PNAS
www.pnas.org

Supplementary Information for

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

Scott L. Brincat*, Markus Siegel*, Constantin von Nicolai, Earl K. Miller

Earl K. Miller
Email: ekmiller@mit.edu

This PDF file includes:

Supplementary text
Figs. S1 to S7
Table S1
References for SI reference citations

www.pnas.org/cgi/doi/10.1073/pnas.1717075115

Supplementary Information Text

SI Methods

Electrophysiological data collection. Parylene-coated tungsten electrodes were acutely inserted into target brain regions each day using a manual microdrive system. Neuronal activity was referenced to animal ground, amplified by a high-impedance headstage, filtered to extract spiking activity, digitized, and streamed to disk by an integrated electrophysiological system (Recorder or Omniplex, Plexon). The filtered signal was threshold-triggered to separate neuronal spikes from background noise, and individual spike waveforms were manually sorted offline into isolated neurons (Offline Sorter 3, Plexon). Neurons were included in analyses only for the duration of time in which they were well isolated from background noise and other neurons.

Behavioral paradigm. Two adult rhesus macaques (*Macaca mulatta*), one male and one female, were trained to perform a cued multidimensional categorization task. All stimuli were displayed on a color-calibrated CRT monitor at 100 Hz vertical refresh rate. An infrared-based eye-tracking system (Eyelink II, SR Research) continuously monitored eye position at 240 Hz. Behavioral control was handled by the MonkeyLogic program (www.monkeylogic.net) (1).

Each trial (Fig. 1A) was initiated when the animal fixated on a central dot (± 1.2° visual angle). Following a 500 ms fixation period, a centrally-presented visual cue (1000 ms) instructed the monkey to perform one of two task rules: color categorization ("greenish" vs. "reddish") or motion categorization ("upward" vs. "downward") of a subsequent centrally-presented colored, moving random-dot stimulus. Within each stimulus, all dots had the same color and moved in the same direction (100% coherence). The monkeys reported the stimulus category via a saccade towards a target to the left or right. The stimulus-response mapping for each task rule was fixed (color rule: greenish→left, reddish→right; motion rule: upward→left, downward→right). The monkeys were free to respond at any point (up to 3 s) after the random-dot stimulus onset.

The cues were four different gray shapes; two of them instructed the color rule, while the other two instructed the motion rule (Fig. 1B). Using two distinct cues for each rule allowed dissociation of neural activity related to the visual cue shapes and the task they instructed. Each motion category consisted of two directions (Fig. 1C; upward: 90°, 30°; downward: -30°, -90°), and each color category consisted of two hues (greenish: 90°, 30°; reddish: -30°, -90° hue angle; colors were defined in CIE L*a*b* space and had the same luminance and saturation). Additional tested values on one or more category boundaries were not included in the present analyses, as we focus here on unambiguously classifiable stimuli.

The monkeys performed both categorization tasks with high accuracy: 94% correct on average for motion, 89% for color. Across all trials and sessions in the current dataset, the

median reaction time was 274 ms, with only 1% of all responses occurring earlier than 200 ms.

General data analysis. For most analyses, spike trains were converted into smoothed rates via convolution with a Hann function (half-width 125 ms; nearly equivalent to a 50 ms SD Gaussian, but with finite extent). To summarize results, we pooled rates or other derived statistics within empirically-defined epochs of interest for each task domain and area. The start of each epoch was set by the onset latency of total explained variance for a given task variable and area (see "Variance-partitioning model" section below), as estimated by the time point where the population total variance first attained 25% of its maximal value across all time points. To capture the full temporal extent of relevant neural activity, without introducing confounding factors, the ends of cue/task epochs for all areas were set at the onset of the random-dot stimulus, and the ends of motion- and color-related epochs were set at the earliest 1 percentile of the distribution of all behavioral reaction times (200 ms). All analyses were performed using custom Matlab code.

Hypothesis testing. All hypothesis tests used non-parametric bootstrap methods that do not rely on specific assumptions about the distributions of data values, unless otherwise noted. One-sample tests (Fig. 3B,E; 4B,E; 5B,E) resampled the statistic of interest with replacement 10,000 times from the neuronal population. The resulting distribution, offset by the actual observed statistic value, is an empirical estimate of the distribution of the statistic under the null hypothesis, and the one-tailed significance level was computed as the proportion of resampled values greater than the actual observed value. Two-sample tests (Fig. 3C,F; 4C,F; 5C,F) were computed similarly, but with separate bootstrap resampling from each compared population, evaluating the difference in the statistic of interest between populations, and computing the two-tailed significance level as the proportion of absolute resampled values exceeding the absolute observed value. All presented standard errors were estimated as the 68% confidence intervals across bootstrap samples.

Variance-partitioning model. Our primary goal was to characterize the categoricality of the population of neurons in each cortical region—the degree to which they reflected the raw sensory stimuli or their abstracted categorical meaning. We quantified this by fitting each neuron's spike rates, at each time point, with a linear model that partitioned rate variance across n trials into between-category and within-category effects:

$$\begin{aligned}\mathbf{rate} = {} & \beta_{\mathrm{cue1}}\mathbf{x}_{\mathrm{cue1}} + \beta_{\mathrm{cue2}}\mathbf{x}_{\mathrm{cue2}} + \beta_{\mathrm{cue3}}\mathbf{x}_{\mathrm{cue3}} \\ & + \beta_{\mathrm{motion1}}\mathbf{x}_{\mathrm{motion1}} + \beta_{\mathrm{motion2}}\mathbf{x}_{\mathrm{motion2}} + \beta_{\mathrm{motion3}}\mathbf{x}_{\mathrm{motion3}} \\ & + \beta_{\mathrm{color1}}\mathbf{x}_{\mathrm{color1}} + \beta_{\mathrm{color2}}\mathbf{x}_{\mathrm{color2}} + \beta_{\mathrm{color3}}\mathbf{x}_{\mathrm{color3}} \\ & + \beta_{\mathrm{choice}}\mathbf{x}_{\mathrm{choice}} + \beta_0\mathbf{1}\end{aligned}$$

In this equation, $\mathbf{rate} \in \mathbb{R}^{n\times 1}$ is a trial-length vector of spike rates for a given neuron and time point. For each task domain (rule cues, motions, and colors), the model included three orthogonal contrast terms (Fig. 2A). The first term $\mathbf{x}_{*1} \in \{-1,1\}^{n\times 1}$ (where * indicates cue, motion, or color) contrasted the two stimulus items (cue shapes, motion directions, or colors) in one categorical grouping (rule, motion category, or color category) against the two items in the other categorical grouping ([A or B] – [C or D] in Fig. 2A). That is, it was a length n vector with a value of 1 for each trial where the

stimulus item was A or B, and –1 for each trial where the stimulus was C or D. Thus, it reflected the actual task-relevant grouping of stimulus items into categories, and its associated fitted scalar coefficient $\beta_{*1} \in \mathbb{R}$ captured between-category variance. The other terms $\mathbf{x}_{*2}, \mathbf{x}_{*3} \in \{-1,1\}^{n\times 1}$ were contrasts between the two other possible, non-task-relevant paired groupings of items ([A or C] – [B or D] and [A or D] – [B or C] in Fig. 2A). Thus, their associated coefficients $\beta_{*2}, \beta_{*3} \in \mathbb{R}$ together captured within-category variance. The model also includes a term described below that accounted for behavioral choice ($\mathbf{x}_{\text{choice}}$), and a constant term $\mathbf{1} \in \{1\}^{n\times 1}$ (i.e. an all-ones vector) and associated coefficient (β_0) to account for the overall condition-independent mean rate. The effect of each model term was quantified in terms of its percent of total data variance explained, measured from the difference in residual variance between the full model and a reduced model with the given term deleted (2). Note that this is an extension of the analysis strategy used to model task rule coding in our previous publication (3) to all three categorical domains in this task.

