Gradual progression from sensory to task-related processing in cerebral cortex

Scott L. Brincat[a,1], Markus Siegel[a,b,1], Constantin von Nicolai[b], and Earl K. Miller[a,2]

[a]The Picower Institute for Learning and Memory, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; and [b]Centre for Integrative Neuroscience & MEG Center, University of Tübingen, 72076 Tübingen, Germany

Edited by Peter L. Strick, University of Pittsburgh, Pittsburgh, PA, and approved June 14, 2018 (received for review September 29, 2017)

Somewhere along the cortical hierarchy, behaviorally relevant information is distilled from raw sensory inputs. We examined how this transformation progresses along multiple levels of the hierarchy by comparing neural representations in visual, temporal, parietal, and frontal cortices in monkeys categorizing across three visual domains (shape, motion direction, and color). Representations in visual areas middle temporal (MT) and V4 were tightly linked to external sensory inputs. In contrast, lateral prefrontal cortex (PFC) largely represented the abstracted behavioral relevance of stimuli (task rule, motion category, and color category). Intermediate-level areas, including posterior inferotemporal (PIT), lateral intraparietal (LIP), and frontal eye fields (FEF), exhibited mixed representations. While the distribution of sensory information across areas aligned well with classical functional divisions (MT carried stronger motion information, and V4 and PIT carried stronger color and shape information), categorical abstraction did not, suggesting these areas may participate in different networks for stimulus-driven and cognitive functions. Paralleling these representational differences, the dimensionality of neural population activity decreased progressively from sensory to intermediate to frontal cortex. This shows how raw sensory representations are transformed into behaviorally relevant abstractions and suggests that the dimensionality of neural activity in higher cortical regions may be specific to their current task.

categorization | cognition | prefrontal cortex | posterior parietal cortex | dimensionality

Neural representations at the earliest stages of cortical processing reflect a relatively faithful copy of sensory inputs, but intelligent behavior requires abstracting the behaviorally relevant elements from sensory inputs. The sensory continuum often needs to be parsed into categories, such as dividing continuous color variations of fruits into "ripe" and "unripe." Arbitrary sensory stimuli can also be functionally associated to acquire the same meaning (e.g., the diverse stimuli grouped into the category "food"). Impaired or atypical categorization is a hallmark of disorders such as autism (1) and schizophrenia (2). Understanding its neural basis could provide pathways to early diagnosis and treatment.

Abstract categorical representations can be found in areas at or near the top of the cortical hierarchy, such as lateral prefrontal cortex (PFC) (3–7), posterior parietal cortex (7–9), and the medial temporal lobe (10). Less well understood are the processing steps that transform bottom-up sensory inputs into these task-related, and thus top-down, representations. We therefore recorded from multiple regions along the cortical hierarchy in macaque monkeys performing a multidimensional categorization task. In a previous report on this dataset (11), we showed evidence that sensory signals flow in a bottom-up direction from visual to frontal cortex, while signals for the monkeys' behavioral choice flow in a top-down direction from frontoparietal to visual cortex.

Here, we exploit the category structure of this task to investigate the degree to which visual representations in six cortical areas reflect bottom-up sensory inputs or the learned categories they are grouped into. The task required binary categorization of stimuli continuously varying along two distinct sensory domains, motion direction and color, and arbitrary grouping of a set of shape cues that signaled which feature (motion or color) should be categorized on each trial. We recorded isolated neurons simultaneously from six cortical areas (Fig. 1*D*) from both the dorsal and ventral visual processing streams, including frontal [lateral PFC and frontal eye fields (FEF)], parietal [lateral intraparietal area (LIP)], temporal [posterior inferior temporal cortex (PIT)], and visual [areas V4 and middle temporal (MT)] cortices. Our results suggest that categorization occurs in a gradual fashion across the cortical hierarchy, reaching its apex in PFC; that categorical coding does not always correlate with classical functional divisions; and that the dimensionality of cortical activity decreases in parallel with the reduction of continuous sensory stimuli to categorical groupings.

Results

Our main interest was to track the transformation of visual inputs from more sensory (bottom-up) representations to task-related (top-down) representations. On each trial of our multidimensional categorization task (Fig. 1*A*), a visual shape cue instructed the monkey whether to categorize a subsequently presented colored, moving random-dot stimulus based on its color ("greenish" vs. "reddish") or direction of motion (upward vs. downward), and report the cued category with a leftward or rightward saccade. Therefore, it probed three different types of sensory inputs: shape, motion, and color. Two of the shapes (arbitrarily chosen) cued motion categorization, while the other two cued color categorization (Fig. 1 *B* and *C*). Thus, four different shapes were arbitrarily

Significance

The earliest stages of processing in cerebral cortex reflect a relatively faithful copy of sensory inputs, but intelligent behavior requires abstracting behaviorally relevant concepts and categories. We examined how this transformation progresses through multiple levels of the cortical hierarchy by comparing neural representations in six cortical areas in monkeys categorizing across three visual domains. We found that categorical abstraction occurred in a gradual fashion across the cortical hierarchy and reached an apex in prefrontal cortex. Categorical coding did not respect classical models of large-scale cortical organization. The dimensionality of neural population activity was reduced in parallel with these representational changes. Our results shed light on how raw sensory inputs are transformed into behaviorally relevant abstractions.

Author contributions: M.S. and E.K.M. designed research; M.S. performed research; S.L.B. analyzed data; S.L.B. and E.K.M. wrote the paper; and S.L.B. and C.v.N. spike-sorted the data.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]S.L.B. and M.S. contributed equally to this work.

[2]To whom correspondence should be addressed. Email: ekmiller@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1717075115/-/DCSupplemental.

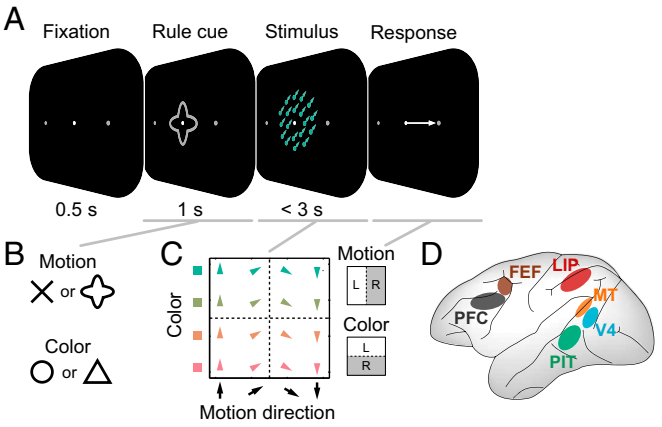

Fig. 1. Experimental design. (*A*) Trial sequence for the multidimensional visual categorization task. On each trial, the monkeys categorized either the motion direction or color of a random-dot stimulus. This stimulus was immediately preceded by a symbolic visual shape cue that instructed which feature (motion or color) to categorize for that trial. The monkey responded with a leftward or rightward saccade during the 3-s stimulus. (*B*) Either of two different cue shapes was used to instruct each task rule so as to dissociate cue- and task-rule–related activity. (*C*) Stimuli systematically sampled motion direction (upward to downward) and color (green to red). Each color category comprised two distinct colors, and each motion category comprised two distinct motion directions (additional ambiguous stimuli on the category boundaries were not analyzed here due to our focus on categoricality). Dashed lines indicate category boundaries. For each task rule, the two categories had a fixed mapping to a leftward (L) or rightward (R) saccadic response. (*D*) Illustration of sampled brain regions: lateral PFC, FEF, LIP, PIT, V4, and MT.

grouped into pairs by virtue of them cueing the same task rule, and four continuously varying colors and directions of motion were arbitrarily divided by a sharp boundary (Fig. 1*C*) [additional colors/motions on category boundaries (11) were excluded from the present analyses, which require an unambiguous category assignment].

