Categorical Biases in Human Visual Cortex

Edward F. Ester[1*], Thomas C. Sprague[2], John T. Serences[3]

[1]*Department of Psychology, Center for Complex Systems and Brain Sciences, and FAU Brain Institute, Florida Atlantic University*
[2]*Department of Psychological and Brain Sciences, University of California, Santa Barbara*
[3]*Department of Psychology, Neurosciences Graduate Program, and Kavli Institute for Brain and Mind, University of California, San Diego*

*Correspondence:

Edward Ester
Department of Psychology, Center for Complex Systems and Brain Sciences, and FAU Brain Institute
Florida Atlantic University
777 Glades Rd.
Boca Raton, FL. 33431
eester@fau.edu

Acknowledgments: Funding provided by NIH R01 EY025872 and a James S. McDonnell Foundation award to J.T.S.. E.F.E., conceived and designed the experiment, collected and analyzed the data, and wrote the paper. T.C.S. provided conceptual input during all phases of the project and edited the paper. J.T.S. supervised all phases of the project. The authors thank Kelvin Lam for assistance with data collection for Experiment 2. The authors declare no competing interests.

Abstract

Categorization refers to the process of mapping continuous sensory inputs onto discrete concepts. Humans, nonhuman primates, and rodents can readily learn arbitrary categories defined by low-level visual features such as hue and orientation, and behavioral studies indicate that such learning distorts perceptual sensitivity for category-defining features such that discrimination performance for physically similar exemplars from different categories is enhanced while discrimination performance for equally similar exemplars from the same category is reduced. These distortions may result from systematic biases in neural representations of discriminanda that begin at the earliest stages of visual processing. We tested this hypothesis in two experiments where human observers learned to classify a set of oriented stimuli into two discrete groups. After behavioral training, we used multivoxel pattern analysis and an inverted encoding model to visualize and quantify population-level neural representations of stimulus orientation from noninvasive measurements of human brain activity (fMRI and EEG) in early retinotopic visual cortical areas. These analyses revealed that during category discrimination neural representations of oriented stimuli in early visual areas were systematically biased towards the center of the appropriate category. These shifts were strongest for orientations near the category boundary, predicted participants' overt category judgments, and emerged rapidly after stimulus onset. Collectively, these results suggest that category information can bias processing at very early stages of the visual processing hierarchy.

Categorization refers to the process of mapping continuous sensory inputs onto discrete and behaviorally relevant concepts. It is a cornerstone of flexible behavior that allows organisms to generalize existing knowledge to novel stimuli and to discriminate between physically similar yet conceptually different stimuli. Many real-world categories are defined by a combination of low-level visual properties such as hue, luminance, spatial frequency, and orientation. For example, a forager might be tasked with determining whether a food source is edible vs. inedible based on subtle variations in color, shape, size, and texture. Humans and other animals can readily learn arbitrary novel categories defined by low-level visual properties (*1*-*2*), and such learning "distorts" perceptual sensitivity for the category-relevant feature such that discrimination performance for physically similar yet categorically distinct exemplars is increased (i.e., acquired distinctiveness; *3*-*4*) and discrimination performance for equally similar exemplars in the same category is reduced (i.e., acquired similarity; *5*).

Invasive electrophysiological studies have shown that single-unit responses in early visual areas index the physical properties of a stimulus but not its category membership, while single-unit responses in later areas index the category membership of a stimulus regardless of its physical properties (e.g., *6*-*8*). These results have been taken as evidence that category-selective responses are a *de novo* property of higher-order visual areas. However, perceptual distortions following category learning could also reflect changes in how information is represented by sensory neural populations (*9*-*10*). Here, we sought to test this possibility. We applied multivariate analyses to noninvasive measurements of human brain activity (fMRI and EEG) to visualize and quantify population-level representations of oriented stimuli in early visual cortical areas after participants had learned to classify these stimuli into discrete groups. We show that representations of to-be-categorized orientations in visual areas V1-V3 are systematically biased

towards the center of the category to which they belong. These biases were correlated with trial-by-trial variability in overt category judgments and were largest for orientations adjacent to the category boundary where they would be most beneficial for discrimination performance. In a second experiment, we used EEG to generate time-resolved representations of to-be-categorized orientations and show that categorical biases manifest rapidly after stimulus onset. Collectively, our results suggest that category knowledge can alter stimulus processing at very early stages of the visual system.

Results

Experiment 1 - fMRI

We trained eight human volunteers to categorize a set of orientations into two groups, Category 1 and Category 2. The stimulus space comprised a set of 15 oriented stimuli, spanning 0-168° in 12° increments (Figure 1A-B). For each participant, we randomly designated one of these 15 orientations as a category boundary such that the seven orientations anticlockwise to the boundary were assigned membership in Category 1 and the seven orientations clockwise to the boundary were assigned membership in Category 2. Each participant completed a one-hour training session prior to scanning. Each participant's category boundary was kept constant across all behavioral training and scanning sessions. Many participants self-reported that they learned the rule delineating the categories in one or two 5-minute blocks of trials. Consequently, task performance measured during scanning was extremely high, with errors and slow responses present only for exemplars immediately adjacent to the category boundary (Figure 1C-D). During each scanning session, participants performed the category discrimination task and an orientation model estimation task where they were required to report the identity of a target letter embedded within a rapid stream presented at fixation while a task-irrelevant grating flickered in the background. Data from this task were used to compute an unbiased estimate of orientation selectivity for each voxel in visual areas V1-hV4v/V3A (see below).

We first examined whether category training increased the similarity of activation patterns evoked by exemplars from the same category (i.e., acquired similarity). We tested this by training a linear decoder (support vector machine) to discriminate between activation patterns associated with exemplars at the center of each category (48° from the boundary), then used the trained classifier to predict the category membership of exemplars immediately adjacent to the

category boundary (±12°; Figure 2A). This analysis was performed separately for the orientation mapping and category discrimination tasks. We reasoned that if category training homogenizes activation patterns evoked by members of the same category, then decoding performance should be superior during the category discrimination task relative to the orientation mapping task. This is precisely what we observed (Figure 2B). For example, near-boundary decoding performance in V1 was reliably above chance during the category discrimination task ($p < 0.0001$, FDR-corrected bootstrap test), but not during the orientation mapping task when the category boundary was irrelevant and the oriented stimulus was unattended ($p = 0.38$). Importantly, the absence of robust decoding performance during the orientation mapping task cannot be attributed to poor signal, as a decoder trained and tested on activation patterns associated with exemplars at the center of each category (Figure 2C) yielded above-chance decoding during both behavioral tasks (Figure 2D; M = 0.58 and 0.69 for the mapping and discrimination tasks, respectively; $p < 0.01$). Collectively, these results suggest that category training can alter population-level responses at very early stages of the visual processing hierarchy.