Across-trial mean neural responses for the four stimulus items in each domain (task cues, motions, or colors), for each neuron and time point, can be considered as four-dimensional vectors within the four-dimensional vector space ($\mathbb{R}^4$) of all possible sets of item responses. If the three contrast terms described above are also considered as four-dimensional vectors (i.e. for items [A, B, C, D], $\mathbf{x}_{*1}$ would be [1 1 -1 -1], $\mathbf{x}_{*2}$ would be [1 -1 1 -1], and $\mathbf{x}_{*3}$ would be [1 -1 -1 1]), then along with the constant term ([1 1 1 1]), they constitute an orthogonal basis for this space. (This follows from the fact that they are mutually orthogonal and contain the same number of vectors as the dimension of the space (4)). This entails that any possible set of neural responses to the four items can be expressed as a linear combination of these terms. Hence, they capture all variance reflecting cue, motion, or color selectivity in the task. Since the first term ($\mathbf{x}_{*1}$) perfectly captures between-category variance, it follows that the other terms together capture all within-category variance. Thus, the three contrast terms partition the total variance in each task domain into between-category and within-category effects.

Within the context of each task rule, motion and color categories were—by design—inextricably linked with the monkey's behavioral choice (e.g., for the color rule, greenish and reddish colors always mandated leftward and rightward saccades, respectively). For this reason, it is effectively impossible to dissociate category and choice effects within the context of each individual task rule, and to directly compare category effects across the two rules. When both rules are considered together, this identity link is broken, permitting dissociation of choice and category effects. However, there remains a partial (50%) correlation between choice and stimulus category. Therefore, it was critical to ensure that any activity reflecting choice (or subsequent motor preparation processes) was not spuriously interpreted as categorical coding. To partition out choice effects, we also included an additional model term $\mathbf{x}_{\text{choice}} \in \{-1,1\}^{n\times 1}$ contrasting the two possible behavioral choices, with a value of 1 for rightward and –1 for leftward saccades. Its associated coefficient β_{choice} accounted for any variance due to choice, and thus category effects are measured in terms of their additional variance explained once choice effects are already accounted for (2). Choice effects were analyzed in detail in our previous publication on this dataset (3), and were therefore not examined further in this work.

Each model, 11 parameters in total, was fit via ordinary least squares separately for each neuron and time point. We used the bias-corrected ω^2 formulation for explained variance (5).

Categoricality index. For each task variable (task cues, motions, or colors), the sum of explained variances for all model three terms sets an upper bound for the between-category variance alone—they can be equal only for a perfectly categorical population with zero within-category variance (Fig. 2B, top; simulated in Fig. S1A, right). A purely sensory-driven population would instead have equal variance for all three terms. For example, a neuron with a preferred response for only a single stimulus item would have equal variance for all three contrasts, as each item appears once in every contrast (simulated in Fig. S1A, left). A population of neurons with preferred responses for arbitrary pairs of stimulus items—without regard to the task-relevant groupings—would also have equal expected variance for all three contrasts (simulated in Fig. S1A, center). Thus, the mean of all three contrasts provides an expected lower bound for the between-category variance in purely sensory populations (Fig. 2B, bottom). (Note that, for populations specifically encoding within-category differences, between-category variance can be less than the mean within-category variance.)

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. 2C).

$$\text{categoricalityIndex} = \frac{\int_t \left(\langle \text{betweenCategory} \rangle_{pop} - \langle \text{sensory} \rangle_{pop} \right)}{\int_t \left(\langle \text{total} \rangle_{pop} - \langle \text{sensory} \rangle_{pop} \right)}$$

Here, $\text{betweenCategory} = v_1$ is the between-category (term 1) explained variance; $\text{total} = v_1 + v_2 + v_3$ is the sum of explained variances for all three terms; and $\text{sensory} = (1/3)(v_1 + v_2 + v_3)$ is the mean of explained variances across all three terms. Each quantity is computed separately for each neuron and time point. The means $\langle \cdot \rangle_{pop}$ are over all sampled neurons in each area's population, and the integrals $\int_t (\cdot)$ are over all time points within a pre-defined epoch for each area and task domain. This index is a specific measure of how categorical a neural population is, and ranges from 1 for a perfectly categorical 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.)

In addition to the population categoricality index described above, we also computed indices for each individual neuron (Fig. S3). Here, all of the population means in the above equation were simply replaced by the corresponding quantities for each neuron. This analysis was restricted to neurons found to have significant total domain variance ($p < 0.01$; *F*-test), since the index is undefined—and the concept of categoricality is not meaningful—in the absence of any rate variance across the constituent items.

The statistic of category information used in our previous publication (3) was a debiased version of the between-category variance, equal to the between-category variance (v_1) minus the mean within-category variance ($(1/2)(v_2 + v_3)$). It can be shown that our

categoricality index is equivalent to normalizing this statistic by the total domain variance $(v_1 + v_2 + v_3)$:

$$\frac{\text{betweenCategory} - \text{sensory}}{\text{total} - \text{sensory}} = \frac{v_1 - (1/3)(v_1 + v_2 + v_3)}{(v_1 + v_2 + v_3) - (1/3)(v_1 + v_2 + v_3)}$$
$$= \frac{3v_1 - (v_1 + v_2 + v_3)}{3(v_1 + v_2 + v_3) - (v_1 + v_2 + v_3)}$$
$$= \frac{2v_1 - v_2 - v_3}{2v_1 + 2v_2 + 2v_3}$$
$$= \frac{v_1 - (1/2)(v_2 + v_3)}{v_1 + v_2 + v_3}$$

Thus, our index can also be viewed as a normalized contrast of between-category and within-category variance.

Neural simulations. To validate our analysis, we measured its properties on synthesized neural activity, where different coding attributes could be manipulated under experimental control. Note that the goal here is only to ensure the analysis behaves as expected in a simple, plausible toy model where ground truth is known, not to make any claims that our simulations accurately reflect actual cortical coding. We simulated populations of 200 neurons with Poisson rates modulated by both selectivity for categories and sensory tuning for stimulus items. Category selectivity in all task domains was simulated by a binary step function. A preferred Poisson rate was set for all items in one category, and a non-preferred rate for all items in the other category, with the preferred category randomized across simulated units. For task cues, stimulus tuning was simulated by setting a preferred rate for one or two cues randomly selected for each unit, without regard to categorical divisions. As a simple, plausible model of tuning for stimulus motion or color, we simulated each unit with a von Mises (circular Gaussian) tuning curve, with its center randomly oriented within the full range (0–360°) of motion directions or hue angles. In some simulations, we also modelled binary selectivity for behavioral choice (leftward vs. rightward saccades), with the preferred choice randomized across units.

Simulations evaluated four distinct measures of categorical coding: the normalized categoricality index introduced here (Fig. S1A–D), raw between-category variance (Fig. S1E,F), the debiased between-category variance statistic used in our previous publication (3) (Fig. S1G,H), and the category preference index (Fig. S1I,J).

In reported simulations, we expressed the tested selectivity in all population units, but manipulated its strength (the difference in Poisson rate between preferred and non-preferred items). Similar results were obtained if strength was fixed and the proportion of units expressing the tested selectivity was manipulated instead. All of the presented results also generalize well from the specific population size, tuning functions, proportion of units expressing selectivity, and response model assumed in these simulations. We evaluated simulated populations under the same task structure and analysis as the real data, and report means and standard deviations across 100 independent simulations.

These results are presented in the “Neural simulations” section of SI Results.

Category/choice consistency analysis. As an additional control to show our results are not confounded by behavioral choice, we compared color and motion category preferences under the two task rules, where category and choice coding make clear, opposing predictions. We fit models containing only the motion and color between-category terms, separately for the motion and color rules (variables here have the same interpretation as in equation 1):

$$\mathbf{rate} = \beta_{\mathrm{motion1}}\mathbf{x}_{\mathrm{motion1}} + \beta_{\mathrm{color1}}\mathbf{x}_{\mathrm{color1}} + \beta_0\mathbf{1}$$

Two models were fit to each neuron's spike rate pooled over the random-dot stimulus epoch. One model was fit to trials where the motion rule was instructed, another to trials where the color rule was instructed. For each term and task rule, we computed explained variance as above, but with a sign reflecting the preferred category: negative for neurons preferring downward motion or reddish color, and positive for upward motion or greenish color. Note that when a given task domain is task-relevant (i.e. motion during the motion rule, or color during the color rule), labeling by preferred category is equivalent to labeling by preferred choice (see "Variance-partitioning model" section above for details).