We exploited the mapping in each domain from two stimulus items (cue shapes, directions, or colors) to each categorical grouping (task rule, motion category, or color category) to dissociate stimulus-related (sensory) and task-related (categorical) effects. Purely categorical neural activity would differentiate between categories (i.e., have "preferred" responses for both items in the same category) but show no differences between items within each category. Purely sensory activity would, instead, differentiate between stimulus items without regard to the learned categorical divisions.

We quantified this by fitting each neuron's spike rate, at each time point, with a linear model that partitioned across-trial rate variance into between-category and within-category effects (details are provided in *SI Appendix*, *SI Methods*). The model included three orthogonal contrast terms for each task domain (Fig. 2*A*). One contrast (blue) reflected the actual task-relevant grouping of stimulus items (cue shapes, directions, or colors) into categories, and thus captured between-category variance. The other contrasts (gray) reflected the two other possible non–task-relevant paired groupings of items and captured all within-category variance. Together, these three terms capture all data variance in the given task domain. We wished to measure how much of that variance, for each domain and studied brain region, was attributable to categorical coding: its categoricality. Note that simply measuring the between-category variance would result in a biased estimate of categoricality; it is nonzero even for neural populations with sensory tuning for single stimulus items or for arbitrary subsets of items (*SI Appendix*, Fig. S1 *E* and *F*).

Instead, we estimated where the between-category variance of each neural population fell between the predictions of purely sensory and purely categorical coding. Note that the sum of variances for all three model terms bounds the between-category variance; they can be equal only for a perfectly categorical population with zero within-category variance (Fig. 2*B*, *Top*). A purely sensory-driven population would, instead, have equal variance for all three contrasts; thus, between-category variance would equal the average of all three terms (Fig. 2*B*, *Bottom*). To measure where neural populations fall between these extremes, we computed a "categoricality index" equal to the area between the between-category and sensory (lower-bound) time series, expressed as a fraction of the full area between the total (upper-bound) and sensory (lower-bound) time series (Fig. 2*C*). It can be shown this is also equivalent to the between-category variance minus the average of within-category variance terms [the statistic used in our prior publication (11)], normalized by the total domain variance (details are provided in *SI Appendix*, *SI Methods*). This index is a specific measure of how categorical a neural population is, and ranges from 1 for a perfectly categorical

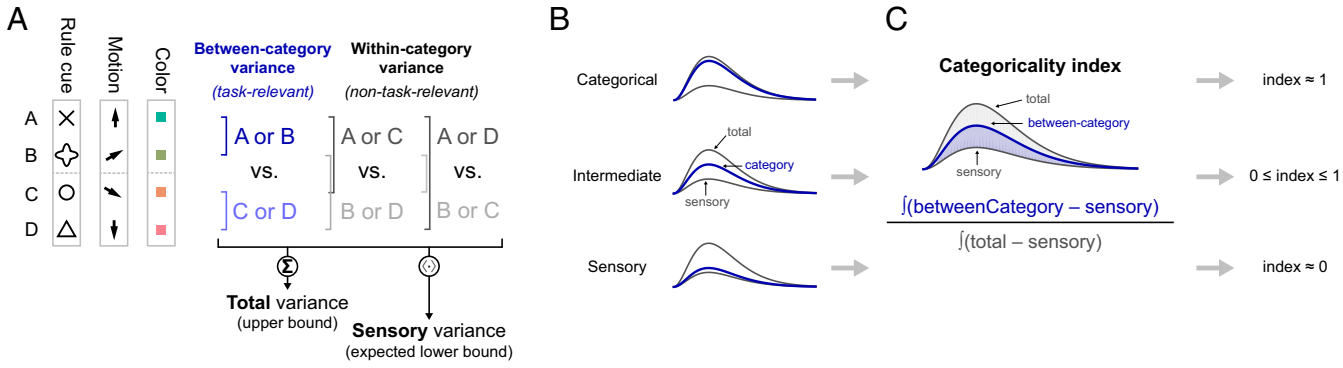

Fig. 2. Illustration of analysis methods. (*A*) Spike rate variance for each task variable (task cues/rules, motion directions, and colors) was partitioned into three orthogonal contrasts. One contrast (blue) reflected the actual task-relevant grouping of stimulus items (cue shapes, directions, or colors) into categories, and thus captured between-category variance. The other contrasts (gray) reflected the two other possible non–task-relevant paired groupings of items, and, together, captured all within-category variance. An additional term in the analysis (not depicted) partitions out variance related to the behavioral choice (left vs. right saccade). Details are provided in *SI Appendix*, *SI Methods*. (*B*, *Top*) The sum of variances for all three contrasts (Σ in *A*) bounds the between-category variance; the total and between-category variances can be equal only for a perfectly categorical neural population with zero within-category variance. (*B*, *Bottom*) Purely sensory-driven population would, instead, have equal variance for all three contrasts, and thus between-category variance would equal the average (⟨ · ⟩ in *A*) of all three contrasts. (*C*) The categoricality index measured where actual neural populations fell between these extremes, in terms of the area between the between-category and sensory lower-bound time series, expressed as a fraction of the full area between the upper and lower bounds. Values of 0 and 1 correspond to purely sensory and purely categorical populations, respectively. Details are provided in *SI Appendix*, *SI Methods*.

population to 0 for a purely sensory population. [Negative values are possible if within-category variance is greater than between-category variance (i.e., for populations that specifically reflect within-category differences).]

Within the context of each task rule, motion and color categories were, by design, inextricably linked with the monkey's behavioral choice (e.g., under the color rule, greenish and reddish colors always mandated leftward and rightward saccades, respectively). Although this identity relationship is broken when both task rules are considered together, there remains a partial correlation between these task variables. To partition out choice effects from the category effects of interest, we also included in the fitted models a term reflecting behavioral choice. Category effects are thus measured in terms of their additional variance explained once choice effects are already accounted for (12).

To validate our analysis, we first assayed its properties on synthesized neural activity with known ground truth (details are provided in *SI Appendix*, *SI Methods* and *SI Results*). We show that our categoricality index reliably reports the relative weights of simulated sensory and categorical signals (*SI Appendix*, Fig. S1*A*), that it is relatively insensitive to coupled changes in both sensory and categorical signals in concert (*SI Appendix*, Fig. S1 *B* and *C*), and that it is relatively insensitive to simulated choice effects (*SI Appendix*, Fig. S1*D*).

Shape Information. We first examined representations of the four cue shapes instructing the task rule in effect on each trial. All sampled cortical areas (MT, V4, PIT, LIP, FEF, and PFC) conveyed significant information about the rule cues, as measured by the total spike rate variance explained by cues (Fig. 3 *A* and *B*; $P < 0.01$, one-sample bootstrap test). The strongest cue shape information was in areas PIT and V4 (Fig. 3*C*; $P < 0.01$ for all comparisons with other areas, two-sample bootstrap test), consistent with their well-established role in shape processing (13).

To measure task-related (top-down) information about the task rule instructed by the cues, we partitioned out spike rate variance due to effects between task rules ("between-category variance") and between cues instructing the same rule ("within-category variance"). Areas MT, V4, and PIT all exhibited between-category variances (Fig. 3*D*, colored curves) that hewed closely to values expected from a pure bottom-up sensory representation of shape (Fig. 3*D*, lower gray curves). We summarized these results with a categoricality index that measures how categorical the information conveyed by each neural population is, ranging continuously from purely sensory (0) to purely categorical (1). Task-rule categoricality indices for each of these visual areas (Fig. 3*E*) did not differ significantly from zero ($P > 0.01$). This was true for both V4 and PIT, areas where we found strong overall cue information, as well as for MT, where there was weaker cue information. Thus, visual areas MT, V4, and PIT contained a primarily bottom-up sensory representation of the shape cues. Note that this result differs from the strong V4 and PIT task-rule signals in our prior publication on this dataset (11). This is primarily due to differences in the specific questions addressed by each study, and can be reconciled by the fact that V4 and PIT do contain some task-rule signals, but these signals constitute a very small fraction of the total cue variance in these areas (details are provided in *SI Appendix*, *SI Results*).