To better understand how category training influences orientation-selective activation patterns in early visual cortical areas, we used an inverted encoding model (*11*) to generate model-based reconstructed representations of stimulus orientation from these patterns. For each visual area (e.g., V1), we first modelled voxel-wise responses measured during the orientation mapping task as a weighted sum of idealized orientation channels, yielding a set of weights that characterize the orientation selectivity of each voxel (Figure 3A). In the second phase of the analysis, we reconstructed trial-by-trial representations of stimulus orientation by combining these weights with the observed pattern of activation across voxels measured during each trial of the category discrimination task, resulting in single-trial reconstructed channel response function

that contains a representations of stimulus orientation for each ROI on each trial (Figure 3B). Finally, we sorted trial-by-trial reconstructions according to category membership such that any bias would manifest as a clockwise (rightward) shift of the orientation representation towards the center of Category 2 and quantified biases towards this category using a curve-fitting analysis (Methods).

Note that stimulus orientation was irrelevant during the orientation mapping task used for model weight estimation. We therefore reasoned that voxel-by-voxel responses evoked by each oriented stimulus would be largely uncontaminated by its category membership. Indeed, the logic of our analytical approach rests on the assumption that orientation-selective responses are quantitatively different during the orientation mapping and category discrimination tasks: if identical category biases are present in both tasks then the orientation weights computed using data from either task will capture that bias and reconstructed representations of orientation will not exhibit any category shift. This is precisely what we observed when we used a cross-validation approach to reconstruct stimulus representations separately for the orientation mapping and category discrimination tasks (Supplementary Figure 1).

As shown in Figure 4, reconstructed representations of orientation in visual areas V1, V2, and V3 were systematically biased away from physical stimulus orientation and towards the center of the appropriate category (shifts of 22.13°, 26.65°, and 34.57°, respectively; $P < 0.05$, bootstrap test, false-discovery-rate [FDR] corrected for multiple comparisons across regions; see Supplementary Figure 2 for separate reconstructions of Category 1 and Category 2 orientations and Supplementary Figure 3 for participant-by-participant reconstructions plotted by visual area). Similar, though less robust biases were also evident in hV4v and V3A (mean shifts of 9.73° and 6.45°, respectively; $p > 0.19$). A logistic regression analysis established that

categorical biases in V1-V3 were strongly correlated with variability in overt category judgments (Supplementary Figure 4). That is, trial-by-trial category judgments were more strongly associated with the responses of orientation channels near the center of each category rather than those near the physical orientation of the stimulus. Importantly, because the location of the boundary separating categories 1 and 2 was randomly selected for each participant, it is unlikely that categorical biases shown in Figure 4 reflect intrinsic biases in stimulus selectivity in early visual areas (e.g., due to oblique effects; *12*).

The category biases shown in Figure 4 may be the result of an adaptive process that facilitates task performance by enhancing the discriminability of physically similar but categorically distinct stimuli. To illustrate, consider a hypothetical example where an observer is tasked with discriminating between two physically similar exemplars on opposite sides of a category boundary (Supplementary Figure 5). Discriminating between these alternatives should be challenging as each exemplar evokes a similar and highly overlapping response pattern. However, discrimination performance could be improved if the responses associated with each exemplar are made more separable via acquired distinctiveness following training (or equivalently, an acquired similarity between exemplars adjacent to the category boundary and exemplars near the center of each category). Similar changes would be less helpful when an observer is tasked with discriminating between physically and categorically distinct exemplars, as each exemplar already evokes a dissimilar and non-overlapping response. From these examples, a simple prediction can be derived: categorical biases in reconstructed representations of orientation should be largest when participants are shown exemplars adjacent to the category boundary and progressively weaker when participants are shown exemplars further away from the category boundary.

We tested this possibility by sorting stimulus reconstructions according to the angular distance between stimulus orientation and the category boundary (Figure 5). As predicted, reconstructed representations of orientations adjacent to the category boundary were strongly biased by category membership, with larger biases for exemplars nearest to the category boundary (μ = 42.62°, 24.16°, and 20.12° for exemplars located 12°, 24°, and 36° from the category boundary, respectively; FDR-corrected bootstrap $p < 0.0015$), while reconstructed representations of orientations at the center of each category exhibited no signs of bias (μ = -3.98°, $p = 0.79$; the direct comparison of biases for exemplars adjacent to the category boundary and in the center of each category was also significant; $p < 0.01$). Moreover, the relationship between average category bias and distance from the category boundary was well-approximated by a linear trend (slope = -14.38°/step; $r^2 = 0.96$). Thus, category biases in reconstructed representation are largest under conditions where they would facilitate behavioral performance and absent under conditions where they would not.

Category-selective signals have been identified in multiple brain areas, including portions of lateral occipital cortex, inferotemporal cortex, posterior parietal cortex, and lateral prefrontal cortex (*6-10; 13-14*). We identified category selective information in many of these same regions using a whole-brain searchlight-based decoding analysis where a classifier was trained to discriminate between exemplars from Category 1 and Category 2 (independently of stimulus orientation; Figure 6 and Methods). Next, we used the same inverted encoding model described above to reconstruct representations of stimulus orientation from activation patterns measured in each area during each of the orientation mapping and category discrimination tasks (reconstructions were computed using a "leave-one-participant-out" cross-validation routine to ensure that reconstructions were independent of the decoding analysis used to define category-

selective ROIs). We were able to reconstruct representations of stimulus orientation in many of these regions during the category discrimination task, but not during the orientation mapping task (where stimulus orientation was task-irrelevant; Supplementary Figure 6). This is perhaps unsurprising as representations in many mid-to-high order cortical areas are strongly task-dependent (e.g., *15*). As our analytical approach requires an independent and unbiased estimate of each voxel's orientation selectivity (e.g., during the orientation mapping task), this meant that we were unable to probe categorical biases in reconstructed representations in these regions.