Choice coding predicts consistent motion and color category preferences (signs and magnitudes) when they are each task-relevant based on the currently instructed rule (i.e. when they both map to the same choice and resulting saccade direction). Category coding instead predicts consistent motion category preferences across both task rules, and the same for color category preferences. Consistency is visualized with scatterplots and quantified with Spearman rank correlation for each of these conditions. These results are presented in the "Category/choice consistency analysis" section of SI Results.

Preferred condition analysis. To qualitatively confirm our results using an alternative, more standard method of measuring category effects, we computed population average spike rates for preferred and non-preferred categories and stimulus items within each category. For each neuron and task variable, the preferred and non-preferred category (task rule, motion category, or color category) and the preferred and non-preferred constituent stimulus item within each category (rule cue, motion direction, or color) were determined from the average spike rate within the relevant time epoch for each area and task variable. Spike rate time series were averaged across the population of neurons in each area separately within each of the resulting four sorted conditions (preferred item in preferred category, non-preferred item in preferred category, preferred item in non-preferred category, non-preferred item in non-preferred category). To avoid circularity in this analysis, we used a condition-balanced ten-fold cross-validation method. Preferred conditions were estimated from 90% of trials and the resulting sorting was used for population averaging of rate series across the remaining 10% of trials, with each partition having approximately balanced trial numbers across the four conditions. This was performed for each disjoint subset of 10% of trials, and the final presented results are the average of these ten cross-validation folds.

In this analysis, the signature of categorical coding is a large difference in rate between the preferred and non-preferred category, with minimal rate differences between items within each category. For sensory representations, rate differences between and within categories would be of similar magnitude. Therefore, to summarize these results we

computed a category preference index, a contrast index comparing differences in preference-sorted mean rates between categories vs. within categories.

$$\text{preferredCondIndex} = \frac{\int_t \left(\langle d_{between} \rangle_{pop} - \langle d_{within} \rangle_{pop} \right)}{\int_t \left(\langle d_{between} \rangle_{pop} + \langle d_{within} \rangle_{pop} \right)}$$

In this equation, $d_{between}$ is the difference in rate between the preferred category and the non-preferred category for each neuron,

$$d_{between} = \left(\text{rate}_{prefCtg,prefItem} + \text{rate}_{prefCtg,nonprefItem)} \right)/2 - \left(\text{rate}_{nonprefCtg,prefItem} + \text{rate}_{nonprefCtg,nonprefItem)} \right)/2$$

and d_{within} is the difference in rate between the preferred item and the non-preferred item within each category, averaged across both categories, for each neuron.

$$d_{within} = \left[\left(\text{rate}_{prefCtg,prefItem} - \text{rate}_{prefCtg,nonprefItem)} \right) + \left(\text{rate}_{nonprefCtg,prefItem} - \text{rate}_{nonprefCtg,nonprefItem)} \right) \right]/2$$

Both the between- and within-category differences were rectified at zero, as negative rate differences have little meaning in a preference-sorted analysis (these were rare, but possible because of the cross-validation procedure). These results are presented in the "Preferred condition analysis" section of SI Results.

Population dimensionality analysis. To measure the dimensionality of population activity, we adapted a method developed by Machens and colleagues (6), and extended by Rigotti and colleagues (7). This method estimates the dimensionality of the space spanned by population activity vectors by determining the number of principal components required to describe them. Spike rates were computed within each of the time epochs enumerated above, and the resulting data for each area was pooled across all electrodes and recording sessions into "pseudo-populations". To extrapolate results to larger neural populations, we generated artificial neurons via specific relabeling of conditions, under the assumption that the full underlying population likely includes neurons with similar activity distributions, but slight differences in tuning for rule cues, motion directions, and colors (7). Labels were swapped between the two visual cues associated with each task, between the two colors within each color category, and/or between the two directions within each motion category. Note that these label reassignments both maintain the semantic logic of the paradigm and would not alter the information carried by either sensory or categorical neurons, as estimated above. Artificial neural populations were generated by all permutations of performing each of these relabeling operations or not, resulting in a 64-fold multiplication of the actual populations (2^2 colors × 2^2 directions × 2^2 cues). Extrapolated populations of different sizes were generated by randomly subsampling from this full population of real and artificial neurons, separately for each cortical area. Dimensionality values obtained from analyzing only the actual recorded populations (Fig. 6A,C, squares) generally fell within the distribution of those obtained from the extrapolated populations, suggesting that this procedure did not substantially alter the results.

For this analysis, we characterized neural activity within the full space of 64 task conditions (4 rule cues × 4 motion directions × 4 colors) and within the random-dot stimulus epoch, when all 64 conditions were differentiated. An additional analysis was

performed within the reduced space of 16 motion directions × colors. To avoid PCA being dominated by a few highly active neurons, trial spike rates for each neuron were preprocessed by z-scoring them by the mean and SD across all trials, with a regularization constant of 0.1 added to each SD (7). For each area and extrapolated population size, the vector of mean spike rates across all neurons for each condition was computed, and the dimensionality of the space spanned by each resulting set of 64 neural population activity vectors was estimated as the number of their principal components (eigenvalues) significantly greater than the estimated distribution of principal components due to finite sampling noise. Noise was estimated by computing the difference between randomly chosen trials within each condition, weighted by its expected contribution to the estimated condition mean (6), and submitting the resulting set of 64 noise vectors to PCA. This procedure was repeated 1000 times, randomly sampling within-condition trial pairs each iteration, to generate a distribution of noise-derived eigenvalues. Dimensionality was estimated as the number of data eigenvalues exceeding 95% of the distribution of first (largest) eigenvalues computed from the sampled noise (7). This entire procedure was repeated 50 times, selecting with replacement a different random subsample of area neurons in each iteration, to estimate bootstrap standard errors of the resulting dimensionality statistic. Qualitatively similar results were obtained using the method of Rigotti et al. 2013 based on the number of binary classifications that can be successfully decoded from population activity vectors.

Cortical organization analysis. To quantitatively relate our results to classical models of large-scale cortical organization, we separately fit the 18 population summary values (6 areas × 3 task variables) of total domain explained variance (i.e., sensory information; Fig. 7A) and categoricality index (Fig. 7B) with a simple two-predictor linear model:

$$\begin{aligned}\mathbf{summaryValue} &= \beta_{hierarchicalLevel}\mathbf{x}_{hierarchicalLevel} \\ &+ \beta_{streamCongruence}\mathbf{x}_{streamCongruence} + \beta_0\end{aligned}$$

Here, $\mathbf{summaryValue} \in \mathbb{R}^{18\times1}$ is the vector of 18 summary values—total domain variances or categoricality indices. The first predictor $\mathbf{x}_{hierarchicalLevel} \in \mathbb{R}^{18\times1}$ was the Felleman and Van Essen hierarchical level of each area (MT,V4: 5; PIT,LIP: 7; FEF: 8; PFC: 10), estimated from the relative distribution of feedforward-type vs. feedback-type anatomical connections between areas (8). Its associated coefficient $\beta_{hierarchicalLevel}$ accounts for any effects of each area's hierarchical level on the given summary value. The second predictor $\mathbf{x}_{streamCongruence} \in \{-1,0,1\}^{18\times1}$ reflected the expected functional congruence of each combination of task variable and area based on the classical divisions of visual areas and functions of the ventral and dorsal processing streams. It took a value of 1 for consistent combinations (e.g. MT and motion, V4 and color), and –1 for inconsistent ones (e.g. MT and color, V4 and motion). For this purpose, rule cue (shape) and color were designated as ventral stream domains, and V4 and PIT ventral stream areas. Motion direction was designated a dorsal stream domain, and MT, LIP, and FEF dorsal stream areas. Since PFC is thought to integrate across both streams (8, 9), all variable/area combinations involving it were set to zero for this predictor. To equate the range of stimulus information and categoricality index values across task variables, values across all areas for each variable were normalized to range from 0 to 1. Prior to this normalization, categorical information values were rectified at zero, thus equating

areas that match the sensory-based prediction with those less than it (reflecting slight biases toward within-category selectivity). Each model (3 parameters in total) was fit via ordinary least squares, and evaluated by the variance explained by each of the two predictors. Errors and significance levels were estimated by bootstrapping these statistics over the contributing neurons 10,000 times.