By contrast, PFC, FEF, and LIP all conveyed task-rule information (Fig. 3*D*, colored curves) well above that predicted from bottom-up sensory signals (Fig. 3*D*, lower gray curves), and had task-rule categoricality indices significantly greater than zero (Fig. 3*E*; $P \leq 1 \times 10^{-5}$ for all three areas). PFC exhibited the most categorical task cue representation, significantly greater than all other areas (Fig. 3*F*; $P < 0.01$) except FEF ($P = 0.05$). FEF and LIP had intermediate values between PFC and the group of sensory areas (MT, V4, and PIT). All areas, including PFC, still conveyed less task-rule information than expected from a purely categorical representation (Fig. 3*D*, upper gray curves) and had categoricality indices significantly less than 1 ($P \leq 1 \times 10^{-5}$). This suggests that, unlike the visual areas, areas LIP, FEF, and particularly PFC represented the top-down meaning of the rule cues, although also retaining some sensory information about them as well. Unlike the case with sensory information, where results were predictable from traditional areal divisions (shape-coding ventral stream areas PIT and V4 showed the strongest information), top-down task-rule coding was observed in traditionally non–shape-coding dorsal stream areas LIP and FEF, but not in V4 or PIT.

It seems likely that some trace of the current task rule would have to persist into the stimulus period to influence how the random-dot stimulus was assigned to categories. This can be seen in the rightmost portion of Fig. 3*D*, and we focus in on it in *SI Appendix*, Fig. S2. The results indicate that cue signals in V4 and PIT decrease sharply after cue offset (*SI Appendix*, Fig. S2*A*) and remain sensory in nature (*SI Appendix*, Fig. S2 *D* and *E*). In contrast, cue signals in PFC, FEF, and LIP persist relatively unabated through the stimulus period (*SI Appendix*, Fig. S2*D*) and are generally even more categorical than during the cue period (*SI Appendix*, Fig. S2 *E* and *F*). These results indicate that task-rule signals in frontoparietal regions (PFC, FEF, and LIP) may be involved in guiding categorical decisions about the random-dot stimuli.

Motion Direction Information. Next, we turned to the random-dot stimuli that the animals had to categorize according to their direction of motion or their color. Both motion direction and color varied along a continuum, but the animals had to group them into upward/downward or greenish/reddish. Much as with the rule cues (as discussed above), we would expect bottom-up sensory signals to reflect the actual direction or color, whereas task-related signals should divide them into their relevant categories. To discriminate signals related to stimulus category and behavioral choice, we combined data across both task rules (as discussed above; details are provided in *SI Appendix*, *SI Methods*).

First, we began with motion. Analogous to the rule cues, there were four distinct motion directions, grouped into two categories: upward (90° and 30°) and downward (−30° and −90°) categories, with rightward motion (0°) serving as the category boundary (Fig. 1*C*). All areas conveyed significant information about motion direction, as measured by the total spike rate variance explained by motion (Fig. 4 *A* and *B*; $P < 0.01$). The strongest motion information was found in area MT (Fig. 4 *A* and *B*), which was significantly greater than in all other areas (Fig. 4*C*; $P < 0.01$) except PIT ($P = 0.05$), consistent with its classical role in motion processing (14).

When we partitioned total motion variance into between- and within-category effects, between-category variance in MT (Fig. 4*D*, orange curve) closely approximated its sensory prediction (Fig. 4*D*, lower gray curve), and its motion categoricality index was not significantly different from zero (Fig. 4*E*; $P = 0.44$). The same was true of more weakly direction-selective areas V4 ($P \approx 1$), PIT ($P \approx 1$), and FEF ($P = 0.03$). Thus, the motion information carried by MT, V4, PIT, and FEF was largely sensory in nature.

In contrast, PFC and LIP both exhibited between-category variances (Fig. 4*D*, colored curves) considerably greater than their sensory predictions (Fig. 4*D*, lower gray curves) and motion categoricality indices significantly greater than zero (Fig. 4*E*; PFC: $P = 0.004$, LIP: $P \leq 1 \times 10^{-5}$). As with the task instruction cues, PFC showed the most categorical motion signals (Fig. 4*F*; significantly greater than MT, V4, and PIT: all $P < 0.01$; nonsignificant for FEF: $P = 0.06$ and LIP: $P = 0.38$). All areas, including PFC, remained significantly below the predictions of a purely categorical representation (Fig. 3*D*, upper gray curves; $P \leq 1 \times 10^{-5}$). Thus, areas PFC and LIP conveyed top-down motion categories but retained some sensory direction information as well. Once again, while the sensory results were consistent with traditional areal divisions (MT predictably had the strongest direction information), significant motion category information was observed in one higher level dorsal stream area (LIP), but not in another one (FEF).

