecting the intersection of size and space dimensions (Fig. 3.5.6), to an extent that varied depending on whether the two sources of information instructed the same response (congruent trials) or competing responses (incongruent trials)(Fig. 3.5.12). This provides a correlate of the influence of response congruency on feature integration at the neural level that I documented at the behavioral level in Chapter 2 (Fig. 2.5.8). Fourth, prefrontal neurons encoded relational categories (larger/smaller, left/right) (Figs. 3.5.5-6, 3.5.10-11, 3.5.13) under task conditions that decorrelated categorical relationships from the specifics (features) of the sensory input or parameters (direction) of the motor output. This constitutes one example of a class of abstract neural representations that capture invariant features (such as relationships) that generalize over a broad range of stimulus configurations. The neural basis of that form of abstraction bears on the broader question of how neural systems acquire and deploy generalized knowledge as an operational instance of intelligent behavior. Fifth, a sequence of computations was evident whereby absolute feature signals preceded relational category signals (Figs. 3.5.10, 3.5.13). Sixth, a large population of prefrontal neurons encoded the internally selected categorization

rule and switching dynamics in rule representation at the neural level (Fig. 3.5.7), bore resemblance to switching dynamics at the behavioral level (Fig. 2.5.5). Seventh, I contrasted neural representations of size and space feature information in absolute, relational, and categorical formats across conditions in which the rule in force made each dimension either relevant or irrelevant to category membership. This enabled me to pinpoint where in the sequence of neural computations rule effects first emerged to implement flexibility in the feature-to-category transformation. Our data suggest that rule-dependence emerged at the stage where continuous relational feature signals were converted into categorical relational feature signals, leading up to the behavioral choice (Figs. 3.5.10-11). Neural representations of feature information upstream of this stage exhibited little rule dependence (Figs. 3.5.9-10). This suggests that cognitive control operates at the step where sensory representations are converted into abstract cognitive representations that ultimately control response selection. Eighth, computational flexibility based on the changing rule involved re-routing the flow of signals from different populations of feature and category neurons (as measured by lagged temporal correlation in coded information) (Fig. 3.5.14). Collectively these data relate cognitive control as a computational and behavioral phenomenon to a sequence of physiological operations in prefrontal circuits.

Characteristics of PFC neural activity recorded during RSC performance resolve ambiguities about the nature of the cognitive strategies employed by monkeys to solve the task based on analysis of behavioral data in Chapter 2. For example, based on behavioral data on congruent trials, I was not able to determine whether monkeys switched between processing relevant and irrelevant features on different trials, combined

the two feature dimensions only after they were individually computed to make a response decision, or generated new feature dimensions by integrating the two feature dimensions at a stage of perceptual processing upstream of response selection. The neural data in Chapter 3 favors the latter hypothesis – PFC neurons generated new representations by integrating space and feature dimensions, as indicated by the existence of single neurons preferring compound space-size categories as well as neurons influenced by the space-size interaction (Figs. 3.5.6, 3.5.12). I also noted in Chapter 2 that although monkeys still made errors in the later stage of the task, these errors did not induce them to switch the rule. This raises the question how monkeys determined whether or not to switch the rule following an error, since errors could reflect either of several circumstances: the right rule was applied but the relational discrimination was incorrect, the wrong rule was applied but the relational discrimination was correct, or both rule and discrimination were incorrect. It is of note that I encountered PFC neurons that specifically encoded the difficulty of the perceptual discrimination (Fig. 3.5.4) as represented by the distance of the target-reference relationship from the category boundary (Fig. 2.5.7 A, B). Neural signals encoding perceptual difficulty could provide a basis to evaluate confidence in the reliability of neural signals encoding relational categories. If this hypothesis is correct, monkeys would be more likely to switch their internal rule representation following error feedback if perceptual difficulty was low and confidence in the perception of the categorical relationship was high. By the same logic, monkeys would be less likely to switch their internal rule representation following error feedback if perceptual difficulty was high and confidence in the perception of the categorical relationship was low.

A key feature of the RSC task is that monkeys decided between two competing cognitive strategies based on trial-and-error feedback. That raises the question as to how reinforcement signals interact with cognitive signals in PFC to implement cognitive flexibility. I found that rule-relevant relational category signals in PFC were surprisingly persistent, lasting well past the response and into the feedback period at which time monkeys received visual feedback as to whether their previous response was correct or not (Fig. 3.5.11 C-E). This persistent could enable assignment of credit (reinforcement) to the cognitive strategies that monkeys utilized to make their choices. Namely, monkeys had to determine which cognitive strategy to deploy as the result of an internal deliberation that had to be driven by integrating reward history over trials in which one rule or the other was used to determine choice. Persistence of the rule-relevant relational category signal until trial feedback period might reflect this credit assignment, or namely, the neural mechanism by which reward signals influence cognitive strategy signals to modify which strategies monkeys select on future trials.

I found that PFC neurons encoded not only rule-relevant categories, as one may have predicted, but also rule-irrelevant categories (Fig. 3.5.5). On congruent trials, integration of rule-relevant and irrelevant category information may improve the reliability of the GO/NOGO decision, as the two feature dimensions instruct the same response. However, on incongruent trials, rule-irrelevant categories provide unnecessary information that may actually interfere with making the correct response decision (Cohen et al., 1990). A possible explanation for why PFC still encodes irrelevant categories on incongruent trials even when the current rule is clear can be found in previous psychophysical studies of human categorization performance (Archer, 1954; Hodge,

1959; Rabbitt, 1964; Razik, 1971). These studies demonstrated that irrelevant feature information can influence performance under conditions analogous to the RSC task in which rules were internally selected, and not under conditions in which the rule was externally cued.

Some prior studies have characterized category signals in primate prefrontal cortex under conditions in which the relationship between features and categories remained fixed (Freedman et al., 2001, 2003; Antzoulatos and Miller, 2011). Other studies have characterized category signals in primate prefrontal cortex under conditions in which the relationship between features and categories changed over trials (Cromer et al., 2010; Merchant et al., 2011; Goodwin et al., 2012; Swaminathan and Freedman, 2012; Mante et al., 2013). However, these studies have decorrelated feature and category signals to the degree we believe is achieved by the RSC task, which leverages the order of presentation of the two stimuli to invert the category relationship between them, dissociating the visual pattern of the stimuli from the category to which the stimuli are assigned. This enhanced our ability to dissociate neural signals coding features and categories in prefrontal cortex during RSC performance. That in turn made it possible for us to begin to characterize the sequence of computations involved in the feature-to-category transformation, making it possible for us to identify the point in this sequence of operations where computational flexibility was implemented. The prior study of Mante and colleagues is particularly relevant in this regard. In their task, monkeys alternatively classified a group of moving colored dots according either to the color or direction of motion in accordance with a cued instruction. They found that categorical feature representations were little influenced by the rule, which instead operated to select which

feature dimension was integrated over time by response neurons in order to select a response (Mante et al., 2013). Therefore, executive control was imposed at the level of response selection. In the RSC task, executive control was implemented upstream at the categorization stage (Fig. 3.5.11 A, B), before response selection (Fig. 3.5.11 A, B). This likely reflects differences in designs of the two tasks. In the prior study (Mante et al., 2013) the rule was externally cued and the relationship between features and categories was fixed. In the RSC task, the rule was internally selected and categories could not be derived from features. At present it is not known what differences between the two tasks led to the emergence of different dynamics in prefrontal circuits. However, the two studies together they could begin to define a task space where differences in the statistics of state-action mapping predict which neural dynamics emerge in prefrontal circuits to implement the required computational flexibility.

3.5 Figures

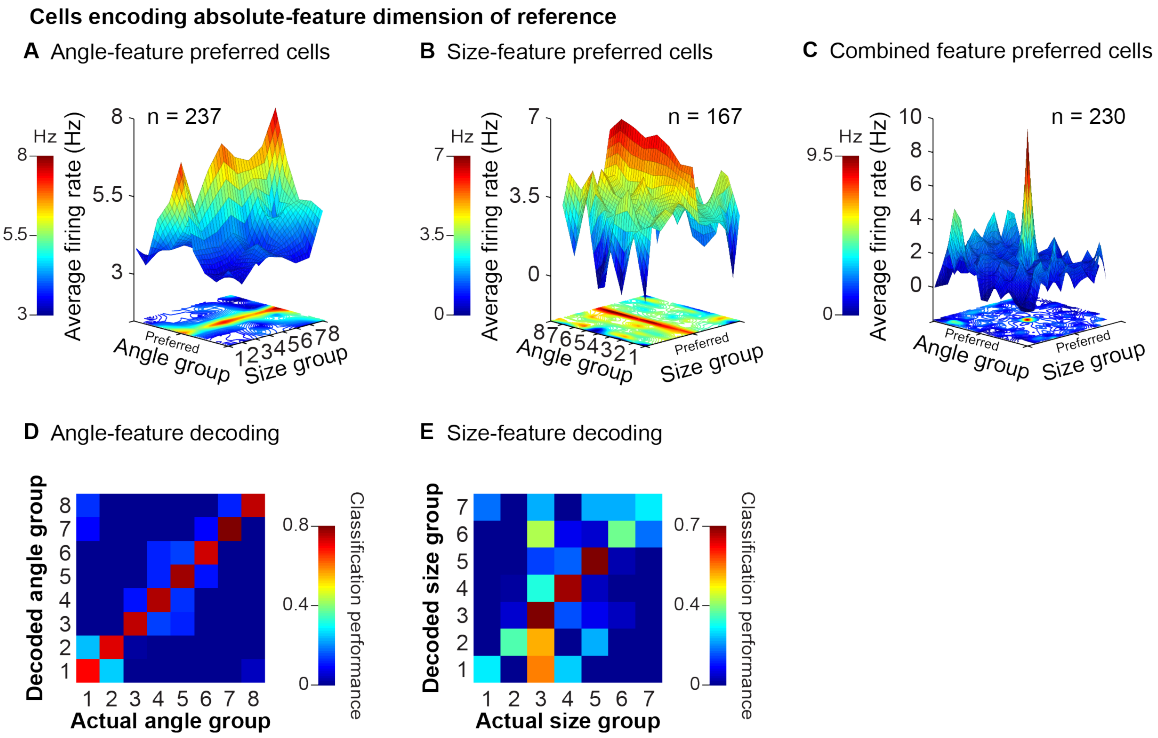

Figure 3.5.1 Population tuning to angle and size as absolute features of the reference stimulus