SI Results

Neural simulations. To confirm that our analysis methods performed as expected, we first assayed their properties on synthesized neural activity with known ground truth (see "Neural simulations" section in SI Methods for details). The basic properties of all analyses considered here are summarized in Table S1. We first confirmed the categoricality index was ≈ 1 for populations of units with ideal binary category selectivity for task rules (Fig. S1A, right), and was ≈ 0 for populations with only sensory selectivity for single task cues (Fig. S1A, left) or for random pairs of cues (Fig. S1A, center). To examine this in more detail, we parametrically varied both the relative strength of simulated sensory and categorical signals (Fig. S1B–C, *x*-axis), and the overall strength of both signals in concert (i.e. simulated differences in spike rates between preferred and non-preferred conditions; Fig. S1B–C, color saturation). For task rule coding, we simulated sensory selectivity for random pairs of cues (Fig. S1B; results are similar for single-cue selectivity). For motion coding, we simulated Gaussian sensory tuning for motion direction (Fig. S1C; note that color and motion are treated identically in these simulations and results therefore generalize to both domains). These simulations confirmed that our categoricality index reliably reflected sensory/category relative strength (curves monotonically increase from left to right), but was relatively invariant to overall signal strength (different-color curves largely overlap). Finally, to confirm that our modeling procedure successfully partitioned category and choice effects, we simulated populations independently varying in both the relative weight of sensory and categorical motion signals (Fig. S1D, *x*-axis), and the strength of choice signals (Fig. S1D, color saturation). These simulations confirmed the categoricality index was also relatively insensitive to choice signal strength. As a positive control, we demonstrated that the variance explained by choice itself did indeed robustly reflect choice signal strength (Fig. S1D, inset).

Raw between-category variance, in contrast, provides a biased and inconsistent measure of categoricality. As expected, between-category variance was high for a simulated population with ideal binary category selectivity for task rules (Fig. S1E, right). However, it also had non-zero values for populations with sensory selectivity for single task cues (Fig. S1E, left) or for random pairs of cues (Fig. S1E, center). Figure S1F shows that between-category variance conflates sensory/categorical relative strength with overall spike rates. Even purely sensory populations (Fig. S1F, leftmost points) have non-zero values that increase proportionately to total signal strength. Results are similar for simulated sensory selectivity for single task cues, and for motion direction and color. Thus, the raw percent of variance explained by categories is an inadequate measure of population categoricality.

In a previous publication, we proposed a debiased between-category variance statistic of categorical task rule information (3). This measure subtracts within-category variance from the raw between-category variance (see "Categoricality index" section in SI Methods for details), resulting in an expected value of zero for purely sensory representations. This was confirmed using the same set of simulations as above (gray bars in Fig. S1G; leftmost points in Fig. S1H). However, for populations with a mixture of sensory and categorical coding (right portion of Fig. S1H), this statistic is influenced by the overall signal strength. For any given value of sensory/categorical relative weight (*x*-axis value) it is proportional to overall rate, and any given value of this statistic (*y*-axis value) might correspond to a wide range of combinations of categoricality and spike rate. Similar results obtained for other sensory domains. Thus, while this statistic provides a measure of category information that is unbiased for purely sensory activity, it does not in general unambiguously report population categoricality, as defined here.

As an alternative method to measure population categoricality, we computed population average spike rates for preferred and non-preferred categories and within-category stimulus items. We summarized this analysis with a category preference index, a contrast index contrasting between-category and within-category differences in the preference-sorted mean spike rates (see "Preferred condition analysis" section in SI Methods for details). Like the categoricality index, this statistic reports values ≤ 0 for sensory populations (Fig. S1I, left and center) and ~ 1 for categorical populations (Fig. S1I, right), and is relatively invariant to overall signal strength (Fig. S1J).

Category/choice consistency analysis. As an additional control to show our results are not confounded by behavioral choice, we compared color and motion category preferences under the two task rules, where category and choice coding make clear, opposing predictions. Category preference was quantified by the between-category motor or color variance, with a sign reflecting the preferred category: negative for downward or reddish preferring, positive for upward or greenish preferring. Consider the comparison between motion category preference during the motion rule vs. color category preference during the color rule, i.e. when each category is task-relevant and drives behavioral choice (Fig. S5A and S5D–E, left). A neuron encoding choice would appear to have the same sign and magnitude of category preference for both conditions, as these would map to the same choice (saccade direction). A population dominated by choice selectivity would thus be expected to have a strong consistency of signed variance between these conditions, and we therefore refer to this comparison as "choice-consistent". In contrast, category-selective populations would be expected to either carry information about only a single domain or have unrelated preferences across category domains, and thus have little consistency between these conditions.

Now consider instead the comparison between motion category preference during the motion rule vs. during the color rule (Fig. S5B; Fig. S5D–E, center), or between color category preference during the motion rule vs. during the color rule (Fig. S5C; Fig. S5D–E, right). A neuron encoding a given category across both rules would have the same sign and magnitude of preference in both conditions, and a population dominated by category selectivity would thus show strong consistency between these conditions. We therefore refer to these comparisons as "category-consistent". In contrast, a choice-selective neuron would appear to carry category information only when the category was task-relevant,

and a choice-dominated population would thus have inconsistent category preferences across task rules.

To determine the type of coding that is dominant overall in each area, we plotted scatterplots and computed Spearman rank correlations across the entire sampled neural populations for each comparison above. For the choice-consistent comparisons, the scatterplots showed no obvious relationship between motion and color category preference (Fig. S5A), and their correlations were weak (Fig. S5D, left) and only significant for V4 ($p = 0.002$; $p \geq 0.01$ for all other areas, permutation test). In contrast, the category-consistent comparisons had clear structure (Fig. S5B,C) and stronger correlations (Fig. S5D, center and right), which were significant for at least one category domain in all areas ($p < 0.01$, Bonferroni-corrected for two comparisons). This suggests that category signals are generally dominant over choice signals in the studied areas. Note that PFC, FEF, and to some extend LIP did have choice-consistent and category-consistent correlations of similar magnitude. This suggests some heterogeneous mixture of category and choice effects in these populations, consistent with previous reports showing strong choice and saccade signals in these areas (3, 10, 11).

As a final check that our main analysis specifically extracts category effects from these population mixtures, we repeated the correlation analysis for only the subset of neurons carrying significant category information across the full set of trials ($p < 0.01$, F-test on between-category variance for equation 1 model fit to stimulus epoch rates across both rules; Fig. S5A–C, colored dots). There is some circularity in this analysis, in that the criterion used to select neurons (significant category variance across all trials) is not entirely independent of the measure examined (signed category variance under each rule). However, we felt it was helpful to include these results to provide a complete picture of the data and analysis. For the subset of category-selective neurons, choice-consistent correlations remained weak (Fig. S5E, left) and were not significant any area ($p \geq 0.01$), while correlations for the category-consistent conditions were uniformly strong (Fig. S5E, center and right). Together, these results provide another line of evidence that our main analysis correctly measured category signals, rather than confounded choice signals.

Preferred condition analysis. To confirm our results using an alternative method of measuring category effects, we computed population average spike rates for preferred and non-preferred categories and stimulus items within each category (Fig. S6). These results were summarized with a category preference index, a contrast index comparing between-category to within-category differences in mean rates (see "Preferred condition analysis" section in SI Methods for details). The results indicated broad agreement with the categoricality index in the main text. For task cue coding in PFC and FEF, rate differences between the preferred and non-preferred task rules were considerably larger than between cues instructing the same rule (Fig. S6A), and their category preference indices were significantly greater than zero (Fig. S6B). LIP trended in the same direction, but unlike the main text analysis (Fig. 3E in the main text), it was not significant. For MT, V4, and PIT, between- and within-category differences were similar (Fig. S6A), and their indices were not significantly different from zero (Fig. S6B). For motion coding, only PFC, FEF, and LIP showed significant category preference indices (Fig. S6E). Again, this result was overall similar to the main text categoricality results, though in that case FEF failed to reach significance (Fig. 4E in main text). As for the categoricality

index analysis (main text Fig. 5E), color coding exhibited a more categorical representation overall, with all areas except MT having significant category preference indices (Fig. S6H). Note that perfect alignment between results from these two methods would not be expected, due to their many salient differences (Table S1). In the preferred condition analysis, choice effects were not partitioned out, rate was not normalized (and thus high-rate neurons might contribute disproportionately), and overall domain variance was not normalized out. The fact that results were, nevertheless, generally quite similar with an arguably more standard method adds support to our main conclusions.