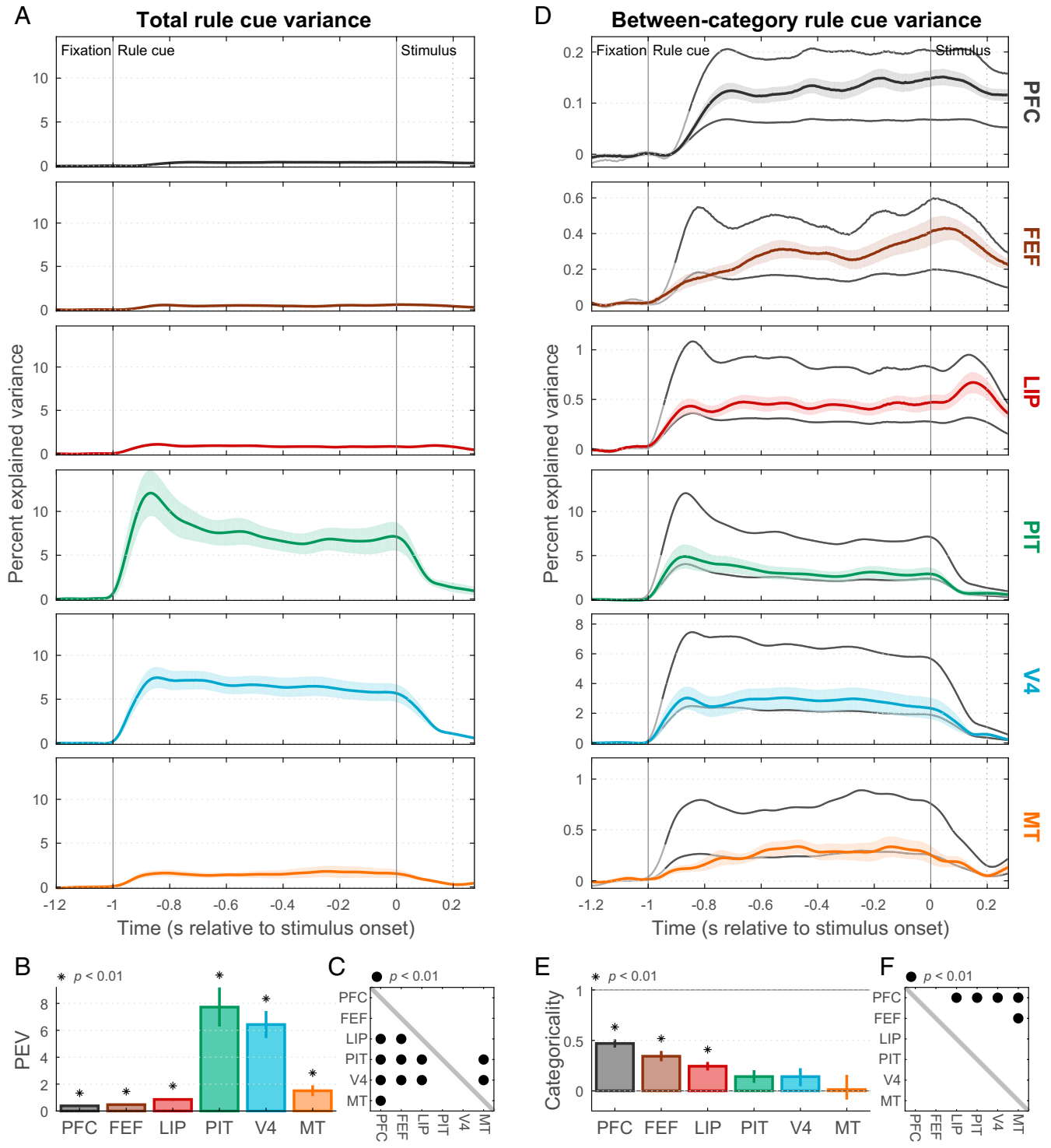

Fig. 3. Task-rule cue representation. (*A*) Population mean (±SEM) total spike rate variance explained by task-rule cues (cue information) in each studied area as a function of within-trial time [referenced to the onset of the random-dot stimulus]. (*B*) Summary (across-time mean ± SEM) of total rule-cue variance for each area. All areas contain significant cue information (*$P < 0.01$). PEV, percent explained variance. (*C*) Cross-area comparison matrix indicating which regions (rows) had significantly greater cue information than others (columns). Dots indicate area pairs that attained significance (•$P < 0.01$). PIT and V4 contain significantly greater cue information than all other areas. (*D*) Mean (±SEM) between-category rule-cue variance (task-rule information; colored curves). Gray curves indicate expected values of this statistic corresponding to a purely categorical representation of rules (upper line) and to a purely sensory representation of rule cues (lower line). The transitions from light to dark gray in these curves indicate the estimated onset latency of overall cue information, which was used as the start of the summary epoch for each area. Note differences in *y*-axis ranges from *A*. (*E*) Task-rule categoricality index (±SEM) for each area, reflecting where its mean between-category rule-cue variance falls between its expected values for pure sensory (0) and categorical (1) representations. Only PFC, FEF, and LIP are significantly different from zero (*$P < 0.01$). (*F*) Cross-area comparison matrix indicating which regions (rows) had significantly greater task-rule categoricality indices than others (columns) (•$P < 0.01$). PFC was significantly greater than all others, except FEF.

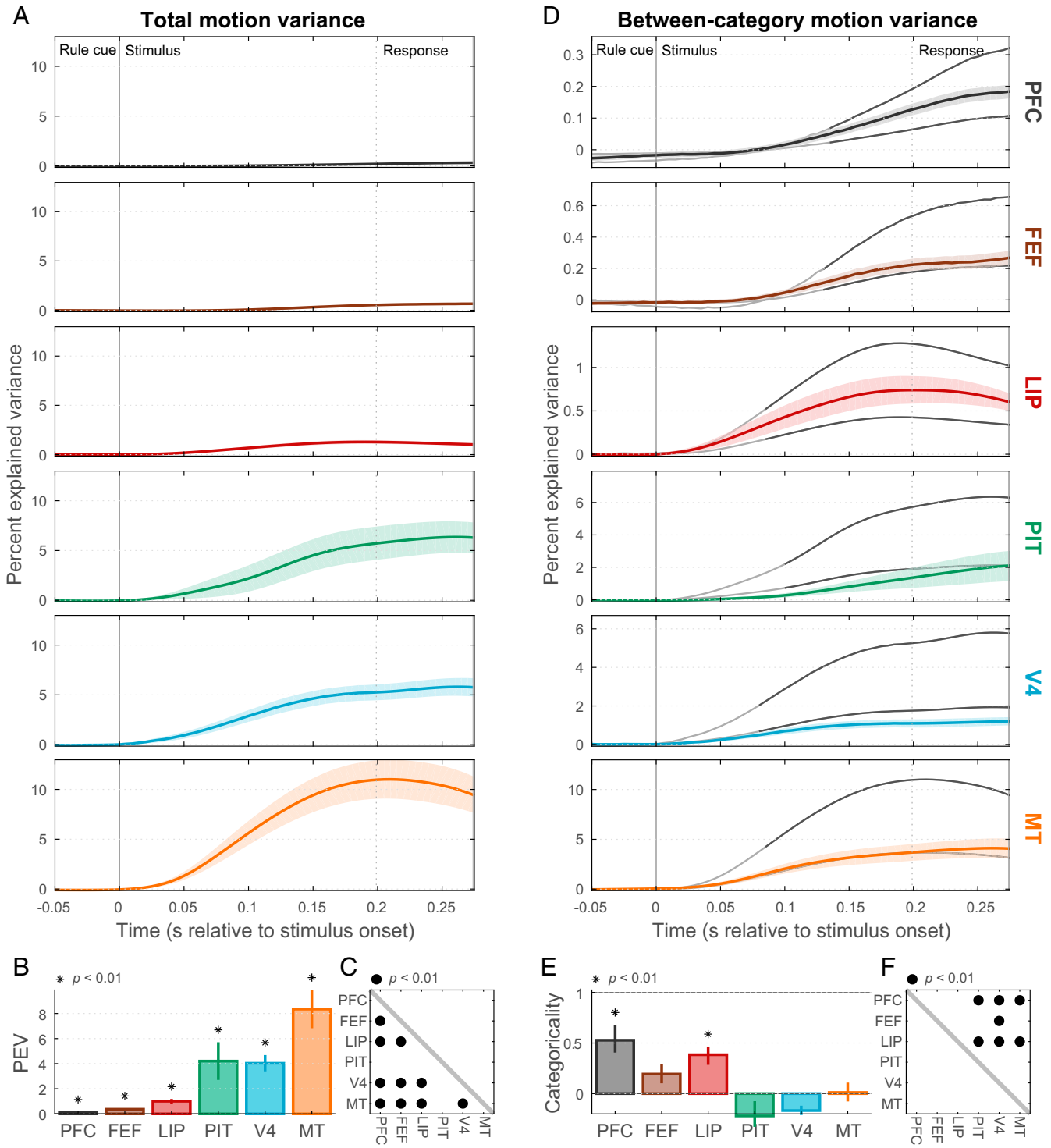

Fig. 4. Motion direction representation. (*A*) Mean (±SEM) total rate variance explained by random-dot stimulus motion directions (motion information) in each area as a function of time (note different time axis from Fig. 3). (*B*) Summary (across-time mean ± SEM) of total motion variance for each area. All areas contain significant motion information (*$P < 0.01$), but it was strongest in MT. PEV, percent explained variance. (*C*) Cross-area comparison matrix indicating which regions (rows) had significantly greater motion information than others (columns). (*D*) Mean (±SEM) between-category motion variance (motion category information). Gray curves indicate expected values for purely categorical (upper line) and purely sensory (lower line) representations of motion direction. (*E*) Motion categoricality index (±SEM) for each area, reflecting where its average between-category motion variance falls between expected values for pure sensory (0) and categorical (1) representations. Only PFC and LIP are significantly different from zero (*$P < 0.01$). (*F*) Cross-area comparison matrix indicating which regions (rows) had significantly greater motion categoricality indices than others (columns) (•$P < 0.01$).

Color Information. Next, we examined neural information about the colors of the random-dot stimuli. As with motion, there were four distinct color hues grouped into two categories, greenish (90° and 30° hue angles) and reddish (−30° and −90°), with the category boundary at yellow (0°; Fig. 1*C*). Once again, significant color information (total color variance) could be found in all

studied areas (Fig. 5 *A* and *B*; $P < 0.01$). Area V4 showed the strongest color information (Fig. 5 *A* and *B*), which was significantly greater than in all other studied areas (Fig. 5*C*; $P < 0.01$), consistent with its established role in color processing (15). Although PIT showed the second strongest stimulus color information, it was not significantly greater than in other areas (Fig. 5*C*; $P > 0.01$), possibly due to insufficient sampling of the relatively sparsely distributed "color patches" in and around this area (16).