Experiment 2 - EEG

Due to the sluggish nature of the hemodynamic response, the category biases shown in Figures 4 and 5 could reflect processes related to decision making or response selection rather than stimulus processing. In a second experiment, we evaluated the temporal dynamics of category biases using EEG. Specifically, we reasoned that if the biases shown in Figures 4 and 5 reflect processes related to decision making, response selection, or motor planning, then these biases should manifest only during a period shortly before the participants' response. Conversely, if the biases are due to changes in how sensory neural populations encode features, they should be evident during the early portion of each trial. To evaluate these alternatives, we recorded EEG while a new group of 27 volunteers performed variants of the orientation mapping and categorization tasks used in the fMRI experiment. On each trial, participants were shown a large annulus of iso-oriented bars that flickered at 30 Hz (i.e., 16.67 ms on, 16.67 ms off; Figure 7A). During the orientation mapping task, participants detected and reported the identity of a target letter (an X or a Y) that appeared in a rapid series of letters over the fixation point. Identical displays were used during the category discrimination task, with the caveat that

participants were asked to report the category of the oriented stimulus while ignoring the letter stream.

The 30 Hz flicker of the oriented stimulus elicits a standing wave of frequency-specific sensory activity known as a steady-state visually-evoked potential (SSVEP, *16*; Figure 7B). The coarse spatial resolution of EEG precludes precise statements about the cortical source(s) of these signals (e.g., V1, V2, etc.). However, to focus on visual areas (rather than parietal or frontal areas) we deliberately entrained stimulus-locked activity at a relatively high frequency (30 Hz). Our approach was based on the logic that coupled oscillators can only be entrained at high frequencies within small local networks, while larger or more distributed networks can only be entrained at lower frequencies due to conduction delays (*17*). Indeed, a topographic analysis showed that evoked 30 Hz activity was strongest over a localized region of occipitoparietal electrode sites. (Figure 7C). As in Experiment 1, participants learned to categorize stimuli with a high degree of accuracy, with errors and slow responses present only for exemplars adjacent to a category boundary (Figure 7D-E)

We computed the power and phase of the 30 Hz SSVEP response across each 3,000 ms trial and then used these values to reconstruct a time-resolved representation of stimulus orientation (*18*). Our analysis procedure followed that used in Experiment 1: In the first phase of the analysis, we rank-ordered scalp electrodes by 30 Hz power (based on a discrete Fourier transform spanning the 3000 ms trial epoch, averaged across all trials of both the orientation mapping and category discrimination tasks). Responses measured during the orientation mapping task were used to estimate a set of orientation weights for the 32 electrodes with the strongest SSVEP signals (i.e., those with the highest 30 Hz power; see Figure 6C) at each timepoint. In the second phase of the analysis, we used these timepoint-specific weights and corresponding

responses measured during each trial of the category discrimination task across all electrodes to compute a time-resolved representation of stimulus orientation (Figure 7A-B).

Importantly, the bandpass filter used to isolate 30 Hz activity will distort temporal characteristics of the raw EEG signal by some extent. We estimated the extent of this distortion by generating a 3 second, 30 Hz sinusoid with unit amplitude (plus 1 second of pre-and post-signal zero padding) and running it through the same filters used in our analysis path. We then computed the time at which the filtered signal reach 25% of maximum. For an instantaneous filter, this should occur at exactly 1 second (due to the pre- and post-signal zero-padding). We estimated a signal onset of ~877 ms, or 123 ms prior to the start of the signal. Therefore, if reconstruction amplitude is greater than zero at time t, then we can conclude that the pattern of scalp activity used to generate the stimulus reconstruction contained reliable orientation information at time $t \pm 125$ ms. The same logic applies to estimates of reconstruction bias as the reconstructions are based on data filtered using the same parameters. Importantly, we also verified that there was no categorical bias in stimulus reconstructions prior to stimulus onset (see Supplementary Figure 7), nor were categorical biases present when we reconstructed stimulus representations using data from the orientation mapping and category discrimination tasks separately (Supplementary Figure S8)

We reasoned that if the categorical biases shown in Figures 4 and 5 reflect processes related to decision making or response selection, then they should emerge gradually over the course of each trial. Conversely, if the categorical biases reflect changes in sensory processing, then they should manifest shortly after stimulus onset. To test this possibility, we computed a temporally averaged stimulus reconstruction over an interval spanning 0 to 250 ms after stimulus onset (Figure 8B). A robust category bias was observed ($M = 23.35°$; $p = 0.014$; bootstrap test)

suggesting that the intent to categorize a stimulus modulates how neural populations in early visual areas respond to incoming sensory signals. An analysis of pre-trial activity revealed no such bias (Supplementary Figure 7), suggesting that our findings cannot be explained by temporal smearing of pre-stimulus activity.

Discussion

Our findings suggest that category learning changes how sensory neural populations code stimulus-specific information at the earliest stages of the visual system. The results of Experiment 1 showed that representations of a to-be-categorized stimulus encoded by population-level activity in early visual cortical areas are systematically biased by their category membership. These biases were correlated with overt category judgments and were largest for exemplars adjacent to the category boundary. The results of Experiment 2 are consistent with the hypothesis that category biases reflect changes in how sensory neural populations code category-defining information by demonstrating that robust category biases are present almost immediately after stimulus onset.

Several candidate mechanisms may be responsible for the category biases reported here. For example, one possibility is that category training recruits a gain mechanism that enhances the responses of neural populations that maximally discriminate between exemplars from each category (e.g., feature-similarity gain; *19*). A second possibility is that category training causes task-dependent shifts in the spectral preferences of sensory neural populations (e.g., *20-21*). These alternatives are not mutually exclusive; nor is this an exhaustive list. Our data cannot resolve these possibilities. For example, several different patterns of single-unit gain changes and/or tuning shifts can produce identical responses in a single fMRI voxel, and different patterns of single-voxel modulation could produce categorical biases in multivariate stimulus reconstructions (see, e.g., *22* for a detailed discussion of this issue). Ultimately, targeted experiments that combine non-invasive measurements of brain activity with careful psychophysical measurements and detailed model simulations will be needed to conclusively identify the mechanisms responsible for the category biases we have reported here. However,

this does not diminish the novelty or importance of our findings: our data suggest that category learning can influence reconstructed representations of stimuli at the earliest stages of the cortical visual processing hierarchy and challenge the widely-held view that sensory cortical areas signal the physical properties of stimuli irrespective of context.

Our findings appear to conflict with results from nonhuman primate research which suggests that sensory cortical areas do not encode categorical information. However, there is reason to suspect that mechanisms of category learning might be qualitatively different in human and non-human primates. For example, our participants learned to categorize stimuli based on performance feedback after approximately 10 minutes of training. In contrast, macaque monkeys typically require six months or more of training using a similar feedback scheme to reach a similar level of performance, and this extensive amount of training may influence how neural circuits code information (e.g., *23*-*24*). Moreover, there is growing recognition that the contribution(s) of sensory cortical areas to performance on a visual task are highly susceptible to recent history and training effects (*23*, *25*-*27*). In one example (*26*), extensive training was associated with a functional substitution of human visual area V3a for MT+ in discriminating noisy motion patches. Thus, training effects may help explain why previous electrophysiological experiments have found category-selective responses in association but not sensory cortical areas.