Comparison with our previous results. Some of our results appear to be somewhat at odds with our prior publication on this dataset (3). In particular, it was previously claimed that V4 and PIT contained strong task rule signals (Fig. 2C in (3)), whereas we claim no significant task rule categoricality in these same areas (Fig. 3D,E). In part, this is due to differences in the studied neural signals—we analyzed isolated single neurons, whereas the previous study analyzed multi-unit signals. We repeated our analysis on multi-unit signals similar to those used previously (pooling together all threshold-crossing spikes on each electrode). Though results were generally quite similar (Fig. S4), multi-units produced a slightly more categorical rule cue representation in V4 and PIT (Fig. S4A), perhaps suggesting stronger signals for task rules exist in the smaller neurons likely to appear only in multi-unit signals.

The primary difference, though, is in the specific questions addressed by each study and their resulting analytical strategies. Our previous study addressed the overall task rule information conveyed by each neural population. Thus, it used a debiased category variance statistic, equal to the difference between between-category variance and the mean within-category variance. This statistic is unbiased for purely sensory representations—it has an expected value of zero (simulated in Fig. S1G). But for populations with any mixture of category effects, it reflects both categoricality and total domain variance (simulated in Fig. S1H). Thus, it can be high for populations—like rule cue representations in V4 and PIT—with strong total domain variance, even if only a very small fraction of that variance is categorical. As expected, recomputing our main findings with this statistic produced quite different results from ours (Fig. S7), which qualitatively appear to look like a mixture of categoricality (Fig. 3E,4E,5E) and total domain variance (Fig. 3B,4B,5B). Here, we instead addressed how *categorical* neural representations are, independent of their overall information. Our categoricality index measures this (Fig. S1B,C) by normalizing total domain variance out of the above statistic. Under this statistic, V4 and PIT contain task cue representations that are only weakly categorical (Fig. 3D,E). 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 (12, 13).

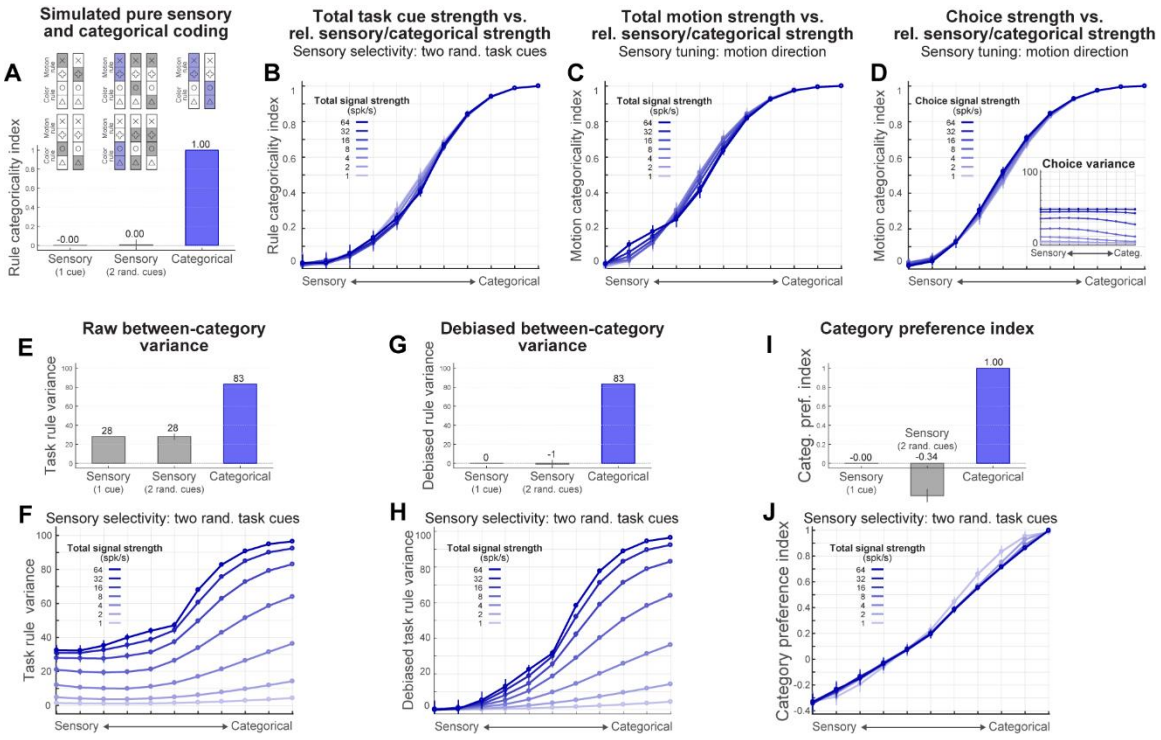

Fig. S1. Validation of analysis methods with neural simulations. (A) Mean (± SD across independent simulations) categoricality index computed on simulated neural populations. Diagrams at top reflect simulated patterns of neural activity across the four task cues—white indicates non-preferred responses, blue indicates preferred responses consistent with task rule coding, and gray indicates preferred responses inconsistent with task rules. Results for three simulated populations are shown: one with sensory selectivity for single task cues (left), one with sensory selectivity for random pairs of cues (center), and one with pure categorical selectivity for pairs of cues that map to the same task rule (right). The categoricality index veridically reports values of ~0 for the two sensory populations and ~1 for the categorical population. (B–C) Mean (± SD) categoricality index for simulated populations parametrically varying in the relative strength of sensory and categorical signals (*x*-axis) and in the overall strength of both signals in concert (rate difference between preferred and non-preferred conditions; color saturation of plots). Results are shown for simulations with sensory selectivity for random pairs of task cues (B), and sensory tuning for motion direction (C). In all cases, index values reliably reflect sensory/categorical relative strength (curves monotonically increase from left to right), but are relatively insensitive to overall signal strength (different-color curves are mostly overlapping). (D) Mean (± SD) categoricality index for simulated populations parametrically varying in the relative strength of sensory and categorical motion signals (*x*-axis) and in the strength of signals for behavioral choice (rate difference between left and right saccades; color saturation of plots). Index values are not confounded by choice effects. Inset: The variance attributed to choice does reliably track choice signals, as expected. (E–F) Mean (± SD) raw task-rule between-category variance, under the same simulations as in A and B. Raw between-category variance has non-zero values for purely sensory populations (E, left and center) and is proportional to total signal strength (vertical shift of curves in F), indicating it is a biased measure of categoricality, which is confounded with global changes in spike rates. (G–H) Mean (± SD) debiased task-rule between-category variance, under the same simulations as in A and B. This is the statistic used to measure task rule information in our previous publication from this dataset (3), and is equal to between-

category variance minus the mean within-category variance. This statistic is ~0 for purely sensory populations (G, left and center), indicating that it successfully removed the bias in the raw variance. However, for populations with a mixture of category effects, it reflects both categoricality and overall spike rates (vertical offset of curves in H). (I–J) Mean (± SD) task-rule category preference index, under the same simulations as in A and B. Like the categoricality index, this statistic reports values ≤ 0 for sensory populations (I, left and center) and ~ 1 for categorical populations (I, right), and is relatively invariant to overall signal strength (J).