Cortical representations of color were overall much more categorical than those for motion direction and rule cues, possibly due to alignment of the category structure with the red-green opponent signals arising in retinal ganglion cells and prevalent at many levels of the visual system (15). All areas except MT showed significant color categoricality indices (Fig. 5 *D* and *E*; $P < 0.01$ for all areas). PIT, FEF, and PFC all had nearly purely categorical color representations. For each of them, categorical information nearly equaled its upper bound (Fig. 5*D*). Their categoricality indices were significantly greater than those of LIP and V4 (Fig. 5*F*; $P < 0.01$), and those of PIT and FEF were not significantly different from a purely categorical representation (PIT: $P = 0.1$; FEF: $P = 0.04$). By comparison, strongly color-selective area V4, as well as weakly color-selective areas MT and LIP, was much less categorical. Thus, areas MT, V4, and LIP have a relatively bottom-up representation of color, while areas PIT, FEF, and PFC have largely categorized them into binary greenish and reddish categories. Note that while bottom-up biases toward red-green opponent coding might have boosted the overall apparent color categoricality, it is not obvious why such signals would be inherently stronger in higher level areas than in V4. We also found that, once again, overall color information was fairly consistent with traditional areal divisions (V4 predictably had the strongest color information), while color categoricality exhibited a mixed correlation with them: As expected, PIT was strongly categorical for color, but so was traditional dorsal stream area FEF.

Changes in Dimensionality Across Cortex. As a complementary assay of cortical coding properties, we examined the dimensionality of neural population activity using a noise-thresholded principal components analysis method (17) (details are provided in *SI Appendix*, *SI Methods*). This analysis was previously used to measure PFC dimensionality in an object sequence memory task (18). In that context, it was found that PFC neurons contained a high-dimensional representation of task components due to their conjunctive coding of multiple task variables (18). We asked whether high dimensionality is an invariant property of PFC population activity or whether it might be specific to task context. We concatenated all neurons from each studied area into a "pseudo-population" and extrapolated to larger population sizes via a condition relabeling procedure. For each area and population size, we computed the mean spike rate for each of 64 task conditions (four rule cues × four motion directions × four colors) within the random-dot stimulus epoch when all conditions were differentiated. The dimensionality of the space spanned by each resulting set of 64 neural population activity vectors was quantified as the number of principal components (eigenvalues) significantly greater than those estimated to be due solely to noise.

As expected, estimated dimensionality grew with the size of the neural population examined but generally approximated an asymptotic value (Fig. 6*A*) that can be taken as an estimate of the dimensionality of the underlying (much larger) neural population. Clear differences in the estimated dimensionality were observed across areas, as summarized in Fig. 6*B*. The highest dimensionality was observed in visual areas V4, PIT, and MT, presumably reflecting a large diversity of sensory tuning curves in these visual areas. These areas were followed by intermediate-level visual areas LIP and FEF. The lowest dimensionality was observed in PFC. The observed high dimensionality of area PIT was likely due, in part, to the inclusion of two task variables that it carried relatively strong information about: rule cues (shape) and color. When the same analysis was performed in 16D space consisting only of four directions × four colors (Fig. 6 *C* and *D*), PIT dimensionality was greatly reduced and was similar to that of LIP.

Thus, the dimensionality of population activity decreased progressively up the cortical hierarchy in parallel with the gradual shift from sensory to categorical representations. Further, the estimated PFC dimensionality is close to the value that would be expected of a purely categorical representation with binary responses for each task variable (~3 for the three-variable analysis, Fig. 6*B*; ~2 for the two-variable analysis, Fig. 6*D*). These results suggest that high dimensionality is not an invariant property of PFC activity but may be specific to current behavioral demands, to reduce high-dimensional sensory stimuli to binary categories in this case.

Discussion

Abstraction of Sensory Inputs Occurs Progressively Through the Cortical Hierarchy. Our results, summarized in Fig. 7, demonstrate a gradual progression from bottom-up sensory inputs to abstracted, top-down behaviorally relevant signals as the cortical hierarchy is ascended. Across three visual domains (shape, motion direction, and color), lower level visual areas MT and V4 conveyed strong information about sensory stimuli within their preferred domains (Fig. 7*A*) but showed little evidence for any abstraction beyond the raw sensory inputs (Fig. 7*B*). In contrast, higher level area PFC, despite containing relatively weak information overall (Fig. 7*A*), showed strongly abstracted, task-relevant coding across all domains (Fig. 7*B*). In between, intermediate-level visual areas PIT, LIP, and FEF showed mixed representations with partially categorical coding in some domains but not others. These results support models of cortical processing where representational transformations happen gradually across multiple cortical processing steps (19, 20), rather than in a discrete, all-or-nothing fashion.

In all but a few cases (PIT and FEF for color), cortical representations, even in high-level areas, remained significantly less categorical than predicted for a purely categorical neural population. The distribution of categoricality index values across neurons in the studied populations (*SI Appendix*, Fig. S3) suggests two reasons for this. First, even mostly categorical populations contained some residual sensory-coding neurons (index values ≈ 0). Second, all studied populations contained some individual neurons whose activity differentiates both between categories and between items within each category ($0 \leq \text{index} \leq 1$). Thus, the intermediate categorical coding we observed in most areas reflected a mixture of sensory and categorical effects at the level of both single neurons and the neural population. Purely categorical signals might exist in other brain regions, such as the medial temporal lobe (10). However, some multiplexing of categorical and residual sensory signals could also have a functional role, such as permitting behavior to be driven flexibly by multiple levels of abstraction.

Note that the exact values of categoricality indices may be influenced by the particular stimuli used, and are thus, to some degree, specific to this task. For example, the apparent strong categoricality of color representations may have been due to alignment of the category structure with the red-green opponent signals arising in retinal ganglion cells (15). However, it is not obvious why any such bottom-up stimulus biases would be inherently stronger in higher level areas, without being driven by a learned category structure. Thus, while we do not put strong interpretational value on the exact index values for each area and task variable, we believe their relative values accurately portray a progression from sensory-dominated to categorical coding through the cortical hierarchy.

Comparison with Our Previous Results. One result may appear to be somewhat at odds with our prior publication on this dataset (11). We previously claimed that V4 and PIT had strong task-rule information (figure 2C in ref. 11). Here, we claim task-rule