Studies of categorization in non-human primates have typically employed variants of a delayed match to category task, where monkeys are shown a sequence of two exemplars separated by a blank delay interval and asked to report whether the category of the second exemplar matches the category of the first exemplar. The advantage of this task is that it allows experimenters to decouple category-selective signals from activity related to decision making,

response preparation, and response execution: since the monkey has no way of predicting whether the category of the second exemplar will match that of the first, it must wait for the second exemplar appears before preparing and executing a response. However, this same advantage also precludes examinations of whether and/or how top-down category-selective signals interact with bottom-up stimulus-specific signals that may explain the biases reported here. We made no effort to decouple category-selective and decision-related signals in our study. That is, we maintained a consistent response mapping for Category 1 and Category 2 throughout the experiment. This can be viewed as an advantage or a handicap. On the one hand, our experimental approach allowed us to quantify category-selective responses in early visual cortex even though a physical stimulus was present for the duration of each trial. On the other hand, we cannot definitively exclude the possibility that the categorical biases reported here reflect decision- or motor-related processes rather than mechanisms of categorization, although it seems unlikely that manual (non-oculomotor) response signals would be present in early visual areas based on existing data.

Our findings are naturally accommodated by hierarchical predictive coding models of brain function (*27-28*). Fundamentally, these models propose that perception (and related functions such as decision making) are the result of a generative process where sensory signals are initially compared to an internal (generative) model of the environment. This comparison yields a measure of prediction error or surprise that is subsequently minimized to yield the most likely percept. As applied to the current study, bottom-up signals engendered by the stimulus on each trial are initially compared with canonical or template representations of the orientation typical of a given category encoded by upstream cortical areas (e.g., posterior parietal cortex and/or lateral prefrontal cortex), yielding a set of prediction errors (relative to each template

representation). Based on this comparison, top-down feedback from these higher-order cortical areas are relayed to early sensory areas to refine the responses of neural populations in these areas. We emphasize that this account is speculative, and additional studies will be needed to evaluate specific predictions from this framework.

Methods

Experiment 1 - fMRI

Participants. Eight neurologically intact volunteers (AA, AB, AC, AD, AE, AF, AG, and AH; six females) from the UC San Diego community participated in Experiment 1. All participants self-reported normal or corrected-to-normal visual acuity and gave both written and oral informed consent as required by the local Institutional Review Board. Each participant completed a single one-hour behavioral training session approximately 24-72 hours prior to scanning. Seven participants (AA, AB, AC, AD, AE, AF, AG) later completed two 2-hour experimental scan sessions; an eighth participant (AH) later completed a single 2-hour experimental scan session. Participants AA, AB, AC, AD, AE, AF, and AH also completed a single 2-hour retinotopic mapping scan session. Data from this session were used to identify visual field borders in early visual cortical areas V1-hV4/V3A and subregions of posterior intraparietal sulcus (IPS0-3; see *Retinotopic Mapping*, below). Participants were compensated at a rate of $10/hr for behavioral training and $20/hr for scanning.

Setup. Stimulus displays were generated in MATLAB using and rendered using Psychophysics Toolbox extensions (*29*). Displays were projected onto a 110 cm-wide screen placed at the base of the MRI bore, and participants viewed displays via a mirror attached to the MR head coil from a distance of 370 cm.

Behavioral Tasks. In separate runs (where “run” refers to a continuous block of 30 trials lasting 280 seconds) participants performed either an orientation mapping task or a category discrimination task. Trials in both tasks lasted 3 seconds, and consecutive trials were separated by a 5 or 7 s inter-trial-interval (pseudorandomly chosen on each trial).

During the orientation mapping task, participants attended a stream of letters presented at fixation (subtending 1.0˚ x 1.0˚ from a viewing distance of 370 cm) while ignoring a task-irrelevant phase-reversing (15 Hz) square-wave grating (0.8 cycles/deg with inner and outer radii of 1.16˚ and 4.58˚, respectively) presented in the periphery. On each trial, the grating was assigned one of 15 possible orientations (0˚-168˚ in 12˚ increments). Participants were instructed to detect and report the identity of a target ("X" or "Y") in the letter stream using an MR-compatible button box. Only one target was presented on each trial. Letters were presented at a rate of 10 Hz (50% duty cycle, i.e. 50 msec on, 50 msec off), and targets could occur during any cycle from +750 to +2250 msec following the start of each trial.

During category discrimination runs, participants were shown displays containing a circular aperture (inner and outer radii of 1.16˚ and 4.58˚ from a viewing distance of 370 cm) filled with 150 iso-oriented bars (see Figure 1A). Each bar subtended 0.2˚ x 0.6˚ with a stroke width of 8 pixels (1024 x 768 display resolution). Each bar flickered at 30 Hz and was randomly replotted within the aperture at the beginning of each "up" cycle. On each trial, all bars were assigned an orientation from 0˚-168˚ in 12˚ increments. Inspired by earlier work in non-human primates (*8*), we randomly selected and designated one of these orientations as a category boundary such that the seven orientations counterclockwise to this value were assigned membership in "Category 1", while the seven orientations clockwise to this value were assigned membership in "Category 2". Participants were not informed that the category boundary was chosen from the set of possible stimulus orientations. Participants reported whether the orientation shown on each trial was a member of Category 1 or 2 (via an MR-compatible button box). Participants were free to respond at any time during each trial, though the stimulus was always presented for a total of 3000 ms. Each participant was familiarized and trained to

criterion performance on the category discrimination task during a one-hour behavioral testing session completed one to three days prior to his or her first scan session. Written feedback ("Correct!" or "Incorrect") was presented in the center of the display for 1.25 sec. after each trial during behavioral training and MR scanning.