A-C. *Color surface indicates mean population firing rate (z-axis) over combinations of reference size and angle (x and y-axes) in neurons with activity modulated by reference (A) absolute angle (position), (B) absolute size, and (C) the interaction between angle and size. To construct the surfaces, the activity of every neuron was aligned to the preferred value along the feature dimension the neuron encoded in the ANCOVA and averaged over all feature combinations and neurons in the population.* ***D, E.*** *Confusion matrices illustrating the accuracy of angle (D) and size (E) decoding obtained from population activity patterns of angle and size coding neurons.*

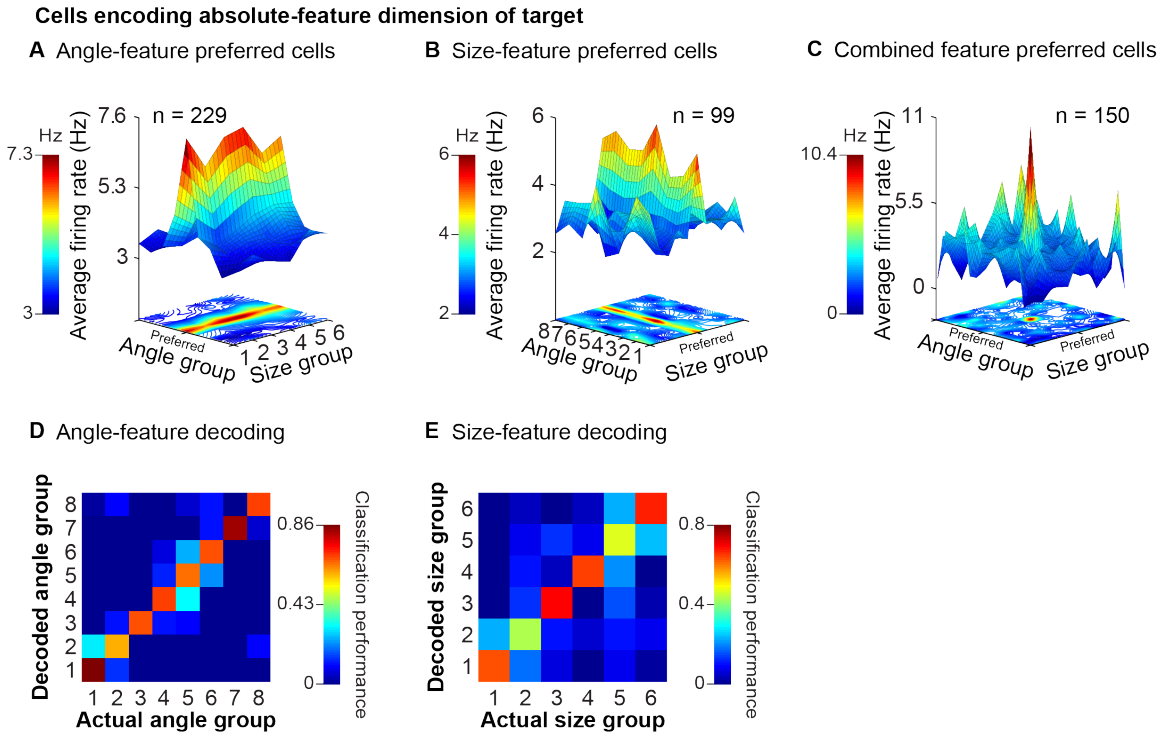

Figure 3.5.2 Population tuning to angle and size as absolute features of the target stimulus

A-E. *Conventions as in Fig. 3.5.1*

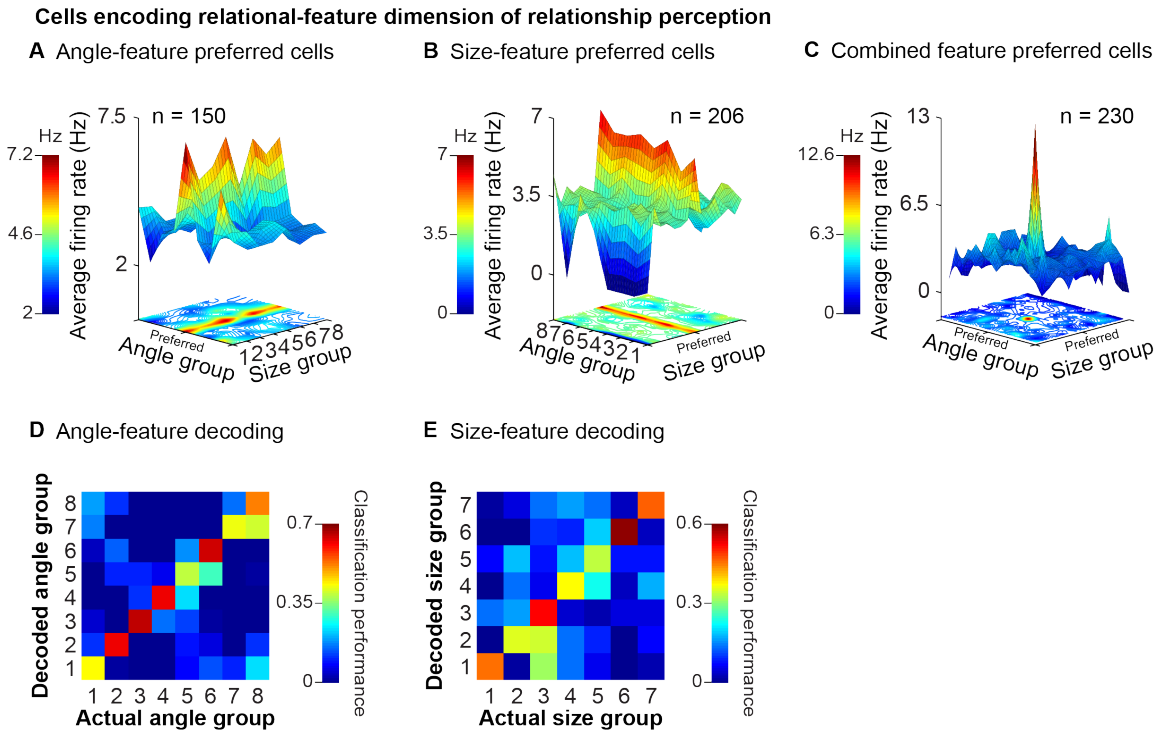

Figure 3.5.3 Population tuning to angle and size as relational features of the target with respect to the reference stimulus

Values along the size and angle group (x and y) axes represent the difference along each feature dimension between the target and the reference stimulus. Other conventions as in 3.5.1

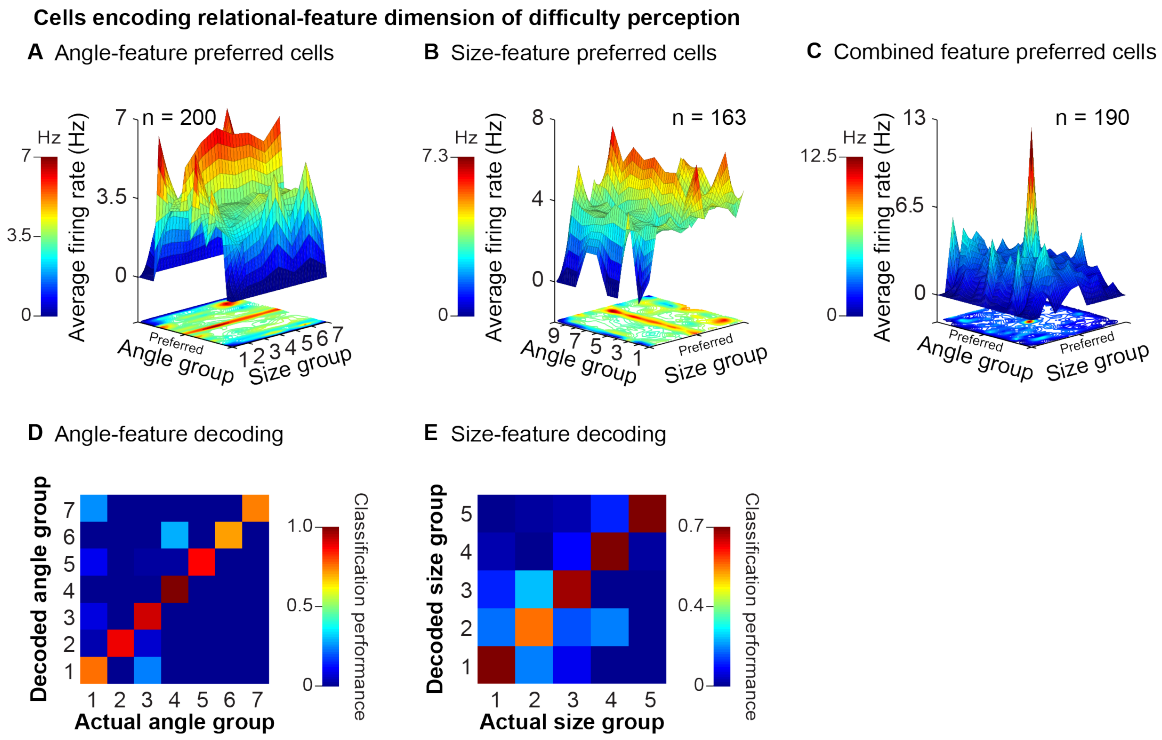

Figure 3.5.4 Population tuning to the difficulty of the relational discrimination between the target and the reference stimulus

Values along the size and angle group (x and y) axes represent the difficulty of the perceptual discrimination of the relation between target and reference stimuli. The perceptual difficulty was defined as the distance from the category boundary. For example, it was more difficult to determine whether the target was larger or smaller than the reference in the case that the two stimuli were were closer in size in comparison to the case that their sizes were farther apart (reference-target pairs of equal size defined the category boundary of the larger/smaller discrimination). Likewise, it was more difficult to determine whether the target was to the left or right of the reference in the case that the two stimuli were aligned in horizontal position in comparison to the case that the stimuli were farther apart (reference-target pairs in which the two stimuli had

the same horizontal position defined the category boundary of the left/right discrimination). The activity of all neurons was aligned to their preferred difficulty level along the feature dimension they encoded by ANCOVA and averaged over all neurons in the population. Other conventions as in 3.5.1