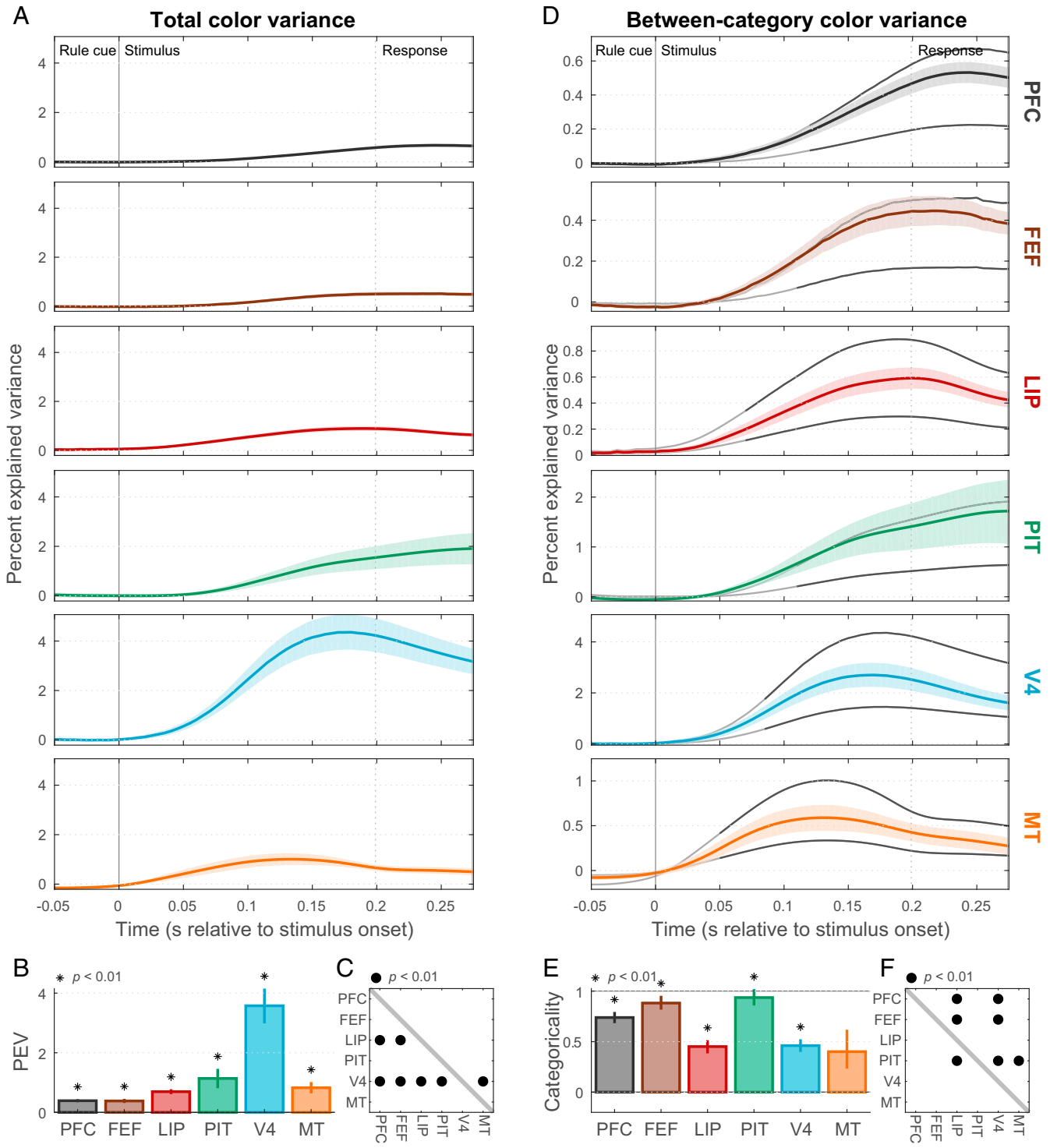

Fig. 5. Color representation. (*A*) Mean (±SEM) total rate variance explained by random-dot stimulus colors (color information) in each area. (*B*) Summary (across-time mean ± SEM) of color information for each area. All areas contain significant information (*$P < 0.01$), but V4 carried the strongest color information. (*C*) Cross-area comparison matrix indicating which regions (rows) had significantly greater color information than others (columns) (•$P < 0.01$). (*D*) Mean (±SEM) between-category color variance (color category information). Gray curves indicate expected values for purely categorical (upper line) and purely sensory (lower line) representations of color. (*E*) Color categoricality index (±SEM) for each area. All areas except MT had indices significantly greater than zero (*$P < 0.01$). (*F*) Cross-area comparison matrix indicating which regions (rows) had significantly greater color categoricality indices than others (columns) (•$P < 0.01$).

categoricality is weak and nonsignificant in these areas (Fig. 3 *D* and *E*). This difference lies primarily in the specific questions addressed by each study. Our previous study addressed the overall task-rule information conveyed by each neural population. It therefore used a statistic that measured a debiased version of between-category variance for task cues. This measure could be high for populations conveying strong information about a domain, such as the representations of rule cue shapes in V4 and PIT, even if only a very small fraction of that information is categorical (simulations in *SI Appendix*, Fig. S1*H*).

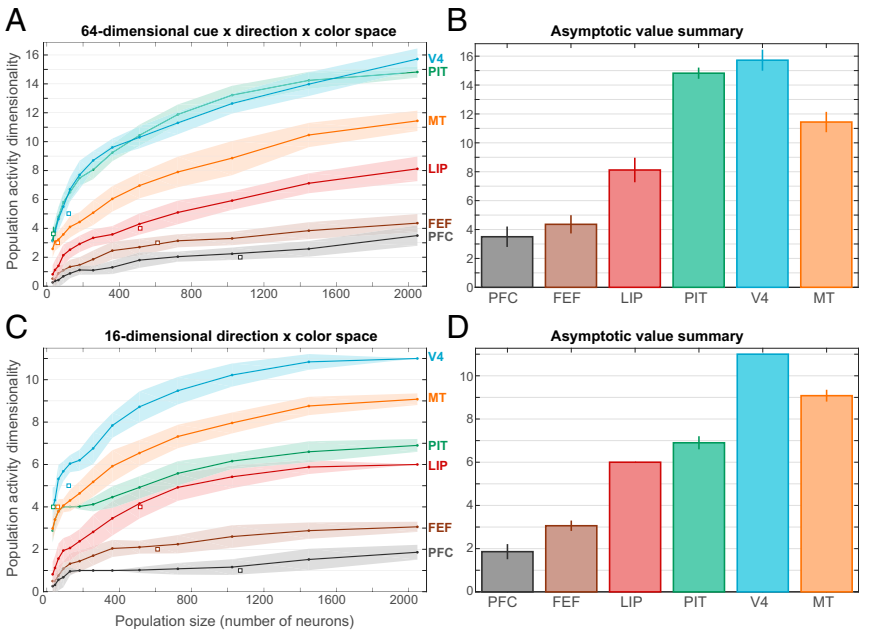

Fig. 6. Population activity dimensionality. (*A*) Dimensionality (mean ± SEM) of neural population activity as a function of extrapolated population size for each studied area. Dimensionality was estimated by noise-thresholded principal components analysis within a 64D rule cue × motion direction × color space (details are provided in *SI Appendix, SI Methods*). Values for the actual recorded neural populations (white squares ± SEM) were largely consistent with those from the extrapolated populations. (*B*) Summary of asymptotic dimensionality values (±SEM) in 64D space. (*C*) Dimensionality (mean ± SEM) of population activity as a function of population size for each studied area. Dimensionality was estimated within a reduced 16D motion direction × color space. (*D*) Summary of asymptotic dimensionality values (±SEM) in 16D space. V4 and MT have the highest dimensionality, followed by PIT and LIP and then by FEF and PFC.

Here, we instead addressed how categorical neural representations are. We used a statistic that normalizes out overall information to measure categoricality per se (*SI Appendix*, Fig. S1 *B* and *C*). Thus, we can reconcile results from the two studies by concluding that V4 and PIT contain strong information about task cues but only a small fraction of that information is categorical. In contrast, despite the weaker overall task cue information in PFC and FEF, a substantial fraction of that information reflects the learned task-rule categories. This definition accords well with both intuitive notions of categoricality and those previously proposed (3, 10). As elaborated below, it is further supported by its tight correspondence to the anatomically defined cortical hierarchy.