Across either one (N = 1) or two (N = 7) scan sessions, each participant completed 7 (N = 1), 13 (N = 1), 14 (N = 1), 15 (N = 1) or 16 (N = 4) runs of the RSVP and category discrimination tasks.

fMRI Acquisition and Preprocessing. Imaging data were acquired with a 3.0T GE MR 750 scanner located at the Center for Functional Magnetic Resonance imaging on the UCSD campus. All images were acquired with a 32 channel Nova Medical head coil (Wilmington, MA). Whole-brain echo-planar images (EPIs) were acquired in thirty-five 3 mm slices (no gap) with an in-plane resolution of 3 x 3 mm (192 x 192 mm field-of-view, 64 x 64 mm image matrix, 90° flip angle, 2000 ms TR, 30 ms TE). During retinotopic mapping scans (see below) EPIs were acquired in 31 3mm thick oblique slices (no gap) positioned over posterior visual and parietal cortex with an in-plane resolution of 2 x 2 mm (192 x 192 mm field-of-view, 96 x 96 mm image matrix, 90° flip angle, 2250 ms TR, 30 ms TE). EPIs were coregistered to a separate anatomical scan collected during the same scan session (FSPGR T1-weighted sequence, 11 ms TR, 3.3 ms TE, 1100 ms TI, 172 slices, 18° flip angle, 1 mm^3 resolution), unwarped (FSL software extensions), slice-time-corrected, motion-corrected, high-pass-filtered (to remove first-, second- and third-order drift), transformed to Talairach space, and normalized (z-score) on a scan-by-scan basis. Data from separate scan sessions were co-registered to a high-resolution anatomical image (FSPGR T1-weighted sequences; parameters as described above) collected during the retinotopic mapping session.

Retinotopic Mapping. Retinotopically organized visual areas V1-hV4v/V3A were defined using data from a single retinotopic mapping run collected during each experimental scan session. Participants fixated a small dot at fixation while phase-reversing (8 Hz) checkerboard wedges subtending 60° of polar angle (at maximum eccentricity) were presented along the horizontal or vertical meridian (alternating with a period of 40 seconds; i.e., 20 seconds of horizontal stimulation followed by 20 seconds of vertical stimulation). To identify visual field borders, we constructed a general linear model with two boxcar regressors, one marking epochs of vertical stimulation and another marking epochs of horizontal stimulation. Each regressor was convolved with a canonical hemodynamic function ("double gamma" as implemented in BrainVoyager QX). Next, we generated a statistical parametric map marking voxels with larger responses during epochs of vertical relative to horizontal stimulation. This map was projected onto a computationally inflated representation of each participant's cortical surface for visualization to aid in the definition of the borders of visual areas V1, V2v, V2d, V3v, V3d, hV4v, and V3A. Data from V2v and V2d were combined into a single V2 ROI, and data from V3v and V3d were combined into a single V3 ROI. ROIs were also combined across cortical hemispheres (e.g., left and right V1) as no asymmetries were observed and the stimulus was presented in the center of the visual field.

Seven participants (AA, AB, AC, AD, AE, AF, and AH) completed a separate two-hour retinotopic mapping scan; data from this session were used to identify retinotopically organized regions of visual and inferior parietal sulcus (IPS0-3). During each task run, participants were shown displays containing a rotating wedge stimulus (period 24.75 or 36 sec) that subtended 72˚ of polar angle with inner and outer radii of 1.75 and 8.75˚, respectively. In alternating blocks, the wedge contained a 4 Hz phase-reversing checkerboard or field of moving dots and participants

were required to detect small, brief, and temporally unpredictable changes in checkerboard contrast or dot speed. Six participants completed between 8 and 14 task runs. To compute the best polar angle for each voxel in IPS we shifted the signals from counterclockwise runs by twice the estimated hemodynamic response function (HRF) delay (2 x 6.75 s = 13.5 s), removed data from the first and last full stimulus cycle, and reversed the time series so that all runs reflected clockwise rotation. We next computed the power and phase of the response at the stimulus' period (either 1/24.75 or 1/36 Hz) and subtracted the estimated hemodynamic response function delay (6.75 seconds) to align the signal phase in each voxel with the stimulus' location. Maps of orientation preference (computed via cross-correlation) were projected onto a computationally inflated representation of each participant's grey-white matter boundary to aide in the identification of visual field borders separating V1-hV4/V3A and IPS0-3. Data from the dorsal and ventral aspects of V2 were combined into a single ROI. An identical approach was used to define a single V3 ROI.

An eighth participant (AG) chose not to participate in an additional retinotopic mapping session. For this participant, we estimated visual field borders for visual areas V1-hV4/V3A. using data from the retinotopic mapping run collected during two experimental scan sessions. Given limited data, we did not attempt to define IPS regions IPS0-3 for this participant.

Decoding Categorical Biases in Visual Cortex. We used a linear decoder to examine whether fMRI activation patterns evoked by exemplars adjacent to the category boundary and at the center of each category were more similar during the category discrimination task relative to the orientation mapping task (i.e., acquired similarity). In the first phase of the analysis, we trained a linear support vector machine (LIBSVM implementation; *30*) to discriminate between the oriented exemplars at the center of each category (48° from the boundary) using data from the

orientation mapping and category discrimination tasks. To ensure internal reliability, we implemented a "leave-one-run-out" cross validation scheme where data from all but one scanning run was used to train the classifier and data from the remaining scanning run were used for validation. This procedure was repeated until data from each scan had served as the validation set, and the results were averaged across permutations. Next, we trained a linear classifier on activation patterns evoked by exemplars at the center of each category boundary and used the trained classifier to predict the category membership of exemplars adjacent to the category boundary. If category learning increases the similarity of activation patterns evoked by exemplars within the same category, then within-category decoding performance should be superior during the category discrimination task relative to the orientation mapping task.

Inverted Encoding Model. A linear inverted encoding model (IEM) was used to recover a model-based representation of stimulus orientation from multivoxel activation patterns measured in early visual areas (*11*). The same general approach was used during Experiment 1 (fMRI) and Experiment 2 (EEG). Specifically, we modeled the responses of voxels (electrodes) measured during the orientation mapping task as a weighted sum of 15 orientation-selective channels, each with an idealized response function (half-wave-rectified sinusoid raised to the 14^{th} power). The maximum response of each channel was set to unit amplitude; thus units of response are arbitrary. Let B_1 (m voxels or electrodes x n_1 trials) be the response of each voxel (electrode) during each trial of the RSVP task, let C_1 (k filters x n_1 trials) be a matrix of hypothetical orientation filters, and let W (m voxels or electrodes x k filters) be a weight matrix describing the mapping between B_1 and C_1:

$$B_1 = W C_1$$

In the first phase of the analysis, we computed the weight matrix W from the voxel-wise (electrode-wise) responses in B_1 via ordinary least-squares:

$$W = B_1 C_1^T (C_1 C_1^T)^{-1}$$

Next, we defined a test data set B_2 (*m* voxels or electrodes x n_2 trials) using data from the category discrimination task. Given W and B_2, a matrix of filter responses C_2 (*k* filters x *n* trials) can be estimated via model inversion:

$$C_2 = (W^T W)^{-1} W^T B_2$$

C_2 contains the reconstructed response of each modeled orientation channel (the channel response function; CRF) on each trial of the category discrimination task. This analysis can be considered a form of model-based directed dimensionality reduction whereby activity patterns are transformed from their original measurement space (fMRI voxels; EEG electrodes) into a modeled information space (orientation-selective channels). Importantly, results from this method cannot be used to infer any changes in orientation tuning – or any properties of neural responses - occurring at the single neuron level, and only assay the information content of large-scale patterns of neural activity (*22*) Additionally, while it is the case that arbitrary linear transforms can be applied to the basis set, model weights, and reconstructed channel response function (*31*), results are uniquely defined for a given model specification. Trial-by-trial CRFs were multiplied by the original basis set to recover a full 180-degree function, circularly shifted to a common center (0°) and sorted by category membership so that any category bias would manifest as a clockwise shift (i.e., towards the center of Category 2).