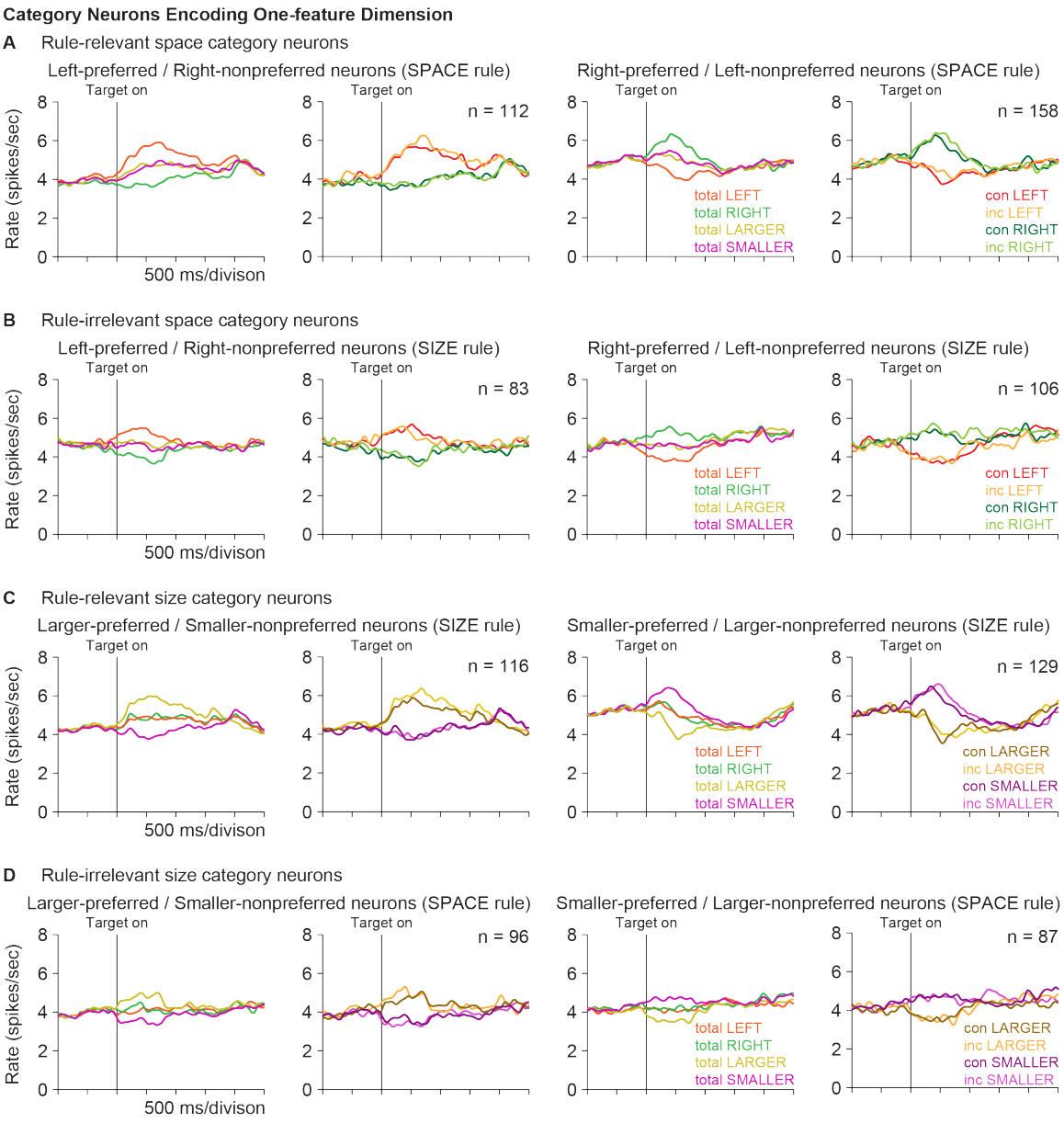

Figure 3.5.5 Average population activity of relational category neurons encoding categories along one feature dimension

Spike density functions (SDFs) illustrate average firing rate as a function of time in the trial in one dimensional relational category neurons. One-dimensional category neurons were identified by ANCOVA as having activity that differed significantly between relational categories along either the space or size dimensions but not both. SDFs in

each panel indicate population activity on subsets of trials indicated by color. ***A, B.*** *Activity of spatial category neurons preferring the left (leftmost two columns) and the right (rightmost two columns) spatial categories under the SPACE rule (A, rule-relevant) or under the SIZE rule (B, rule-irrelevant). Panels in the first and third column illustrate SDFs on trials when the target belonged to the left, right, larger and smaller categories (SDFs of different color). Activity is modulated by space and not size categories. Panels in the second and fourth columns illustrate activity on left and right trials further divided by response congruence (SDFs of different color). Congruence had little effect on activity.* ***C, D.*** *Activity of size category neurons preferring the larger (leftmost two columns) and the smaller (rightmost two columns) size categories under the SIZE rule (C, rule-relevant) or under the SPACE rule (D, rule-irrelevant). Panels in the first and third column illustrate SDFs on trials when the target belonged to the left, right, larger and smaller categories (SDFs of different color). Activity is modulated by size and not space categories. Panels in the second and fourth columns illustrate activity on larger and smaller trials further divided by response congruence (SDFs of different color). Congruence had little effect on activity.*

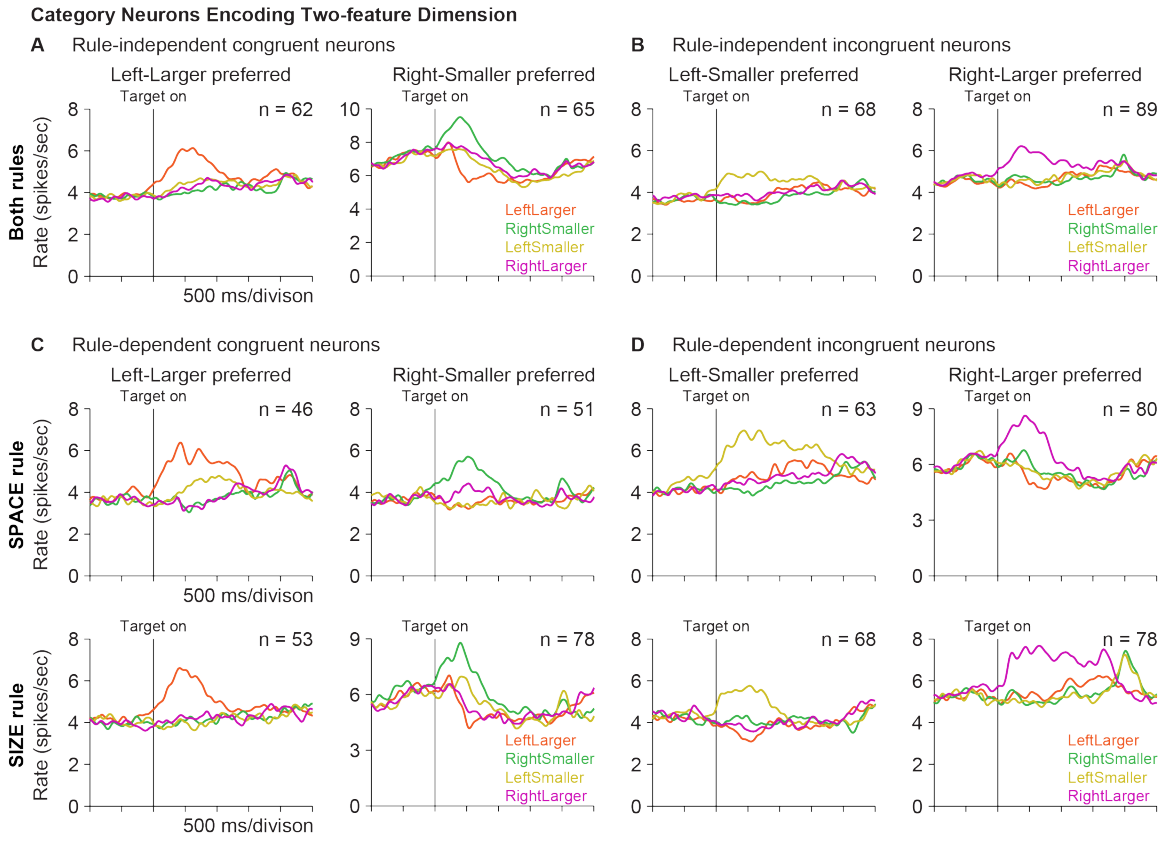

Figure 3.5.6 Average population activity of relational category neurons encoding categories along two feature dimensions

Two-dimensional category neurons were identified by ANCOVA as having activity that differed significantly between relational categories along both the space and size dimensions as main factors. These neurons preferred compound categories that were conjunctions of size and space (for example, 'Left-Larger' refers to trials in which the target was both to the left of and larger than the reference stimulus). ***A, B.*** *Activity of rule-independent two-dimensional category neurons preferring the response congruent compound categories Left-Larger, Right-Smaller (A, left and right panels), and the response incongruent categories Left-Smaller, Right-Larger (B, left and right panels).* ***C, D.*** *Activity of rule-dependent two-dimensional category neurons preferring the response*

congruent compound categories Left-Larger, Right-Smaller (A, left and right panels), and the response incongruent categories Left-Smaller, Right-Larger (B, left and right panels). Upper and lower rows illustrate population activity divided by the rule (SPACE, upper row; SIZE, lower row).