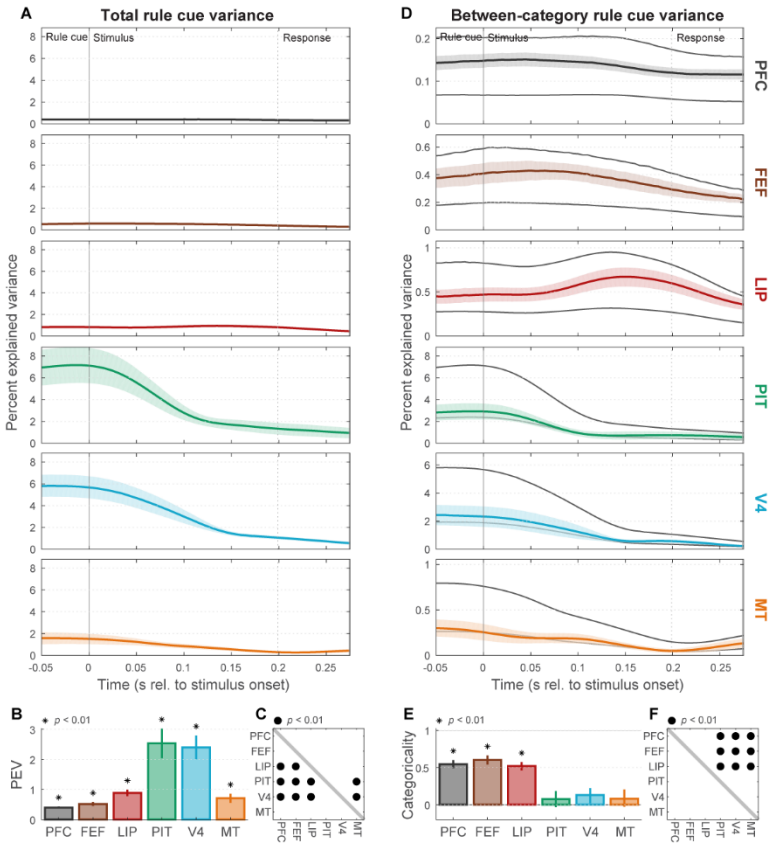

Fig. S2. Task rule cue coding during stimulus period. (A) Population mean (± SEM) total rule cue variance (cf. Fig. 3A) during the random-dot stimulus period. Cue information in visual areas V4, PIT, and MT drops toward baseline soon after cue offset. (B) Summary (across-time mean ± SEM) of total stimulus-period rule cue variance for each area (cf. Fig. 3B). All areas retain significant cue information during the stimulus period ($p < 0.01$). (C) Indicates which regions (rows) had significantly greater cue information than others during the stimulus period (cf. Fig. 3C). (D) Mean (± SEM) between-category rule cue variance (task rule information). (E) Stimulus-period task rule categoricality index (± SEM) for each area (cf. Fig. 3E). PFC, FEF, and LIP remain significantly categorical ($p < 0.01$)—they continue to convey task rule information through the stimulus period. (F) Indicates which regions (rows) had significantly greater task rule categoricality indices than others (columns; $p < 0.01$; cf. Fig. 3F).

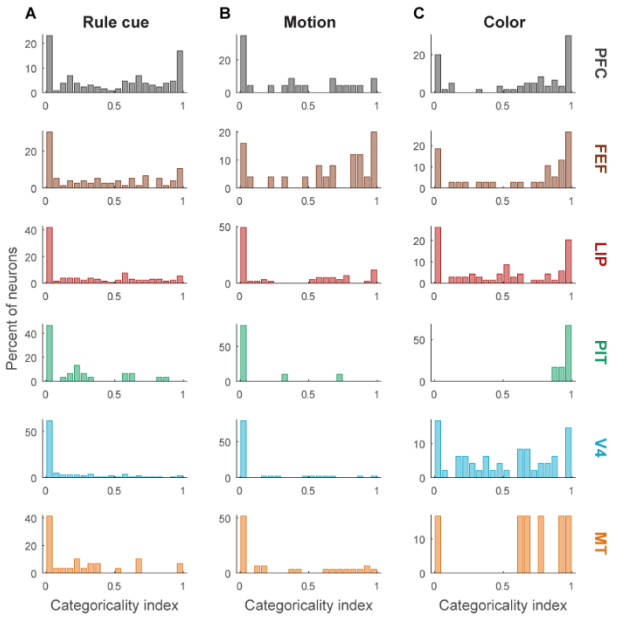

Fig. S3. Distribution of single-neuron categoricality. (A) Population distributions of rule cue categoricality indices computed on single neurons in each area. (B) Population distributions of motion categoricality indices computed on single neurons in each area. (C) Population distributions of color cue categoricality indices computed on single neurons in each area. For all three domains, areas with high population categoricality indices (cf. Fig. 3E,4E,5E) have a large proportion of nearly purely categorical single neurons (index ≈ 1). However, in almost all cases (except for PIT color coding) there also remains a residual subpopulation of sensory-driven neurons (index ≈ 0), as well as single neurons whose activity reflects a mixture of between-category and within-category effects ($0 \leq \text{index} \geq 1$).

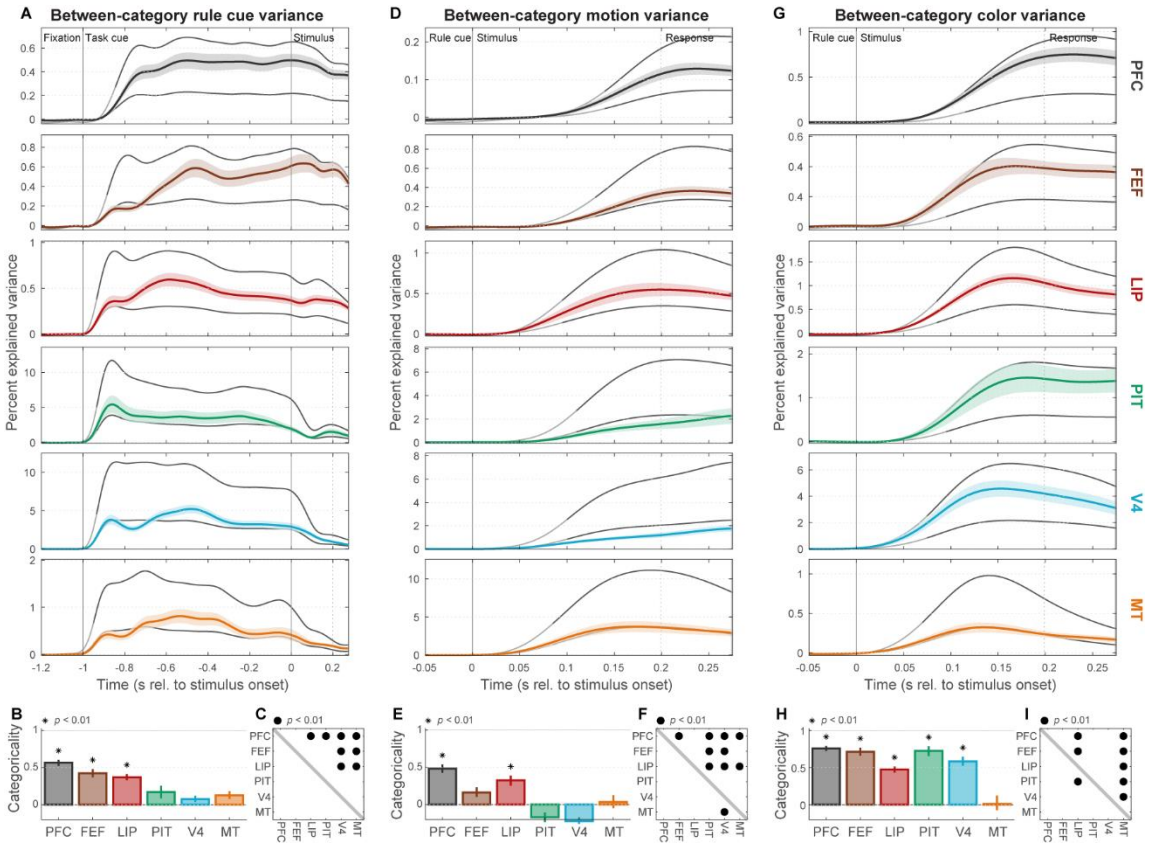

Fig. S4. Main results are similar for multi-unit signals. (A) Population mean (± SEM) multi-unit between-category rule cue variance (cf. Fig. 3D). Multi-unit signals were computed by pooling together all threshold-crossing spikes on each electrode. (B) Summary (± SEM) of multi-unit rule cue categoricality index (cf. Fig. 3E). (C) Cross-area significance matrix for multi-unit rule cue categoricality indices (cf. Fig. 3F). (D) Population mean (± SEM) multi-unit between-category motion variance (cf. Fig. 4D). (E) Summary (± SEM) of multi-unit motion categoricality index (cf. Fig. 4E). (F) Cross-area significance matrix for multi-unit motion categoricality indices (cf. Fig. 4F). (G) Population mean (± SEM) multi-unit between-category color variance (cf. Fig. 5D). (H) Summary (± SEM) of multi-unit color categoricality index (cf. Fig. 5E). (I) Cross-area significance matrix for multi-unit color categoricality indices (cf. Fig. 5F).