During Experiment 1, the inverted encoding model was applied to normalized (z-scored) multivoxel activation patterns averaged over time (4-6 seconds after stimulus onset). During

Experiment 2, the model was applied to instantaneous multi-electrode activity patterns at the stimulus' flicker frequency of 30 Hz. To isolate stimulus-specific responses, the epoched EEG timeseries at each electrode was bandpass filtered from 29 to 31 Hz (zero-phase forward and reverse 3rd order Butterworth), and the analytic representation of the resulting time series was computed using a Hilbert transformation. To visualize and quantify orientation-selective signals from frequency-specific responses, we first constructed a complex-valued data set $B_1(t)$ (*m* electrodes x n_{train} trials). We then estimated a complex-valued weight matrix W(t) (*m* channels x *k* filters) using $B_1(t)$ and a basis set of idealized orientation-selective filters C_1. Finally, we estimated a complex-valued matrix of channel responses $C_2(t)$ (*m* channels x n_{test} trials) given W(t) and complex-valued test data set $B_2(t)$ (*m* electrodes x n_{test} trials) containing the complex Fourier coefficients measured during the category discrimination task. Trial-by-trial and sample-by-sample response functions were shifted in the same manner described above so that category biases would manifest as a rightward (clockwise) shift towards the center of Category B. We estimated the evoked (i.e., phase-locked) power of the response at each filter by computing the squared absolute value of the average complex-valued coefficient for each filter after shifting.

Quantification of Bias in Orientation Representations. To quantify categorical biases in reconstructed model-based CRFs, these functions were fit with an exponentiated cosine function of the form:

$$f(x) = \alpha\left(e^{k(\cos(\mu - x) - 1)}\right) + \beta$$

where, *x* is a vector of channel responses and α, β, *k* and μ correspond to the amplitude (i.e., signal over baseline), baseline, concentration (the inverse of bandwidth) and the center of the function, respectively. Fitting was performed using a multidimensional nonlinear minimization algorithm (Nelder-Mead).

Category biases in the estimated center of each construction (μ) during the category discrimination task were quantified via permutation tests. For a given visual area (e.g., V1) we randomly selected (with replacement) stimulus reconstructions from eight of eight participants. Specifically, we computed a "mean" reconstruction by randomly selecting (with replacement) and averaging reconstructions from all participants. The mean reconstruction was fit with the cosine function described above, yielding point estimates of α, β, k, and μ. This procedure was repeated 1,000 times, yielding 1,000 element distributions of parameter estimates. We then computed the proportion of permutations where a μ value less than 0 was obtained to obtain an empirical *p*-value for categorical shifts in reconstructed representations. The same analysis was used to quantify category biases in Experiment 2 (EEG).

Searchlight Decoding of Category Membership. We used a roving searchlight analysis (*32-33*) to identify cortical regions beyond V1-V3 that contained category-specific information. We defined a spherical neighborhood with a radius of 8 mm around each grey matter voxel in the cortical sheet. We next extracted and averaged the normalized response of each voxel in each neighborhood over a period from 4-8 seconds after stimulus onset (this interval was chosen to account for typical hemodynamic lag of 4-6 seconds). A linear SVM (LIBSVM implementation; *30*) was used to classify stimulus category using activation patterns within each neighborhood. To classify category membership independently of orientation, we designated the three orientations immediately counterclockwise to the category boundary (see Figure 1) as members of Category 1 and the three orientations immediately clockwise of the boundary as members of Category 2. We then trained our classifier to discriminate between categories using data from all but one task run. The trained classifier was then used to predict category membership from activation patterns measured during the held-out task run. This procedure was repeated until each

task run had been held out, and the results were averaged across permutations. Finally, we repeated the same analysis using the three Category 1 and Category 2 orientations adjacent to the second (orthogonal) category boundary (see Figure 1) and averaged the results across category boundaries.

We identified neighborhoods encoding stimulus category using a leave-one-participant-out cross validation approach (*34*). Specifically, for each participant (e.g., AA) we randomly selected (with replacement) and averaged classifier performance estimates from each neighborhood from each of the remaining 7 volunteers (e.g., AB-AH). This procedure was repeated 1000 times, yielding a set of 1000 classifier performance estimates for each neighborhood. We generated a statistical parametric map (SPM) for the held-out participant that indexed neighborhoods where classifier performance was greater than chance (50%) on 97.5% of permutations (false-discovery-rate corrected for multiple comparisons across neighborhoods). Finally, we projected each participant's SPM onto a computationally inflated representation of his or her grey-white matter boundary and used Brain Voyager's "Create POIs from Map Clusters" function with an area threshold of 25 mm^2 to identify ROIs supporting above-chance category classification performance. Because of differences in cortical folding patterns, some ROIs could not be unambiguously identified in all 8 participants. Therefore, across participants, we retained all ROIs that were shared by at least 7 out of 8 participants (see Supplementary Figure S5). Finally, we extracted multivoxel activation patterns from each ROI and computed model-based reconstructions of channel response functions during the RSVP and category tasks using a leave-one-run-out cross-validation approach. Specifically, we used data from all but one task run to estimate a set of orientation weights for each voxel in each ROI. We then used these weights and activation patterns measured during the held-out task run to estimate a channel

response function, which contains a representation of stimulus orientation. This procedure was repeated until each task run had been held out, and the results were averaged across permutations. Note that each participant's ROIs were defined using data from the remaining 7 participants. This ensured that participant-level reconstructions were statistically independent of the searchlight method used to define ROIs encoding category information.