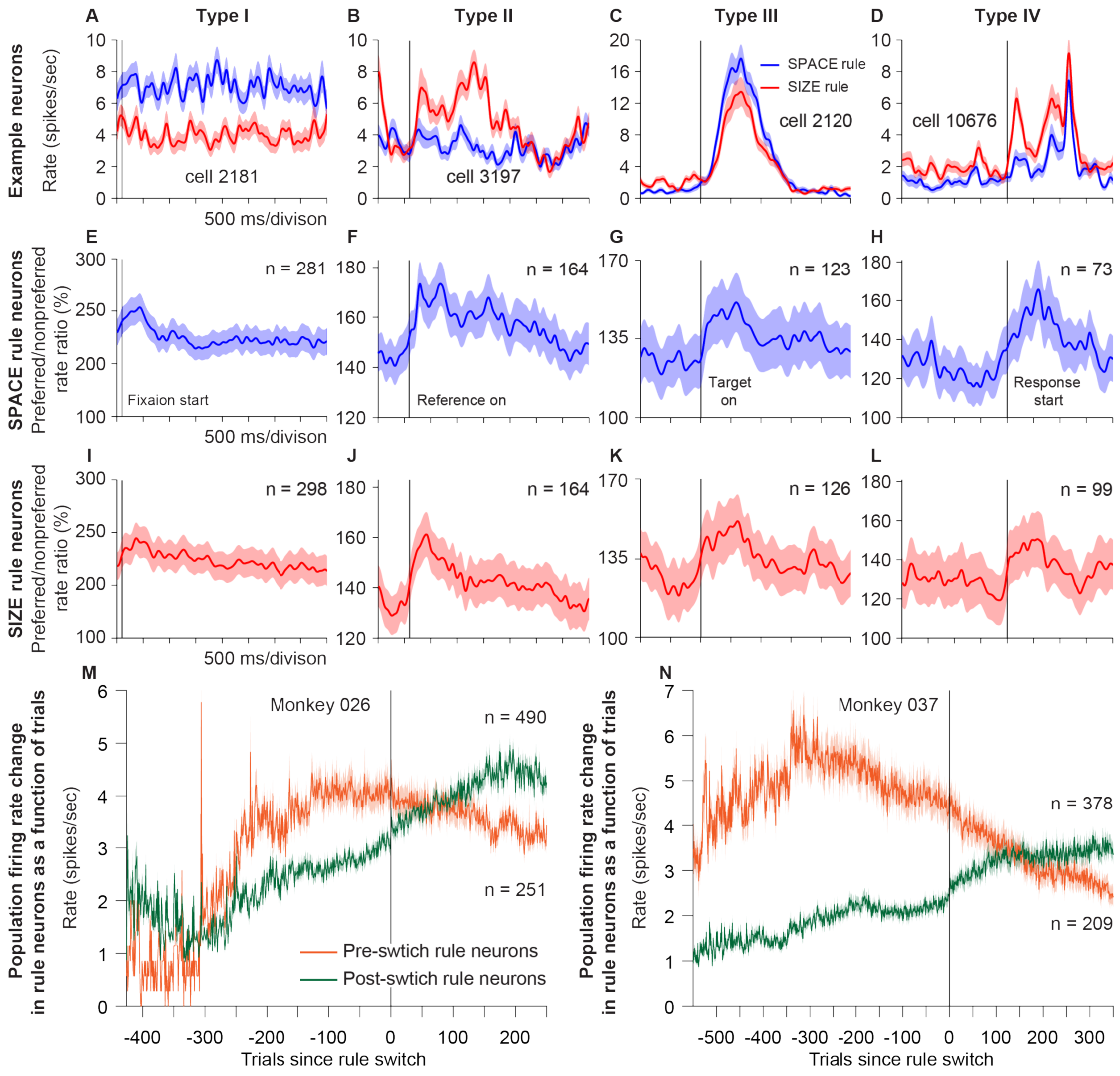

Figure 3.5.7 Average population activity of rule neurons

Rule neurons were identified by ANCOVA as having activity that differed significantly between SIZE and SPACE rules during the fixation period (Type I), the reference period (Type II), the target period (Type III) and the response period (Type IV). ***A-D.*** *Single neuron examples of each of the four types of rule neuron. SDFs illustrate neural activity under the SPACE rule (blue) and the SIZE rule (red) in each neuron. Type I rule neurons exhibited tonically elevated firing rate on preferred rule trials without additional modulation in activity time-locked to RSC task events (A). Types II-IV neurons exhibited*

elevated firing rate on preferred rule trials that rode on top of additional modulation in activity that was time locked to(B) reference onset, (C) target onset, and (D) the go signal. ***E-H.*** *Type I-IV rule neurons preferring the SPACE rule. SDFs illustrate activity on the SPACE rule normalized to activity on the SIZE rule in each neuron and then averaged over the population.* ***I-L.*** *Type I-IV rule neurons preferring the SIZE rule. SDFs illustrate activity on the SIZE rule normalized to activity on the SPACE rule in each neuron and then averaged over the population.* ***M, N.*** *Average population activity on preferred rule trials of all rule neurons as a function of trial number relative to the switch trial when the rule switched from SPACE to SIZE or vice versa within the block of trials administered during neural recording. Population activity of neurons preferring the rule in the first half of the block ('Pre-switch rule neurons') illustrated in orange, population activity of neurons preferring the rule in the second half of the block ('Post-switch rule neurons') is shown in green. The functions represent the population average firing rate of all rule neurons on preferred rule trials at the trial position relative to the switch trial indicated.*

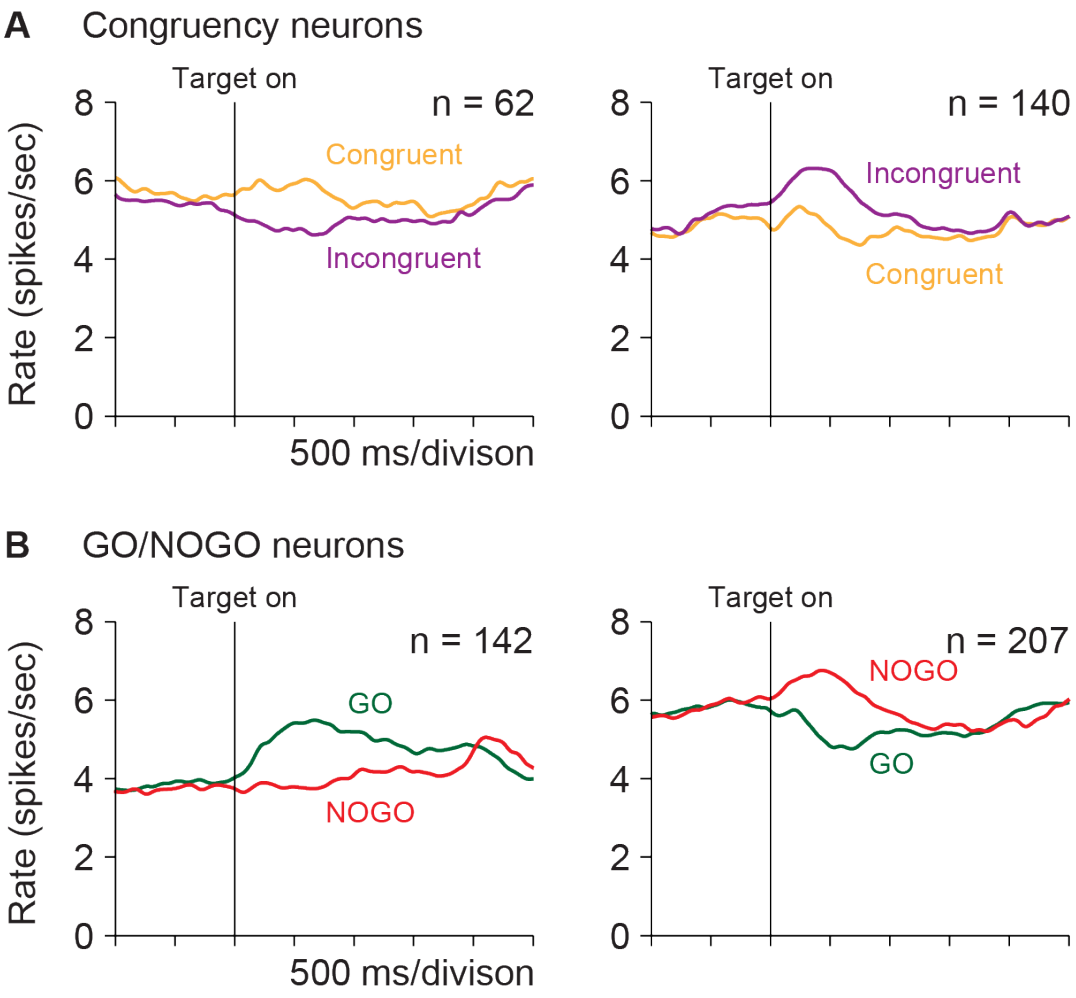

Figure 3.5.8 Average population activity of neurons encoding response congruency and the GO/NOGO response

A, B. *Spike density functions (SDFs) illustrate average firing rate as a function of time in the trial in (A) congruency and (B) response neurons. SDFs of different color indicate population activity on trials with (A) congruent (yellow) and incongruent (purple) compound categories, and (B) GO (green) and NOGO (red) responses.*

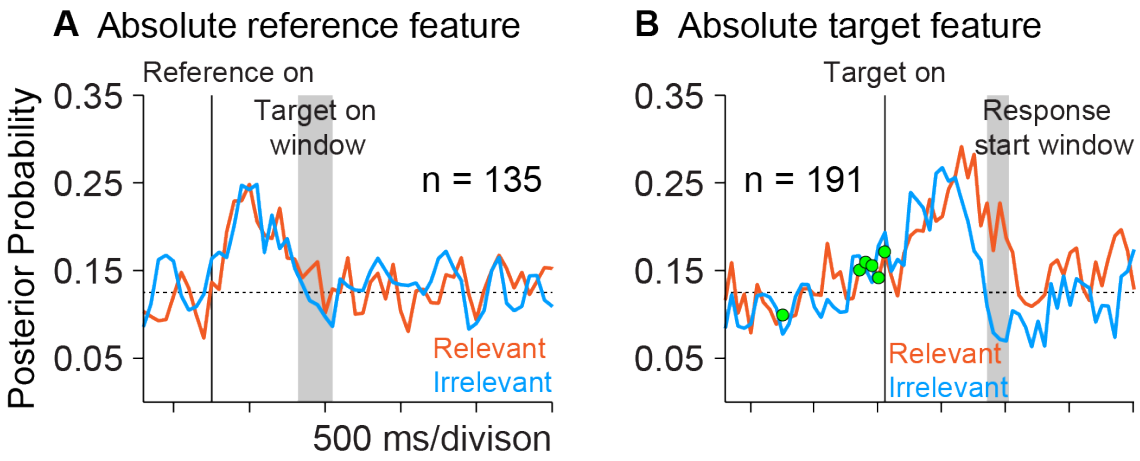

Figure 3.5.9 Time-resolved population decoding of absolute reference and target features

Functions plot the mean posterior probability obtained in a decoding analysis applied to the activity of neural populations that encoded the absolute size or space features of the reference or target stimuli in an ANCOVA. Absolute features refer to the features of individual reference and target stimuli (in contrast to relational features that capture differences in features between reference and target stimuli). Firing rates were measured in a sequence of 50 ms time bins and decoding performed on firing rate measurements within a 3-bin sliding window passed through the trial on subsets of trials in which the feature encoded by neurons was relevant (orange) or irrelevant (blue) to category membership based on the rule in force. Green dots indicate a significant difference between posterior probabilities on rule-relevant and irrelevant trials ($p < 0.05$, FDR corrected). ***A.*** *Decoding reference absolute size or space (position) features from neurons encoding the absolute features of the reference stimulus.* ***B.*** *Decoding target absolute size or space (position) features from neurons encoding the absolute features of the target stimulus.*