Graded Cortical Functional Specialization. There is a long-standing debate about the degree to which the functions of different cortical regions are specialized or overlapping. The cortex's broad interconnections and remarkable resistance to localized damage argue for more distributed, overlapping representations (21), while many studies find evidence for seemingly circumscribed functions in at least some cortical regions (22, 23). We find evidence supporting both points of view. Information about task variables was not distributed homogeneously across cortical regions. For each variable, one or two areas clearly conveyed much stronger information (Fig. 7*A*) than others. These results were largely predictable from the classical functions of visual cortical areas. Area MT was dominant for motion direction,

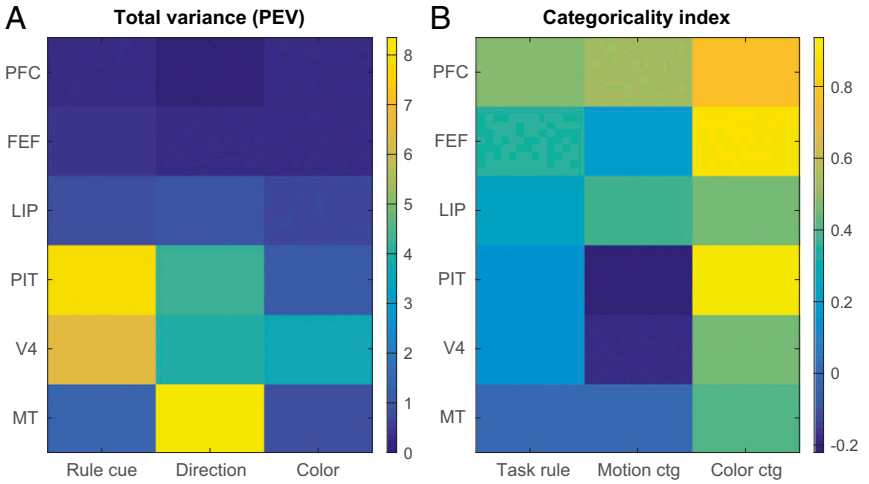

Fig. 7. Summary of results. (*A*) Mean total variance in each studied area explained by rule cues, motion directions, and colors. PEV, percent explained variance. (*B*) Categoricality indices for each studied area for task rules, motion categories, and color categories.

consistent with its well-established role in motion processing (14). V4 was dominant for color, consistent with many reports of its robust color selectivity (15). PIT and V4 both showed strong sensory information about rule cue shapes consistent with their well-established role in shape processing (13). Thus, these results support the idea of specialized cortical representations and reconfirm some of the classical functional divisions between ventral stream and dorsal stream areas using an experimental paradigm where multiple areas from both streams were tested simultaneously across multiple stimulus domains.

On the other hand, significant information about all examined experimental variables was found in all sampled areas, supporting the idea of broadly distributed cortical representations. The fact that color and shape information can be found in dorsal stream areas MT, LIP, and FEF and motion information can be found in ventral stream areas V4 and PIT argues against any absolute functional dichotomies between cortical processing streams, consistent with previous reports (24–26). We believe this body of results supports cortical models with graded functional specialization, where cortical areas have clear innate or learned biases to represent certain attributes but retain coarse, distributed information about nonspecialized attributes (27, 28).

While the overall strength of task-related information accorded well with classical divisions, the degree of top-down categorical abstraction painted a somewhat different picture. Dorsal stream area FEF exhibited a strongly categorical representation of color and task rule (derived from cue shape) but a noncategorical, sensory representation of motion direction. LIP was predictably categorical for motion but also showed a moderately categorical representation for task rule (shape) and color. While areas V4 and PIT were somewhat more predictable (they were relatively categorical for color but not at all for motion direction), they unexpectedly exhibited little to no categorical coding for the cue shape-derived task rule.

We quantified these observations by explaining the summary data in Fig. 7 with two predictors related to large-scale cortical organization: (*i*) the anatomically derived hierarchical level of each area (29) and (*ii*) the expected functional congruence of each combination of task variable and area, positive for those consistent with classical dorsal/ventral processing stream divisions (e.g., MT and motion, V4 and color) and negative for inconsistent combinations (e.g., MT and color, V4 and motion) (details are provided in *SI Appendix, SI Methods*). We found that both the decrease in sensory information and the increase in categorical coding across cortical areas were well explained by their anatomical hierarchical level ($P \leq 1 \times 10^{-5}$ for both), with only a marginally significant difference between them ($P = 0.04$). In contrast, only sensory information was also significantly explained by classical processing stream divisions ($P \leq 1 \times 10^{-5}$), while categoricality index values were not ($P = 0.11$), with a significant difference between them ($P = 0.003$). Thus, while our sensory information results confirm classical areal divisions, the degree of categorical coding is not well explained by them. These results suggest cortical regions may form different functional networks for bottom-up vs. top-down functions, putatively reflecting primarily feedforward and feedback/recurrent circuits, respectively.

Categorization Reached Its Apex in PFC. Many studies have now reported neural correlates of categories in PFC (3, 5–7), as well as in area LIP and related areas of posterior parietal cortex (6, 7, 9). A recent study suggested that LIP might contain a stronger representation of categories that could drive categorical processing in PFC (6). For all examined domains, we found that PFC exhibited a degree of categorical abstraction either greater than all other studied areas (task rule and motion) or not significantly different from the other most categorical area (color). For all domains, the prefrontal representation was more categorical than LIP, although this difference was significant only for task rule and color, and not for motion direction. On the other hand, despite being less categorical than PFC, LIP did also exhibit a significantly categorical representation for all tested domains, which was not the case for any other studied area besides PFC. We interpret these results to mean that PFC does play an important, and perhaps the most important, role in categorization (also ref. 7). However, LIP clearly also plays a central role, and categorization likely involves reciprocal interactions between these areas as well as others (30, 31).

Cortical Dimensionality May Be Task-Specific. We found a progression from high-dimensional population activity in the visual areas (V4 and MT) to low-dimensional populations in the frontal areas (PFC and FEF), paralleling the change in categoricality. We interpret this to reflect a shift from a large diversity of sensory tuning curves in visual cortex to nearly binary categorical responses in PFC.

At first glance, however, these results might seem at odds with a recent report showing prefrontal population activity is high dimensional (18). That study found that PFC neurons tend to exhibit "nonlinear mixed selectivity" for specific conjunctions of task variables, and, consequently, PFC population activity had a dimensionality near the theoretical maximum (24 dimensions) for the studied task. However, that study employed a task involving encoding and maintenance in working memory of a sequence of visual objects and responding via either a recollection or recall probe (32). Thus, correct performance required remembering which of 12 different sequences was shown and which of two modes of behavioral output was mandated. By contrast, the task used here emphasized dimensionality reduction. First, four visual cues were grouped into either of two task instructions. Next, 16 random-dot stimuli (four colors × four directions) were mapped onto binary color or motion categories, depending on the currently instructed task rule. Finally, the deduced category was translated into a binary response. Thus, this task, unlike the previous one, emphasized reduction of high-dimensional sensory inputs to lower dimensional abstractions. Our results therefore suggest the possibility that prefrontal dimensionality may flexibly reflect current cognitive demands (33). Inputs may be expanded to higher dimensions when decisions depend on multiple variables but reduced to lower dimensionality when categorical abstraction is required. Thus, PFC dimensionality, like other PFC coding properties (34), appears to flexibly adapt to behavioral needs.

Methods

Experimental methods are briefly reviewed here, but further details can be found in *SI Appendix, SI Methods*, as well as in our prior publication from this dataset (11). All procedures followed the guidelines of the Massachusetts Institute of Technology Committee on Animal Care and the NIH.

Electrophysiological Data Collection. In each of 47 experimental sessions, neuronal activity was recorded simultaneously from up to 108 electrodes acutely inserted daily into up to six cortical regions (Fig. 1*D*): MT, V4, PIT, LIP, FEF, and lateral PFC. All analyses were based on 2,414 well-isolated single neurons (MT: 60, V4: 121, PIT: 36, LIP: 516, FEF: 612, and PFC: 1,069). The basic analysis was also repeated using multiunit signals (pooling together all threshold-crossing spikes on each electrode), with very similar results (*SI Appendix*, Fig. S4). To minimize any sampling bias of neural activity, we did not prescreen neurons for responsiveness or selectivity. Details are provided in *SI Appendix, SI Methods*.

Behavioral Paradigm. Two adult rhesus macaques (*Macaca mulatta*) were trained to perform a multidimensional categorization task. On each trial (Fig. 1*A*), a visual cue instructed the monkey to perform one of two tasks: color categorization (greenish vs. reddish) or motion categorization (upward vs. downward) of a subsequently presented colored, moving random-dot stimulus. The monkey responded via a saccade toward a target to the left (greenish/upward) or right (reddish/downward). Details are provided in *SI Appendix, SI Methods*.