Within-participant Error Bars. We report estimates of within-participant variability (e.g., ±1 S.E.M.; Figures 3-4; Figure 6) throughout the paper. These estimates discard subject variance (e.g., overall differences in BOLD response amplitude) and instead reflect variance related to the subject by condition(s) interaction term(s) (i.e., variability in estimated channel responses). We used the approach described by Cousineau (*35*): raw data (e.g., channel response estimates) were de-meaned on a participant by participant basis, and the grand mean across participants was added to each participant's zero-centered data. The grand mean-centered data were then used to compute estimates of standard error.

Experiment 2 - EEG

Participants. 28 new participants completed Experiment 2. All participants self-reported normal or corrected-to-normal visual acuity and gave both written and oral informed consent as required by the local Institutional Review Board. Each participant was tested in a single 2.5-3 hour experimental session (the exact duration varied across participants depending on the amount of time needed to set up and calibrate the EEG equipment). Unlike Experiment 1, participants were not trained on the categorization task prior to testing. We adopted this approach in the hopes of tracking the gradual emergence of categorical biases during learning. However, many participants learned the task relatively quickly (within 40-60 trials), leaving too few trials to enable a direct analysis of this possibility. Monetary compensation was provided at a rate of

$15/hr. Data from one participant were discarded due to a high number of EOG artifacts (over 35% of trials); the data reported here reflect the remaining 27 participants.

Behavioral Tasks. Stimulus displays were generated in MATLAB and rendered on an electrically shielded 19-inch CRT monitor (1024 x 768; 120 Hz) via Psychophysics Toolbox software extensions. Participants were seated approximately 55 cm from the display (head position was unconstrained).

In separate runs (where "run" refers to a continuous block of 60 trials lasting approximately 6.5 minutes), participants performed orientation mapping and category discrimination tasks similar to those used in Experiment 1. During both tasks a rapid series of letters (subtending 1.54˚ x 1.54˚ from a viewing distance of 55 cm) was presented at fixation, and an aperture of 150 iso-oriented bars (subtending 0.8˚ x 2.2˚) was presented in the periphery. The aperture of bars had inner and outer radii of 4.76˚ and 20.13˚, respectively. On each trial, the bars were assigned one of 15 possible orientations (again 0˚-168˚ in 12˚ increments) and flickered at a rate of 30 Hz. Each bar was randomly replotted within the aperture at the beginning of each "up" cycle. Letters in the RSVP stream were presented at a rate of 6.67 Hz

During orientation mapping runs, participants detected and reported the presence of a target letter (an X or Y) that appeared at an unpredictable time during the interval from +750 msec to +2250 ms following stimulus onset. Responses were made on a USB-compatible number pad. During category discrimination runs, participants ignored the RSVP stream and instead reported whether the orientation of the bar aperture was an exemplar from category "1" or category "2". As in the neuroimaging experiment, we randomly designated one of the 15 possible stimulus orientations as the category boundary such that the seven orientations counterclockwise to this value were assigned to category 1 and the seven orientations clockwise

to this value were assigned to category 2. Participants could respond at any point during the trial, but the stimulus was presented for a total of 3000 msec. Trials were separated by a 2.5 – 3.25 sec inter-trial-interval (randomly selected from a uniform distribution on each trial). Each participant completed five (N = 5), six (N = 5), or 7 (N = 1) blocks of each task.

EEG Acquisition and Preprocessing. Participants were seated in a dimly lit, sound-attenuated, and electrically shielded recording chamber (ETS Lindgren) for the duration of the experiment. Continuous EEG was recorded from 128 Ag-AgCl^{-} scalp electrodes via a Biosemi "Active Two" system (Amsterdam, Netherlands). The horizontal electrooculogram (EOG) was recorded from additional electrodes placed near the left and right canthi, and the vertical EOG was recorded from electrodes placed above and below the right eye. Additional electrodes were placed over the left and right mastoids. The horizontal and vertical EOG were recorded from electrodes placed over the left and right canthi and above and below the right eye (respectively). Electrode impedances were kept well below 20 kΩ, and recordings were digitized at 1024 Hz.

After testing, the entire EEG time series at each electrode was high- and low-pass filtered (3rd order zero-phase forward and reverse Butterworth) at 0.1 and 50 Hz and re-referenced to the average of the left and right mastoids. Data from both tasks were epoched into intervals spanning -1000 to +4000 msec from stimulus onset; the relatively large pre- and post-stimulus epochs were included to absorb filtering artifacts that could affect later analyses. Trials contaminated by EOG artifacts (horizontal eye movements > ~2° and blinks) were identified and excluded from additional analyses. Across participants an average of 4.21% (±0.99%) and 7.59% (±1.49%) of trials from the orientation mapping and category discrimination tasks were discarded (respectively). Finally, noisy channels (those with multiple deflections ≥ 100 μV over the course

of the experiment) were visually identified and eliminated (mean number of removed electrodes across participants ±1 S.E.M. = 2.29 ± 0.66).

Next, we identified a set of electrodes-of-interest (EOIs) with strong responses at the stimulus' flicker frequency (30 Hz). Data from each task were re-epoched into intervals spanning 0 to 3000 msec around stimulus onset and averaged across trials and tasks (i.e., RSVP and category discrimination), yielding a *k* electrode by *t* sample data matrix. Next, we computed the evoked power at the stimulus' flicker frequency (30 Hz) by applying a discrete Fourier transform to the average time series at each electrode. We selected the 32 electrodes with the highest evoked power at the stimulus' flicker frequency for further analysis. These electrodes were typically distributed over occipitoparietal electrode sites (Figure 7C).

To isolate stimulus-specific responses, the epoched EEG timeseries at each electrode was resampled to 256 Hz and then bandpass filtered from 29 to 31 Hz (zero-phase forward and reverse 3rd order Butterworth). We next estimated a set of complex Fourier coefficients describing the power and phase of the 30 Hz response by applying a Hilbert transformation to the filtered data. To visualize and quantify orientation-selective signals from frequency-specific responses, we first constructed a complex-valued data set $B_1(t)$ (*m* electrodes x n_{train} trials). We then estimated a complex-valued weight matrix W(t) (*m* channels x *k* filters) using $B_1(t)$ and a basis set of idealized orientation-selective filters C_1. Finally, we estimated a complex-valued matrix of channel responses $C_2(t)$ (*m* channels x n_{test} trials) given W(t) and complex-valued test data set $B_2(t)$ (*m* electrodes x n_{test} trials) containing the complex Fourier coefficients measured during the category discrimination task. Trial-by-trial and sample-by-sample response functions were shifted in the same manner described above so that category biases would manifest as a rightward (clockwise) shift towards the center of Category 2. We estimated the evoked (i.e.,

phase-locked) power of the response at each filter by computing the squared absolute value of the average complex-valued coefficient for each filter after shifting. Categorical biases were quantified using the same curve fitting analysis described in the main text.