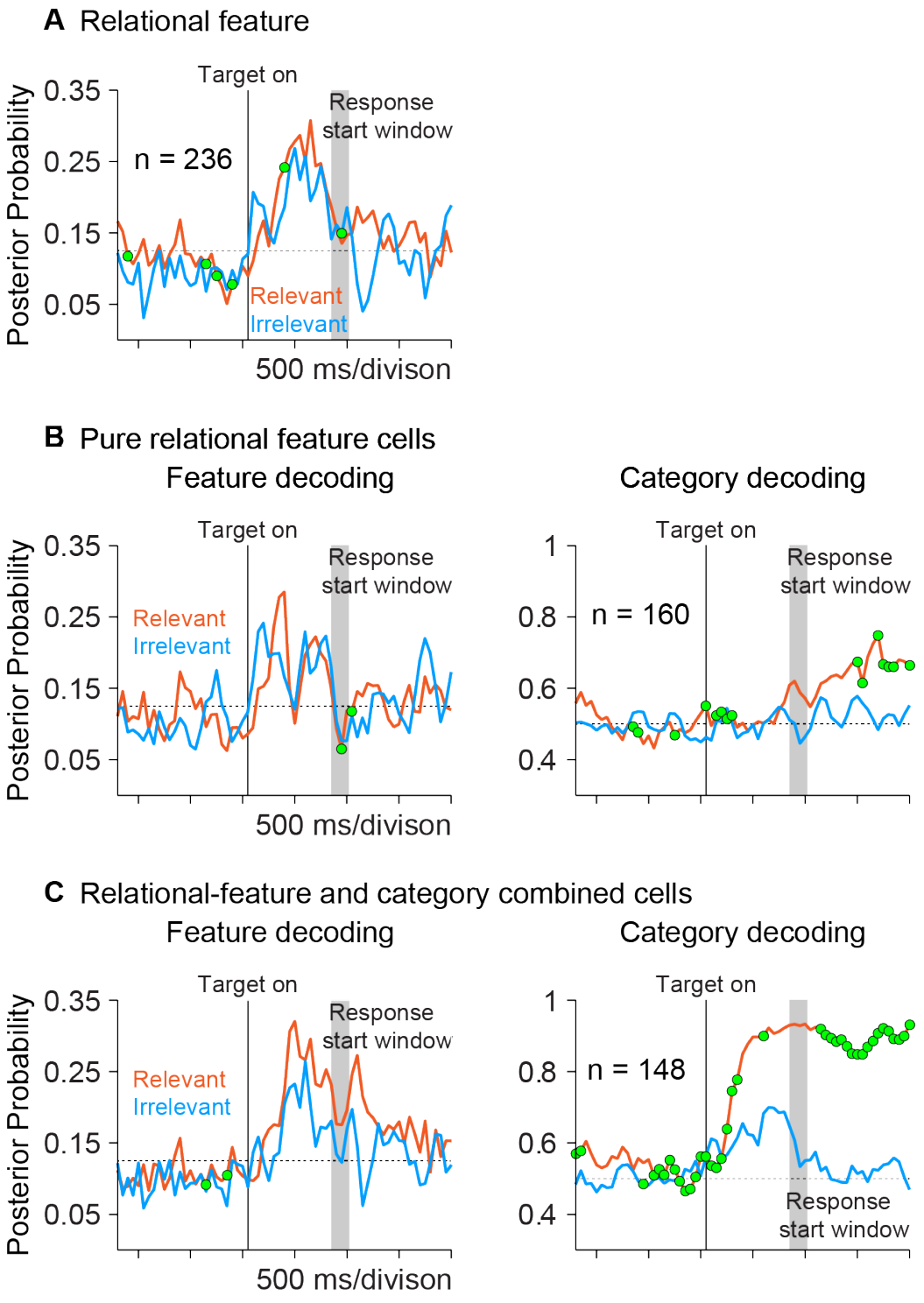

Figure 3.5.10 Time-resolved population decoding of target relational features and categories

Functions plot the mean posterior probability obtained in a decoding analysis applied to the activity of neural populations that encoded the relational size or space features or categories of the target in comparison to the reference stimulus in an ANCOVA. Relational features refer to the difference between the space (position) or size features of the target and reference stimuli. Relational categories refer to the left/right larger/smaller status of the target with respect to the reference. Decoding time courses of different color plot the mean posterior probability on subsets of trials in which the relational feature or category encoded by neurons was relevant (orange) or irrelevant (blue) to category membership based on the rule in force. Other conventions as in Fig. 3.5.9. ***A.*** *Decoding target relational features (size or space) from all neurons encoding the relational features of the target in relation to the reference stimulus.* ***B.*** *Decoding target relational features (left) or categories (right) from all neurons encoding the relational features but not the relational category of the target during the target period.* ***C.*** *Decoding target relational features (left) or categories (right) from all neurons encoding the relational features and the relational category of the target during the target period.*

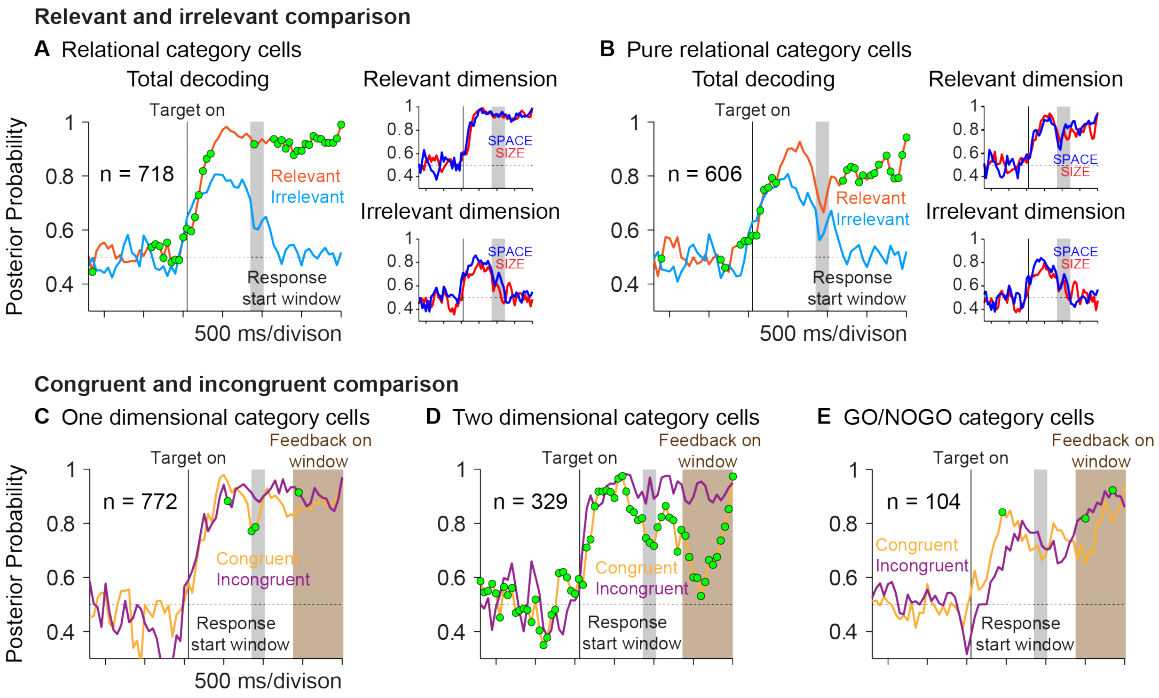

Figure 3.5.11 Time-resolved population decoding of target one and two-dimensional relational categories and the GO/NOGO response

Functions plot the mean posterior probability obtained in a decoding analysis applied to the activity of neural populations that encoded the relational category of the target in comparison to the reference stimulus, or the GO/NOGO status of the required response, in an ANCOVA. Other conventions as in Fig. 3.5.9. ***A, B.*** *Decoding target relational category from (A) all relational category neurons, including those which also encoded relational features, and (B) from pure relational category neurons that did not also encode relational features. Decoding time courses of different color in the panels on the left plot mean posterior probability on trials that the relational category was relevant (orange) or irrelevant (blue) to category membership based on the rule in force. Relevant and irrelevant-rule decoding time courses are further divided into trials under the SPACE rule (blue) and the SIZE rule (red) in the insets.* ***C, D.*** *Decoding relational*

category from (C) one-dimensional category neurons (encoding size but not space dimensions or vice versa by ANCOVA), and (D) two-dimensional category neurons (encoding both size and space dimensions). Decoding time courses of different color plot mean posterior probability on trials that the compound category of the target was response congruent (yellow) or incongruent (purple). ***E.*** *Decoding trial GO/NOGO status from neurons encoding GO/NOGO status in the ANCOVA trials that the compound category of the target was response congruent (yellow) or incongruent (purple).*

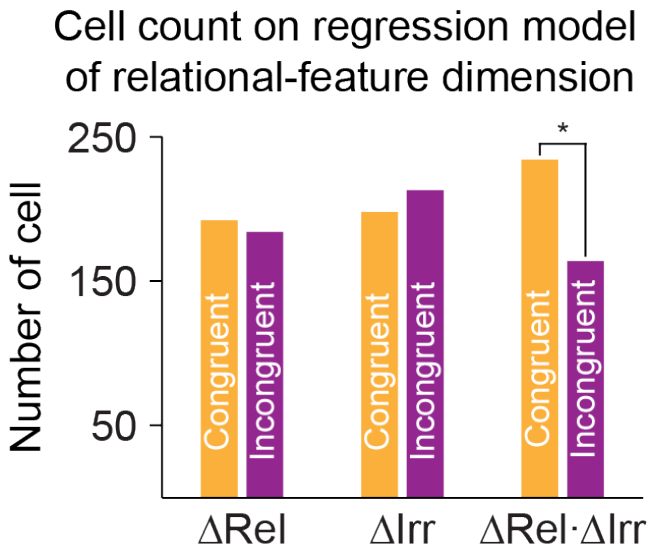

Figure 3.5.12 Numbers of neurons with activity relating to relational features as a function of rule-relevance and response congruence

The number of neurons in which firing rate related significantly to target relational category along the rule-relevant dimension (ΔRel) the rule-irrelevant dimension (ΔIrr), or the interaction between rule-relevant and rule-irrelevant dimensions (ΔRel· ΔIrr) in a linear regression analysis applied to firing rate in the target period ($p < 0.05$). Separate regression analyses were conducted using trials when the target compound category was response congruent (yellow) and incongruent (purple). (Counts for main effects obtained in a model without the interaction term.)