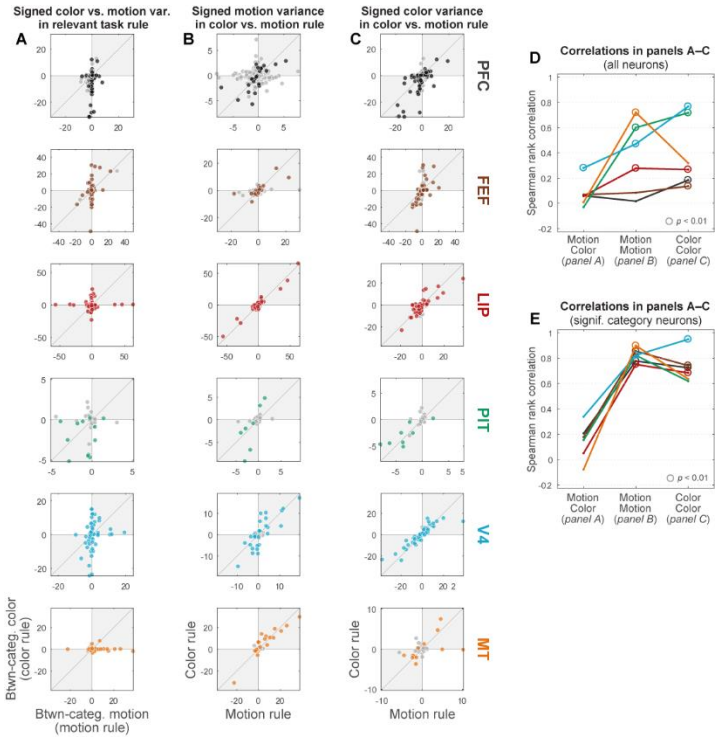

Fig. S5. Category/choice consistency analysis. (A) Signed motion variance for each area neuron under the motion rule (*x*-axis) vs. color variance under the color rule (y-axis), i.e. when each category domain is task-relevant and drives behavioral choice. The sign for each neuron's data point reflects its preferred category—negative for downward-preferring or reddish-preferring, and positive for upward-preferring or greenish-preferring. Neurons with significant between-category motion or color variance ($p < 0.01$; F-test) from the full-session analysis are colored in, while non-significant neurons are light gray. Neurons coding for behavioral choice (or subsequent motor preparation processes) would appear to have consistent motion and color category preferences, and thus would lie near the positive diagonal. (B) Signed between-category motion variance, measured separately within the motion (*x*-axis) and color rule (y-axis) trials. Signs reflect preferred motion categories—negative for downward-preferring, and positive for upward-preferring. Neurons with significant motion variance are colored in. Neurons coding for the same motion category irrespective of the task rule in effect would have consistent values and lie near the positive diagonal. (C) Signed between-category color variance, under the motion (*x*-axis) and color (y-axis) rules. Signs reflect preferred color categories—negative for reddish-preferring, and positive for greenish-preferring. Neurons with significant color variance are colored in. Neurons coding for the same color category irrespective of task rule would have consistent values and lie near the positive diagonal. (D) Spearman rank correlations for all neurons in each scatterplot in panels A–C (*x*-axis). Circles indicate significant correlations ($p < 0.01$; permutation test). Choice coding predicts stronger correlations for the motion/color (left, panel A) condition. Categorical coding predicts stronger correlations for the motion/motion (center, panel B) and color/color (right, panel C) conditions. Results are consistent with categorical coding dominating the overall population in most studied areas, except for PFC and FEF, which appear to contain a heterogeneous mixture of category and choice effects. (E) Correlations for only significant categorical neurons in each scatterplot in panels A–C (*x*-axis). Predictions are same as in D. All areas exhibit correlation patterns consistent with categorical, rather than choice, coding. This

confirms that our variance-partitioning model successfully recovers categorical coding, unconfounded by choice signals.

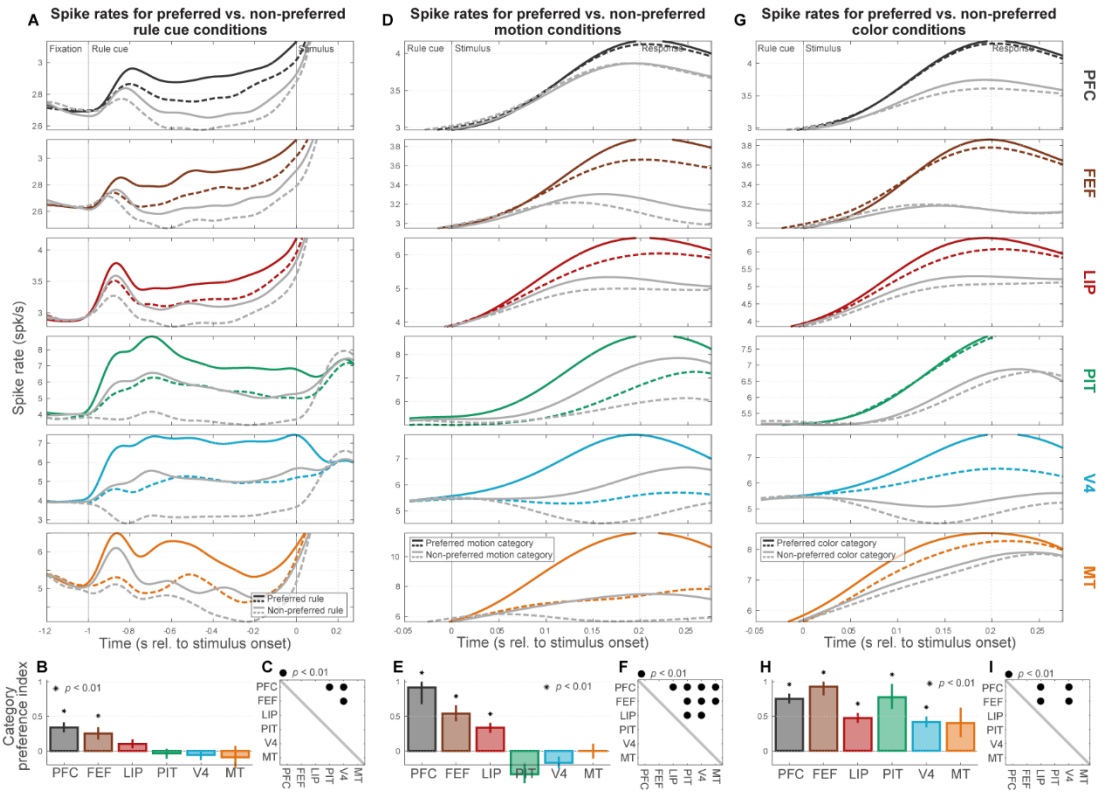

Fig. S6. Preferred condition spike density analysis. (A) Population mean spike rates for preferred (colored lines) and non-preferred (light gray lines) task rules, and for preferred (solid lines) and non-preferred (dashed lines) rule cues within each task rule. This is plotted separately for each studied area as a function of within-trial time (referenced to the onset of the random-dot stimulus). Distinct subsets of trials were used to estimate preferred conditions and compute mean rates with the estimated sorting, to avoid circularity in the analysis. Note that rates for some areas exceed the plotting range outside the time epoch of interest here (the rule cue period). Broadly consistent with the analyses in the main text, rate differences for cues instructing the same task were small compared to differences between tasks in PFC and FEF, whereas for visual areas MT, V4, and PIT these differences were of comparable size. (B) Task rule category preference index (± SEM) for each area, which summarizes results in (A) by contrasting between-category and within-category population rate differences (cf. Fig 3E in the main text). Only PFC and FEF had index values significantly greater than zero ($p < 0.01$, asterisks). (C) Indicates which regions (rows) had significantly greater task rule preferred condition indices than others (columns; $p < 0.01$, dots; cf. main text Fig. 3F). (D) Mean spike rates for preferred (colored lines) and non-preferred motion categories (light gray lines), and for preferred (solid) and non-preferred (dashed) directions within each motion category. Between-category rate differences were large compared to within-category differences in PFC, FEF, and LIP, whereas for visual areas MT, V4, and PIT these differences were comparable. (E) Motion category preference index (± SEM) for each area, summarizing results in (D). (cf. main text Fig 4E). PFC, FEF, and LIP had index values significantly greater than zero ($p < 0.01$, asterisks). (F) Indicates which regions (rows) had significantly greater motion preferred condition indices than others (columns; $p < 0.01$, dots; cf. main text Fig. 4F). (G) Mean spike rates for preferred (colored lines) and non-preferred color categories (light gray lines), and for preferred (solid) and non-preferred (dashed) colors within each color category. Between-category rate differences were large compared to within-category differences in FEF, PIT, and PFC, but less so for areas MT, V4, and LIP. (H) Color category preference index (± SEM) for each area, summarizing results in (G). (cf. main text Fig 5E). FEF,

PIT, PFC, LIP, and V4 all had index values significantly greater than zero ($p < 0.01$, asterisks). (I) Indicates which regions (rows) had significantly greater color preferred condition indices than others (columns; $p < 0.01$, dots; cf. main text Fig. 5F).