General Data Analysis. For most analyses, spike trains were converted into smoothed rates (spike densities). To summarize results, we pooled rates or other derived statistics within empirically defined epochs of interest for each task variable and area (details are provided in *SI Appendix, SI Methods*). Only correctly performed trials were included in the analyses. All hypothesis tests used distribution-free bootstrap methods unless otherwise noted.

Categoricality Analysis. Our primary interest was to characterize each cortical region's categoricality, the degree to which it reflected the raw sensory stimuli or their abstracted meaning (task rule or motion/color category). We quantified this by fitting each neuron's spike rate, at each time point, with a linear model that partitioned across-trial rate variance within each task domain into between-category and within-category effects (Fig. 2*A*). We then computed a categoricality index reflecting where the observed between-category variance for each population fell between the predictions of purely sensory and categorical coding (Fig. 2*B*). Because overall variance within each task domain is effectively normalized out of this index, it reflects a pure measure of the categorical quality of a neural representation, similar to previous measures of category selectivity (3), but taking the reliability of neural coding into account, as it is based on explained variance rather than raw spike rates. Details are provided in *SI Appendix, SI Methods*. Our analysis methods were validated with extensive simulations (*SI Appendix*, Fig. S1) and supported by a separate analysis comparing predictions of category and choice coding (*SI Appendix*, Fig. S5). Details are provided in *SI Appendix, SI Methods* and *SI Results*.

We also measured categoricality by comparing mean spike rates for preferred categories and stimulus items within categories. The relative difference in spike rates between and within categories was generally consistent with our presented results (*SI Appendix*, Fig. S6). Details are provided in *SI Appendix, SI Methods* and *SI Results*.

Population Dimensionality Analysis. To measure the dimensionality of population activity, we estimated the number of principal components required to describe the space spanned by condition-mean neural population activity vectors (17, 18). Epoch spike rates were computed for each trial and neuron, averaged across all trials of each condition, and concatenated across all neurons and sessions into a set of neural pseudo-population vectors for each studied area. Dimensionality was computed as the number of principal components (eigenvalues) of each resulting matrix significantly greater than the estimated distribution of principal components due to noise. Details are provided in *SI Appendix, SI Methods*.

Cortical Organization Analysis. To relate our results to classical models of cortical organization, we fit the data in each of the population summary matrices of Fig. 7 with a two-predictor linear model: (*i*) the Felleman and Van Essen (29) hierarchical level of each area and (*ii*) the expected functional congruence of each combination of task variable and area based on classical functional divisions, positive for consistent combinations (e.g., MT and motion) and negative for inconsistent ones (e.g., MT and color). Details are provided in *SI Appendix, SI Methods*.

ACKNOWLEDGMENTS. We thank Andre Bastos, Mikael Lundqvist, Morteza Moazami, Jefferson Roy, Jason Sherfey, and Andreas Wutz for helpful discussions, and Mattia Rigotti for providing code and advice on the dimensionality analysis. This work was supported by National Institute of Mental Health Grant 5R37MH087027 (to E.K.M.), European Research Council Grant StG335880 (to M.S.), Deutsche Forschungsgemeinschaft Grant DFG SI1332-3/1 (to M.S.), and the Centre for Integrative Neuroscience (Deutsche Forschungsgemeinschaft Grant EXC 307 to M.S.).

1. Klinger LG, Dawson G (1995) A fresh look at categorization abilities in persons with autism. *Learning and Cognition in Autism* (Springer, New York), pp 119–136.
2. Kéri S, et al. (1999) Category learning and perceptual categorization in schizophrenia. *Schizophr Bull* 25:593–600.
3. Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2001) Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* 291:312–316.
4. Wallis JD, Anderson KC, Miller EK (2001) Single neurons in prefrontal cortex encode abstract rules. *Nature* 411:953–956.
5. Roy JE, Riesenhuber M, Poggio T, Miller EK (2010) Prefrontal cortex activity during flexible categorization. *J Neurosci* 30:8519–8528.
6. Swaminathan SK, Freedman DJ (2012) Preferential encoding of visual categories in parietal cortex compared with prefrontal cortex. *Nat Neurosci* 15:315–320.
7. 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.
8. Stoet G, Snyder LH (2004) Single neurons in posterior parietal cortex of monkeys encode cognitive set. *Neuron* 42:1003–1012.
9. Freedman DJ, Assad JA (2006) Experience-dependent representation of visual categories in parietal cortex. *Nature* 443:85–88.
10. Kreiman G, Koch C, Fried I (2000) Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat Neurosci* 3:946–953.
11. Siegel M, Buschman TJ, Miller EK (2015) Cortical information flow during flexible sensorimotor decisions. *Science* 348:1352–1355.
12. Draper NR, Smith H (1998) *Applied Regression Analysis* (Wiley, New York), 3rd Ed.
13. Connor CE, Brincat SL, Pasupathy A (2007) Transformation of shape information in the ventral pathway. *Curr Opin Neurobiol* 17:140–147.
14. Born RT, Bradley DC (2005) Structure and function of visual area MT. *Annu Rev Neurosci* 28:157–189.
15. Conway BR (2009) Color vision, cones, and color-coding in the cortex. *Neuroscientist* 15:274–290.
16. Conway BR, Moeller S, Tsao DY (2007) Specialized color modules in macaque extrastriate cortex. *Neuron* 56:560–573.
17. Machens CK, Romo R, Brody CD (2010) Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. *J Neurosci* 30:350–360.
18. Rigotti M, et al. (2013) The importance of mixed selectivity in complex cognitive tasks. *Nature* 497:585–590.
19. Riesenhuber M, Poggio T (1999) Hierarchical models of object recognition in cortex. *Nat Neurosci* 2:1019–1025.
20. DiCarlo JJ, Zoccolan D, Rust NC (2012) How does the brain solve visual object recognition? *Neuron* 73:415–434.
21. Lashley KS (1963) *Brain Mechanisms and Intelligence: A Quantitative Study of Injuries to the Brain* (Dover, New York).
22. Fodor JA (1983) *The Modularity of Mind: An Essay on Faculty Psychology* (MIT Press, Cambridge, MA).
23. Kanwisher N (2010) Functional specificity in the human brain: A window into the functional architecture of the mind. *Proc Natl Acad Sci USA* 107:11163–11170.
24. Desimone R, Schein SJ (1987) Visual properties of neurons in area V4 of the macaque: Sensitivity to stimulus form. *J Neurophysiol* 57:835–868.
25. Sereno AB, Maunsell JHR (1998) Shape selectivity in primate lateral intraparietal cortex. *Nature* 395:500–503.
26. Peng X, Sereno ME, Silva AK, Lehky SR, Sereno AB (2008) Shape selectivity in primate frontal eye field. *J Neurophysiol* 100:796–814.
27. Schiller PH (1996) On the specificity of neurons and visual areas. *Behav Brain Res* 76: 21–35.
28. Singer W (2013) Cortical dynamics revisited. *Trends Cogn Sci* 17:616–626.
29. Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex* 1:1–47.
30. Crowe DA, et al. (2013) Prefrontal neurons transmit signals to parietal neurons that reflect executive control of cognition. *Nat Neurosci* 16:1484–1491.
31. Antzoulatos EG, Miller EK (2016) Synchronous beta rhythms of frontoparietal networks support only behaviorally relevant representations. *eLife* 5:e17822.
32. Warden MR, Miller EK (2010) Task-dependent changes in short-term memory in the prefrontal cortex. *J Neurosci* 30:15801–15810.
33. Fusi S, Miller EK, Rigotti M (2016) Why neurons mix: High dimensionality for higher cognition. *Curr Opin Neurobiol* 37:66–74.
34. Miller EK, Cohen JD (2001) An integrative theory of prefrontal cortex function. *Annu Rev Neurosci* 24:167–202.