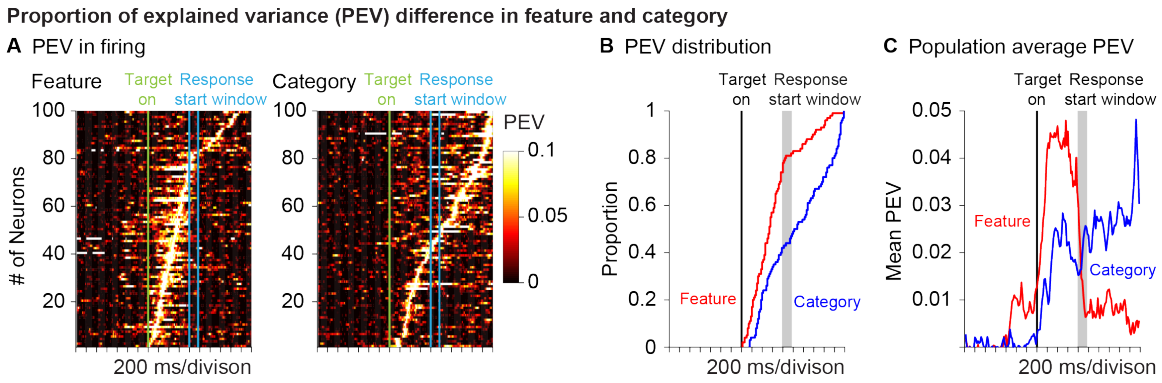

Figure 3.5.13 Single neuron representation of target feature and category in a sliding-window regression analysis

Firing rates of individual neurons within a sliding window (100 ms window, 20 ms steps) were regressed onto target relational feature and category. The results of the regression are expressed as proportion of explainable variance (PEV) in firing rate over trials attributable to target relational feature and category predictors. ***A.*** *Heat maps plot the PEV as a function in time in the trial attributable to target relational feature (left) and category (right). Warmer colors indicate higher PEV values. Populations of feature and category neurons were equated to 100 neurons so that recruitment curves, illustrated by the diagonal bands of warmer color, could be directly compared.* ***B.*** *Functions plot time to peak PEV for target feature (red) and category (blue) regressors derived from data in panel A. Neurons were recruited to encode target relational features significantly earlier than target relational categories (Kolmogorov-Smirnov test, $p < 0.05$).* ***C.*** *Functions plot the population average PEV for target feature (red) and category (blue) regressors derived from data in panel A. During the target period, target PEV was significantly greater than category PEV (permutation test, 1000 iterations, $p < 0.05$)*

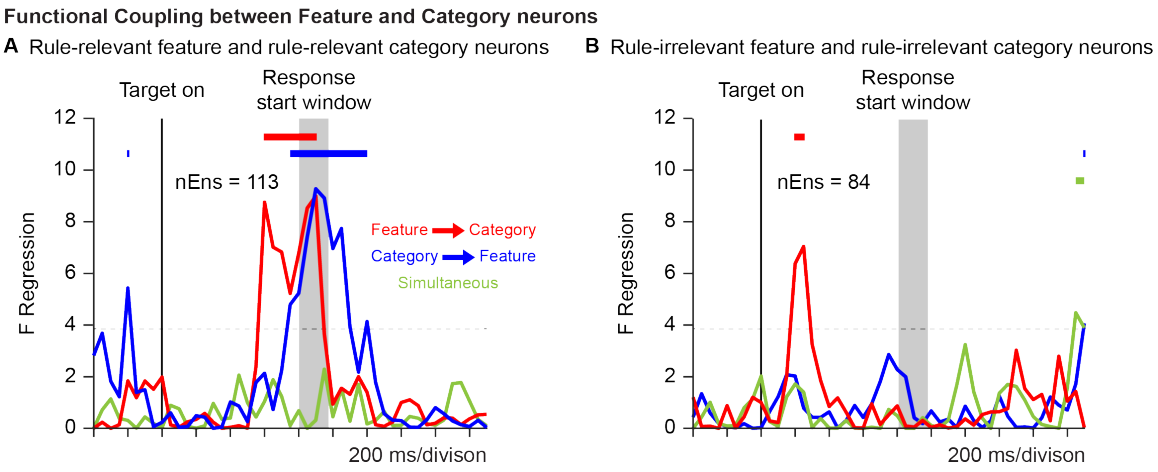

Figure 3.5.14 Functional coupling between feature and category signals evaluated by signal transmission analysis

Signal transmission analysis was applied to measure lagged temporal correlation in information about target absolute features and relational categories encoded by subsets of simultaneously recorded PFC neurons. This analysis proceeded in steps: (1) identification by ANCOVA of subsets of absolute feature and pure category neurons within simultaneously recorded ensembles, (2) decoding target absolute feature and relational category from the firing rates of these neural subsets in 50 ms time bins, (3) pre-whitening of the resulting posterior probability time series by ARIMA modeling, (4) regressing residual (post-ARIMA) posterior probabilities associated with target category onto posterior probabilities associated with target features (or vice versa) within a 400 ms sliding window. The regression was performed either comparing simultaneous time bins (green), with feature data leading category data by one 50 ms time bin (Feature to Category, red), or with category data leading feature data by one 50 ms time bin (Category to Feature, blue). Significant functional coupling at each lag was identified at time points (horizontal bars of corresponding color) for which the original F-statistic

exceeded the 95th percentile of a bootstrap distribution of F-statistics obtained after randomly shuffling trials to break the simultaneity of neural activity in the two groups and repeating the regression analysis. Numbers of ensembles ('nEns') contributing to each transmission analysis are shown. ***A, B.*** *Feature-to-category transmission (red; feature signals leading category signals) and category-to-feature transmission (blue; category signals leading feature signals) when the feature and category dimensions were (A) relevant, or (B) irrelevant to the rule in force.*

4

Conclusion

My dissertation provides both behavioral and neural data that advances our understanding of how cognitive strategies are both developed and dynamically applied to make decisions in a changing environment. I relate computational flexibility in changing environments to physiological dynamics in prefrontal circuits. My study is the first to investigate the neural mechanisms of cognitive control in a rule-based categorization task in which the rule must be determined by an internal decision process based on trial-and-error feedback rather than being explicitly instructed by an external cue. The rule selection categorization (RSC) task is also the first to fully decorrelate stimulus features and stimulus categories by basing categories on relationships between stimuli rather than features intrinsic to any one of the considered individually. That allowed me to track the transformation of feature representations into category representations in prefrontal cortex and determine the stage of processing where cognitive control (as evidenced by rule-dependence in neural signals) first emerged. My task is also one of the first to

parametrically vary the perceptual difficulty of the categorical discrimination along two feature dimensions simultaneously, making it possible for me to determine which sources of sensory input were used to compute category membership. These aspects of RSC task design required monkeys to generate and dynamically adjust abstract cognitive strategies and enabled us to evaluate the flexibility of neural computations in prefrontal cortex to determine how sensory information becomes a category or action at the neuronal and behavioral levels.

In chapter 2, I discussed several major findings deriving from the behavioral data collected during performance of the RSC task. First, monkeys were proven able to learn abstract categorization rules based on relationships between stimuli and to select the rule autonomously without explicit indication. Second, by using trial-and-error feedback, monkeys were able judge what sensory information was relevant or irrelevant based on the current rule. Third, monkeys developed and applied their own strategies of how to integrate or filter out relevant and irrelevant sensory information flexibly depending on whether they instructed the same or conflicting responses. This has implications for prominence of the role of behavioral output in sculpting or constraining cognitive operations in prefrontal cortex, namely that cognition advances in the service of action. In my prior discussion, I draw analogies to the Stroop test (Stroop, 1936; Cohen et al., 1990) to frame interpretation of behavioral and neural correlates of congruency in my experiments in the context of response conflict. In addition, I note similarities between my results and prior psychophysical studies of human categorization performance (Archer, 1954; Hodge, 1959) reporting that choosing a rule in accordance either with an

external cue or an internal decision changes how the brain integrates sensory information to make category judgments.

In chapter 3, I discussed the characterization of feature, rule and category signals during RSC performance and neural correlates of the rule-dependent transformation of feature into category signals in the PFC. This work revealed the following. First, PFC neurons encode absolute and relational features. Second, both absolute and relative feature signals exhibit little rule-dependence, suggesting that cognitive control is imposed at a stage of processing subsequent to feature representation. Third, I report converging behavioral and neural evidence that monkeys do not restrict their analysis of the visual input to relevant feature dimensions as logically required the current rule. Rather, they extract information from both rule-relevant and irrelevant feature dimensions when the two instruct the same conditional response. This shows that the brain can optimize data mining algorithms that operate on the sensory input to support ongoing cognitive operations in a way that is not intuitively obvious a priori. (For example, our expectation was that monkeys would switch processing between space and size feature dimensions entirely when they switched the rule, rather than integrating information from these dimensions when they instructed the same action.) Fourth, under conditions that monkeys internally select rules, rules are robustly encoded by a dominant population of prefrontal neurons. We plan to further investigate how feedback regarding trial success and failure influences rule representations to adapt cognitive strategies over trials. Fifth, prefrontal neurons encode relational categories, and based on the features of the RSC task, I was able to show that these signals are decorrelated from sensory and motor signals in the brain. Emergence of abstract neural representations of this type in

prefrontal cortex is of some interest to the question of how prefrontal cortex supports intelligent behavior by acquiring and applying abstract representations (such as representations of relationships or categories) which generalize over a broad range of specific sensory conditions, do not map to specific motor responses, are therefore decoupled from the specifics of sensorimotor control, but enable rule-based action selection nonetheless. I show the existence of utilization of these signals. A deeper question is how these representations are acquired, and what aspects of the statistics of the environment favor their emergence. Sixth, I provide evidence that information processing operations are staggered in prefrontal cortex, with neurons sequentially encoding absolute features, relative features and relative categories. Seventh, I show relational category signals are the first to exhibit robust rule-dependence, suggesting that executive control is implemented at the stage of computation at which feature signals are transformed into category signals. Eight, I provide evidence that whereas signals coding categories that are relevant and irrelevant to the rule rise together after target onset, the rule-relevant signals persist, whereas the rule-irrelevant signals decay. One consequence of this is that rule-relevant category signals persist after the decision and response are made into the feedback period. These long-lasting category signals could enable assignment of credit (reinforcement) to the recently used cognitive strategy to drive rule learning and adaptation of strategy on future trials. I plan future analyses of this dataset to investigate whether neurons that encode feedback signals transmit information to neurons that encode relational categories and whether the strength of this interaction predicts changes in both the strength of category representation and category choice probability on subsequent trials. Ninth, I provide evidence that transmission of feature-

to-category signals precedes transmission in the opposite direction (category-to-feature), and that the strength of this transmission is strongly influenced by the relevance of the transmitted information to the categorical judgement required under the current rule. The implication here is that changing cognitive strategies can re-route the flow of information between neurons in prefrontal circuits leading to altered behavioral responses to sensory input.