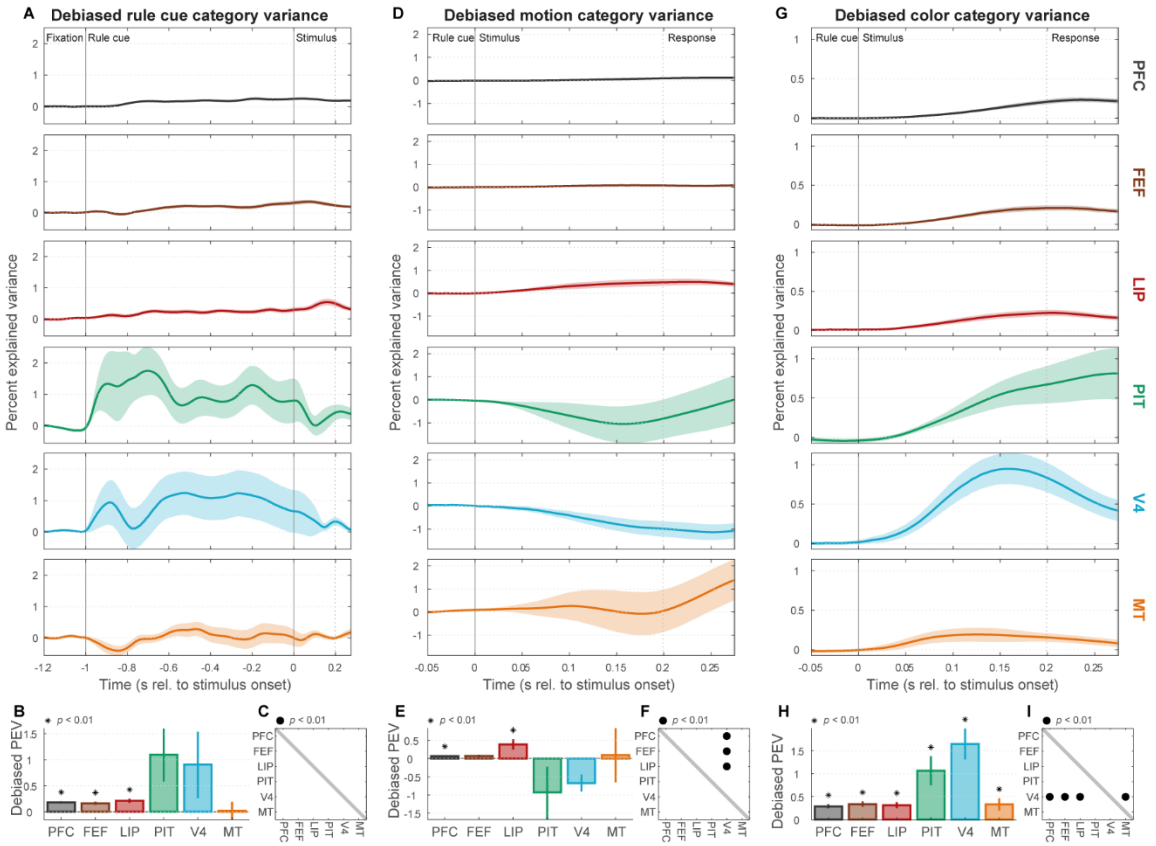

Fig. S7. Debiased category variance results (cf. Siegel et al. 2015). (A) Population mean (± SEM) debiased rule cue category variance, the statistic used to measure task rule information in our previous publication from this dataset (3). As expected, results under this statistic look quite different and appear to reflect both task cue categoricality per se (cf. Fig. 3D) as well as the overall information about task cues (cf. Fig. 3A). (B) Summary (± SEM) of debiased rule cue category variance (cf. Fig. 3B,E). (C) Cross-area significance matrix for rule cue category variance (cf. Fig. 3C,F). (D) Population mean (± SEM) debiased motion category variance (cf. Fig. 4A,D). (E) Summary (± SEM) of debiased motion category variance (cf. Fig. 4B,E). (F) Cross-area significance matrix for motion category variance (cf. Fig. 4C,F). (G) Population mean (± SEM) debiased color category variance (cf. Fig. 5A,D). (H) Summary (± SEM) of debiased color category variance (cf. Fig. 5B,E). (I) Cross-area significance matrix for color category variance (cf. Fig. 5C,F).

Table S1. Comparison of properties of analysis methods used. Table compares basic properties of all analysis methods examined in this paper. A green check indicates the given analysis features the given property; a red x indicates it does not. Analyses listed in table rows are: the categoricality index, proposed in this paper (main text Fig. 3E,4E,5E; simulations in Fig. S1A–D); raw between-category variance (main text Fig. 3D,4D,5D; simulations in Fig. S1E,F); the debiased category variance statistic, proposed in Siegel et. al 2015 (Fig. S7; simulations in Fig. S1G,H); the category preference index (Fig. S6; simulations in Fig. S1I,J).

	Unbiased for pure sensory coding	*Insensitive to overall rate*	*Normalizes across-cell rate differences*	*Partitions out choice effects*
Categoricality index ***(Fig. 3–5)***	✓	✓	✓	✓
Between-category variance	✗	✗	✓	✓
Debiased category variance (Fig. S7)	✓	✗	✓	✓
Category preference index (Fig. S6)	✓	✓	✗	✗

References

1. Asaad WF, Santhanam N, McClellan S, Freedman DJ (2013) High-performance execution of psychophysical tasks with complex visual stimuli in MATLAB. J Neurophysiol 109(1):249–260.
2. Draper NR, Smith H (1998) Applied regression analysis (Wiley, New York). 3rd ed.
3. Siegel M, Buschman TJ, Miller EK (2015) Cortical information flow during flexible sensorimotor decisions. Science 348(6241):1352–1355.
4. Anton H (2010) Elementary Linear Algebra (Wiley, Hoboken, New Jersey). 10th ed.
5. Olejnik S, Algina J (2003) Generalized Eta and Omega Squared Statistics: Measures of Effect Size for Some Common Research Designs. Psychol Methods 8(4):434–447.
6. Machens CK, Romo R, Brody CD (2010) Functional, But Not Anatomical, Separation of "What" and "When" in Prefrontal Cortex. J Neurosci 30(1):350–360.
7. Rigotti M, et al. (2013) The importance of mixed selectivity in complex cognitive tasks. Nature 497(7451):585–590.
8. Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in the primate cerebral cortex. Cereb Cortex 1(1):1–47.
9. Rao SC, Rainer G, Miller EK (1997) Integration of what and where in prefrontal cortex. Science 276(5313):821–824.
10. Kim J-N, Shadlen MN (1999) Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. Nat Neurosci 2(2):176.
11. Johnston K, Everling S (2008) Neurophysiology and neuroanatomy of reflexive and voluntary saccades in non-human primates. Brain Cogn 68(3):271–283.
12. Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2001) Categorical Representation of Visual Stimuli in the Primate Prefrontal Cortex. Science 291(5502):312–316.
13. Kreiman G, Koch C, Fried I (2000) Category-specific visual responses of single neurons in the human medial temporal lobe. Nat Neurosci 3(9):946.