Below, I consider several limitations of this study along with prospective analyses that could provide solutions.

1. Although rule-relevant category signals were not correlated with the direction of the saccadic response, they were partially confounded with the GO/NOGO status of the trial because, in the RSC task, the left and larger categories always required GO responses, while the right and smaller categories always require NOGO responses, and I did not switch the required GO/NOGO for each category during neural recording (in part because the RSC task was at the complexity limit of tasks I could effectively train monkeys to perform, and switching the response contingency would have made the task significantly more challenging). GO/NOGO status is itself an abstract parameter (specifying whether or not to make a saccade, without specifying parameters typically associated with saccade plans such as saccade direction and amplitude), however future work will be required to cleanly parse these signals

2. The neural representation of rule was confounded with elapsed time (or trial number) in the RSC task, because I only was able to train monkeys to switch the rule once per block of trials administered during neural recording. The reason for that is that

the learning dynamics were slow, taking 100-200 trials to asymptote. The protracted time course of rule switching in the RSC task could reflect the overall complexity of the task, the ambiguity intrinsic to error feedback (namely whether the wrong rule was applied or incorrect relationship perceived), or the fact that half of the (response congruent) trials did not provide informative feedback about the correctness of the rule. Nonetheless, any slowly changing parameter (quality of the neural isolation, attention or motivation) may have contributed to the significance of rule signals (particularly in Type I rule neurons). However, rule signals rode on top of modulations in firing rate that were time-locked to stimulus events (in Type II-IV rule neurons), and the rule reliably modulated the persistence (rather than magnitude) of category signals in PFC, so it seems unlikely that rule signals could be attributed entirely to slowly changing variables such as loss of isolation or change in motivation. That said, future studies that investigate repeated switching back and forth between alternative rules would provide a stronger argument.

3. The flexible functional mapping of feature to category signals needs further analysis. In chapter 3, we demonstrated that the rule modulated PFC circuit dynamics, such that feature-to-category transmission was stronger when the rule made feature information relevant to the category discrimination in comparison to when the rule made that information irrelevant. However, the details of flexible information coupling in PFC circuits are still unknown. We do not know for example how information is differently routed through PFC circuits as a function of response congruency, although a difference in dynamics is expected based on the finding that PFC neurons differentially integrate feature information based on response congruency. Thus, it is unclear how feature

signals along different dimensions communicate with category signals differently as a function of the rule or how these signals map to responses. In chapters 2 and 3, I mentioned that irrelevant feature or category signals could either enhance the reliability of action decisions on congruent trials or hinder them on incongruent trials. However, the neural mechanisms of synergy or competition between neural populations encoding different forms of information was not investigated in this work. In addition, although hypotheses about how category information influences the GO/NOGO decision were suggested in chapter 3, how neural representations of relational categories are transformed into response decisions was not investigated.

4. I did not investigate what, or how, information is used to update rules and compel monkeys to switch their cognitive behavior in response to trial-and-error feedback. This is one of the central questions posed by the RSC task design and it will require further analysis to investigate. I suggest above that the persistence of rule-relevant category signals past the decision and response into the subsequent feedback period of the trial might provide a clue as to how credit is assigned to cognitive strategies that were used in the recent past, but testing this will require further work.

I plan to further investigate how monkeys generalize the concepts of relative left, right, larger, and smaller. An analog reference point model has been proposed that assumes that there are internal reference points distributed at the two ends of feature dimensions used to define categories, and the stimulus's mental distance from the reference point is calculated to make a decision about what category the stimulus belongs to (Holyoak, 1978). If true, reaction time in the RSC task should be faster and the accuracy of categorical judgments higher when a stimulus is closer to the reference point.

That effect has been documented in a categorical spatial judgment task in which monkeys judged visual stimuli to be 'high' or 'low' (Fortes et al., 2004). If this model explains category performance in the RSC task, I would predict that monkeys would be faster to report and more accurately judge larger-than categorical relationships between target and reference stimuli when the absolute size of the target was larger compared to when it was smaller (a type of congruity effect), apart from the relationship to the reference. If monkeys generalized relative categories without using internal reference points of this type, I would not expect to observe this form of congruity effect.

In addition, the RSC task could provide a good model to investigate the roles of other brain area such as anterior cingulate cortex (ACC) in conflict monitoring and resolution at the neural level. In human studies, it has been reported that the ACC is active under conditions of high response conflict as imposed by the Stroop task and the GO/NOGO response conflict task (Kawashima et al., 1996; Peterson et al., 1999; Kiehl et al., 2000). Neural recording studies in monkey have failed to find conflict signals in ACC under conditions of high conflict (Nakamura et al., 2005; Hayden et al., 2011). It has been proposed that disagreement between human neuroimaging and monkey neural recording studies of conflict processing relate to the nature of the tasks employed, specifically that tasks used to study conflict in monkeys are relatively constrained and lack the quality of naturalistic decision making in humans (Widge et al., 2019). Often in human decision making, the correct response is not explicitly dictated by external cues, and conflict emerges because multiple cognitive strategies that could apply in any environmental circumstance compete for behavioral control, and must be resolved by an internal process. Many of these characteristics are captured by the RSC task, since the

rule that monkeys use is not instructed by an external cue and they must internally resolve conflict between competing cognitive strategies. Therefore, further investigation of competition between neural populations encoding conflicting strategies or responses in the RSC task may provide a useful model of the neural mechanisms of conflict processing in human decision making.

BIBLIOGRAPHY

Antzoulatos EG, Miller EK (2011) Differences between neural activity in prefrontal cortex and striatum during learning of novel abstract categories. Neuron 71:243–249.

Archer EJ (1954) Identification of visual patterns as a function of information load. J Exp Psychol 48:313–317.

Asaad WF, Rainer G, Miller EK (2000) Task-specific neural activity in the primate prefrontal cortex. J Neurophysiol 84:451–459.

Barbey AK, Koenigs M, Grafman J (2013) Dorsolateral prefrontal contributions to human working memory. Cortex 49:1195–1205.

Blackman RK, Crowe DA, DeNicola AL, Sakellaridi S, MacDonald AW 3rd, Chafee MV (2016) Monkey Prefrontal Neurons Reflect Logical Operations for Cognitive Control in a Variant of the AX Continuous Performance Task (AX-CPT). J Neurosci 36:4067–4079.

Brincat SL, Miller EK (2016) Prefrontal Cortex Networks Shift from External to Internal Modes during Learning. J Neurosci 36:9739–9754.

Buckley MJ, Mansouri FA, Hoda H, Mahboubi M, Browning PGF, Kwok SC, Phillips A, Tanaka K (2009) Dissociable components of rule-guided behavior depend on distinct medial and prefrontal regions. Science 325:52–58.

Burgess PW (2000) Strategy application disorder: the role of the frontal lobes in human multitasking. Psychol Res 63:279–288.

Cavada C, Goldman-Rakic PS (1989a) Posterior parietal cortex in rhesus monkey: II. Evidence for segregated corticocortical networks linking sensory and limbic areas with the frontal lobe. J Comp Neurol 287:422–445.

Cavada C, Goldman-Rakic PS (1989b) Posterior parietal cortex in rhesus monkey: I. Parcellation of areas based on distinctive limbic and sensory corticocortical connections. J Comp Neurol 287:393–421.

Christoff K, Keramatian K, Gordon AM, Smith R, Mädler B (2009) Prefrontal organization of cognitive control according to levels of abstraction. Brain Res 1286:94–105.

Cohen JD, Dunbar K, McClelland JL (1990) On the control of automatic processes: a parallel distributed processing account of the Stroop effect. Psychol Rev 97:332–361.

Cromer JA, Roy JE, Miller EK (2010) Representation of multiple, independent categories in the primate prefrontal cortex. Neuron 66:796–807.

Crowe DA, Averbeck BB, Chafee MV (2010) Rapid sequences of population activity patterns dynamically encode task-critical spatial information in parietal cortex. J Neurosci 30:11640–11653.

Crowe DA, Goodwin SJ, Blackman RK, Sakellaridi S, Sponheim SR, MacDonald AW 3rd, Chafee MV (2013) Prefrontal neurons transmit signals to parietal neurons that reflect executive control of cognition. Nat Neurosci 16:1484–1491.

Croxson PL, Johansen-Berg H, Behrens TEJ, Robson MD, Pinsk MA, Gross CG, Richter W, Richter MC, Kastner S, Rushworth MFS (2005) Quantitative investigation of connections of the prefrontal cortex in the human and macaque using probabilistic diffusion tractography. J Neurosci 25:8854–8866.

DeNicola AL, Park M-Y, Crowe DA, MacDonald AW 3rd, Chafee MV (2020) Differential roles of MD thalamus and prefrontal cortex in decision making and state representation in a cognitive control task measuring deficits in schizophrenia. J Neurosci Available at: http://dx.doi.org/10.1523/JNEUROSCI.1703-19.2020.

D'Esposito M, Postle BR, Ballard D, Lease J (1999) Maintenance versus manipulation of information held in working memory: an event-related fMRI study. Brain Cogn 41:66–86.

D'Esposito M, Postle BR, Rypma B (2000) Prefrontal cortical contributions to working memory: evidence from event-related fMRI studies. Exp Brain Res 133:3–11.

Donahue CH, Lee D (2015) Dynamic routing of task-relevant signals for decision making in dorsolateral prefrontal cortex. Nat Neurosci 18:295–301.

Duncan J, Parr A, Woolgar A, Thompson R, Bright P, Cox S, Bishop S, Nimmo-Smith I (2008) Goal neglect and Spearman's g: competing parts of a complex task. J Exp Psychol Gen 137:131–148.

Ferrier D (1886) The Functions of the Brain. Smith, Elder.

Fitzgerald JK, Swaminathan SK, Freedman DJ (2012) Visual categorization and the parietal cortex. Front Integr Neurosci 6:18.

Fortes AF, Merchant H, Georgopoulos AP (2004) Comparative and categorical spatial judgments in the monkey: "high" and "low." Anim Cogn 7:101–108.

Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2001) Categorical representation of visual stimuli in the primate prefrontal cortex. Science 291:312–316.

Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2003) A comparison of primate prefrontal and inferior temporal cortices during visual categorization. J Neurosci 23:5235–5246.

Fujisawa S, Amarasingham A, Harrison MT, Buzsáki G (2008) Behavior-dependent short-term assembly dynamics in the medial prefrontal cortex. Nat Neurosci 11:823–833.

Fuster J (2015) The Prefrontal Cortex. Academic Press.

Goldman-Rakic P (2011) Circuitry of Primate Prefrontal Cortex and Regulation of Behavior by Representational Memory.

Goodwin SJ, Blackman RK, Sakellaridi S, Chafee MV (2012) Executive control over cognition: stronger and earlier rule-based modulation of spatial category signals in prefrontal cortex relative to parietal cortex. J Neurosci 32:3499–3515.

Harlow JM (1868) Recovery from the Passage of an Iron Bar through the Head. Publications of the Massachusetts Medical Society 2:327–347.

Hayden BY, Heilbronner SR, Pearson JM, Platt ML (2011) Surprise signals in anterior cingulate cortex: neuronal encoding of unsigned reward prediction errors driving adjustment in behavior. J Neurosci 31:4178–4187.

Hebb DO (1939) Intelligence in Man after Large Removals of Cerebral Tissue: Report of Four Left Frontal Lobe Cases. J Gen Psychol 21:73–87.

Hodge MH (1959) The influence of irrelevant information upon complex visual discrimination. J Exp Psychol 57:1–5.

Holyoak KJ (1978) Comparative judgments with numerical reference points. Cogn Psychol 10:203–243.

Ilinsky IA, Jouandet ML, Goldman-Rakic PS (1985) Organization of the nigrothalamocortical system in the rhesus monkey. J Comp Neurol 236:315–330.

Jeon H-A, Anwander A, Friederici AD (2014) Functional network mirrored in the prefrontal cortex, caudate nucleus, and thalamus: high-resolution functional imaging and structural connectivity. J Neurosci 34:9202–9212.

Johnson RA, Wichern DW, Others (2002) Applied multivariate statistical analysis. Prentice hall Upper Saddle River, NJ.

Kawashima R, Satoh K, Itoh H, Ono S, Furumoto S, Gotoh R, Koyama M, Yoshioka S, Takahashi T, Takahashi K, Yanagisawa T, Fukuda H (1996) Functional anatomy of GO/NO-GO discrimination and response selection--a PET study in man. Brain Res 728:79–89.

Kiehl KA, Liddle PF, Hopfinger JB (2000) Error processing and the rostral anterior cingulate: an event-related fMRI study. Psychophysiology 37:216–223.

Klecka WR, Iversen GR, . Klecka WR (1980) Discriminant Analysis. SAGE.

Koechlin E, Ody C, Kouneiher F (2003) The architecture of cognitive control in the human prefrontal cortex. Science 302:1181–1185.

Luria AR (2012) Higher Cortical Functions in Man. Springer Science & Business Media.

MacLeod CM, MacDonald PA (2000) Interdimensional interference in the Stroop effect: uncovering the cognitive and neural anatomy of attention. Trends Cogn Sci 4:383–391.

Mansouri FA, Tanaka K, Buckley MJ (2009) Conflict-induced behavioural adjustment: a clue to the executive functions of the prefrontal cortex. Nat Rev Neurosci 10:141–152.

Mante V, Sussillo D, Shenoy KV, Newsome WT (2013) Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature 503:78–84.

Merchant H, Crowe DA, Robertson MS, Fortes AF, Georgopoulos AP (2011) Top-down spatial categorization signal from prefrontal to posterior parietal cortex in the primate. Front Syst Neurosci 5:69.

Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. Annu Rev Neurosci 24:167–202.

Milner B (1963) Effects of Different Brain Lesions on Card Sorting: The Role of the Frontal Lobes. Arch Neurol 9:90–100.

Miocinovic S, Noecker AM, Maks CB, Butson CR, McIntyre CC (2007) Cicerone: stereotactic neurophysiological recording and deep brain stimulation electrode placement software system. Acta Neurochir Suppl 97:561–567.

Monsell S (2003) Task switching. Trends Cogn Sci 7:134–140.

Monsell S, Taylor TJ, Murphy K (2001) Naming the color of a word: Is it responses or task sets that compete? Mem Cognit 29:137–151.

Moore TL, Schettler SP, Killiany RJ, Rosene DL, Moss MB (2009) Effects on executive function following damage to the prefrontal cortex in the rhesus monkey (Macaca mulatta). Behav Neurosci 123:231–241.

Nakahara K, Hayashi T, Konishi S, Miyashita Y (2002) Functional MRI of macaque monkeys performing a cognitive set-shifting task. Science 295:1532–1536.

Nakamura K, Roesch MR, Olson CR (2005) Neuronal activity in macaque SEF and ACC during performance of tasks involving conflict. J Neurophysiol 93:884–908.

Olejnik S, Algina J (2003) Generalized eta and omega squared statistics: measures of effect size for some common research designs. Psychol Methods 8:434–447.

Pachitariu M, Steinmetz NA, Kadir SN, Carandini M, Harris KD (2016) Fast and accurate spike sorting of high-channel count probes with KiloSort. In: Advances in neural information processing systems, pp 4448–4456.

Parker A, Wilding E, Akerman C (1998) The Von Restorff effect in visual object recognition memory in humans and monkeys. The role of frontal/perirhinal interaction. J Cogn Neurosci 10:691–703.

Passingham RE (1985) Memory of monkeys (Macaca mulatta) with lesions in prefrontal cortex. Behav Neurosci 99:3–21.

Peterson BS, Skudlarski P, Gatenby JC, Zhang H, Anderson AW, Gore JC (1999) An fMRI study of Stroop word-color interference: evidence for cingulate subregions subserving multiple distributed attentional systems. Biol Psychiatry 45:1237–1258.

Petrie A (1952) Personality and the frontal lobes: An investigation of the psychological effects of different types of leucotomy. Available at: http://dx.doi.org/10.1037/14904-000.

Ploner CJ, Ostendorf F, Brandt SA, Gaymard BM, Rivaud-Péchoux S, Ploner M, Villringer A, Pierrot-Deseilligny C (2001) Behavioural relevance modulates access to spatial working memory in humans. Eur J Neurosci 13:357–363.

Rabbitt PM (1964) IGNORING IRRELEVANT INFORMATION. Br J Psychol 55:403–414.

Racz FS, Mukli P, Nagy Z, Eke A (2017) Increased prefrontal cortex connectivity during cognitive challenge assessed by fNIRS imaging. Biomed Opt Express 8:3842–3855.

Razik TA (1971) Relevant and irrelevant dimensions in concept formation. Theory Pract 10:109–116.

Rowe JB, Passingham RE (2001) Working memory for location and time: activity in prefrontal area 46 relates to selection rather than maintenance in memory. Neuroimage 14:77–86.

Rowe JB, Toni I, Josephs O, Frackowiak RS, Passingham RE (2000) The prefrontal cortex: response selection or maintenance within working memory? Science 288:1656–1660.

Roy JE, Buschman TJ, Miller EK (2014) PFC neurons reflect categorical decisions about ambiguous stimuli. J Cogn Neurosci 26:1283–1291.

Sarma A, Masse NY, Wang X-J, Freedman DJ (2016) Task-specific versus generalized mnemonic representations in parietal and prefrontal cortices. Nat Neurosci 19:143–149.

Seo H, Cai X, Donahue CH, Lee D (2014) Neural correlates of strategic reasoning during competitive games. Science 346:340–343.

Shallice T (1982) Specific impairments of planning. Philos Trans R Soc Lond B Biol Sci 298:199–209.

Shallice T, Burgess PW (1991) Deficits in strategy application following frontal lobe damage in man. Brain 114 (Pt 2):727–741.

Simon HA (1975) The functional equivalence of problem solving skills. Cogn Psychol 7:268–288.

Sleezer BJ, Castagno MD, Hayden BY (2016) Rule Encoding in Orbitofrontal Cortex and Striatum Guides Selection. J Neurosci 36:11223–11237.

Stroop JR (1936) Studies of interference in serial verbal reactions. J Exp Psychol 18:643.

Stuss DT, Knight RT (2013) Principles of Frontal Lobe Function. OUP USA.

Swaminathan SK, Freedman DJ (2012) Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. Nat Neurosci 15:315–320.

Szczepanski SM, Knight RT (2014) Insights into human behavior from lesions to the prefrontal cortex. Neuron 83:1002–1018.

Tanji J, Hoshi E (2008) Role of the lateral prefrontal cortex in executive behavioral control. Physiol Rev 88:37–57.

Tsuchida A, Fellows LK (2009) Lesion evidence that two distinct regions within prefrontal cortex are critical for n-back performance in humans. J Cogn Neurosci 21:2263–2275.

van Moorselaar D, Slagter HA (2020) Inhibition in selective attention. Ann N Y Acad Sci Available at: http://dx.doi.org/10.1111/nyas.14304.

Voytek B, Davis M, Yago E, Barceló F, Vogel EK, Knight RT (2010) Dynamic neuroplasticity after human prefrontal cortex damage. Neuron 68:401–408.

Widge AS, Heilbronner SR, Hayden BY (2019) Prefrontal cortex and cognitive control: new insights from human electrophysiology. F1000Res 8 Available at: http://dx.doi.org/10.12688/f1000research.20044.1.

Wutz A, Loonis R, Roy JE, Donoghue JA, Miller EK (2018) Different Levels of Category Abstraction by Different Dynamics in Different Prefrontal Areas. Neuron 97:716–726.e8.

Zanto TP, Gazzaley A (2009) Neural suppression of irrelevant information underlies optimal working memory performance. J Neurosci 29:3059–3066.