As noted above, orientation mapping and category discrimination trials contaminated by EOG artifacts were excluded from this analysis. To obtain an unbiased estimate of orientation selectivity in each electrode, we ensured that the training data set $B_1(t)$ contained an equal number of trials for each stimulus orientation (0-168° in 12° increments). For each participant, we identified the stimulus orientation θ with the N fewest repetitions in the orientation mapping data set after EOG artifact removal. Next, we constructed the training data set $B_1(t)$ by randomly selecting (without replacement) 1:N trials for each stimulus orientation. Data from this training set were used to estimate a set of orientation weights for each electrode and these weights were in turn used to estimate a response for each hypothetical orientation channel during the category discrimination task. To ensure that our method generalized across multiple combinations of orientation mapping trials, we repeated this analysis 100 times and averaged the results across permutations.

References

1. Goldstone, R.L. Perceptual Learning, *Annu Rev Psychol* **49** 585-612 (1998)

2. Ashby, F.G. & Maddox, W.T. Human Category Learning. *Annu Rev Psychol* **56** 149-178 (2005)

3. Goldstone, R.L. Influences of categorization on perceptual discrimination., *J Exp Psychol Gen* **123**, 178-200 (1994)

4. Newell, F.N. & Bulthoff, H.H., Categorical perception of familiar objects. *Cognition* **85**, 113-143 (2002)

5. Livingston, K., Andrews, J. & Harnad, S. Categorical perception effects induced by category learning. *J Exp Psychol Learn Mem Cogn* **24** 732-753 (1998)

6. Sigala, N. & Logothetis, N.K. Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature* **415**, 318-320 (2002)

7. Freedman, D.J., Riesenhuber, M., Poggio, T. & Miller, E.K. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* **291** 312-316 (2001)

8. Freedman, D.J. & Assad, J.A. Experience-dependent representation of visual categories in parietal cortex. *Nature* **443** 85-88 (2006)

9. Folstein, J.R., Palmeri, T.J. & Gauthier, I. Category learning increases discriminability of relevant object dimensions in visual cortex. *Cereb Cortex* **23** 814-823 (2012)

10. Davis, T. & Poldrack, R.A. Quantifying the internal structure of categories using a neural typicality measure. *Cereb Cortex* **24** 1720-1737 (2013)

11. Brouwer, G.J. & Heeger, D.J. Cross-orientation suppression in human visual cortex. *J Neurophysiol* **106** 2108-2119 (2011)

12. Sun, P., Gardner, J.L., Costagli, M., Ueno, K., Waggoner, R.A., *et al*. Demonstration of tuning to stimulus orientation in the human cortex: A high-resolution fMRI study with a novel continuous stimulation paradigm. *Cereb Cortex* **23** 1618-1629 (2013)

13. Pourtois, G., Schwartz, S., Spiridon, M., Martuzzi, R. & Vuilleumier, P. Object representations for multiple visual categories overlap in lateral occipital and medial fusiform cortex. *Cereb Cortex* **19**, 1806-1819 (2008)

14. Mack, M.L., Preston, A.R. & Love, B.C. Decoding the brain's algorithm for categorization from its neural implementation. *Curr Biol* **23** 2023-2027 (2013)

15. Silver, M.A., Ress, D. & Heeger, D.J. Topographic maps of visual spatial attention in human parietal cortex. *J Neurophysiol* **94**, 1358-1371 (2005)

16. Vialatte, F-B, Maurice M, Dauwels J, & Cichocki A (2010) Steady-state visually evoked potentials: Focus on essential paradigms and future perspectives. *Progress Neurobiology* 90:418-438 (2010)

17. Breakspear, M., Heitmann, S. & Daffertshofer, A. Generative models of cortical oscillations: Neurobiological implications of the Kuramoto model. Front Hum Neurosci **4** 190 (2010)

18. Garcia, J.O., Sreenivasan, R. & Serences, J.T. Near-real-time feature-selective modulations in human cortex. *Curr Biol* **23**, 515-522 (2013)

19. Martinez-Trujillo JC, Treue S. Feature-based attention increases the selectivity of population responses in primate visual cortex. Curr Biol 14:744-751 (2004)

20. David, S.V., Hayden, B.Y., Mazer, J.A. & Gallant, J.L. Attention to stimulus features shifts spectral tuning of V4 neurons during natural vision. *Neuron* **59**, 509-521 (2008)

21. Koida, K. & Komatsu, H. Effects of task demands on the responses of color-selective neurons in the inferior temporal cortex. *Nat. Neurosci* **10** 108-116 (2007).

22. Sprague TC, Adam KCS, Foster JJ, Rahmati M, Sutterer DW, Vo VA. Inverted encoding models assay population-level stimulus representations, not single-unit neural tuning. *eNeuro* **5**(3) (2018)

23. Itthipurripat S, Cha K, Byers A, Serences JT. Two different mechanisms support selective attention at different phases of training. *PLOS Biology* (2017)

24. Birman D, Gardner JL. Parietal and prefrontal: categorical differences? *Nat Neurosci* **19** 5-7 (2015)

25. Chen, N., Cai, P., Zhou, T., Thompson, B. & Fang, F. Perceptual learning modifies the functional specializations of visual cortical areas. *Proc Natl Acad Sci USA* **113** 5724-5729 (2016)

26. Liu, L.D. & Pack, C.C. The contribution of area MT to visual motion perception depends on training. *Neuron* **95** 436-446 (2017)

27. Rao, R.P.N., Ballard, D.H. Predictive coding in the visual cortex: A functional interpretation of some extra-classicial receptive-field effects. *Nat Neurosci* **2** 79-87 (1999)

28. Friston, K. The free-energy principle: A unified brain theory? *Nat Rev Neurosci* **11** 127-138 (2010)

29. Kleiner M, Brainard D, Pelli D. (2007) What's new in Psychtoolbox-3. *Perception* 36: 14

30. Chang C-C, Lin CJ (2011) LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol* 2(27):1:27

31. Gardiner JL, Liu T. Inverted encoding models reconstruct an arbitrary model response, not the stimulus. *eNeuro* (in press)

32. Ester EF, Sprague TC, Serences JT (2015) Parietal and frontal cortex encode stimulus-specific mnemonic representations during visual working memory. *Neuron* 87: 893-905

33. Ester EF, Sutterer DW, Serences JT, Awh E. (2016) Feature-selective attentional modulations in human frontoparietal cortex *J Neurosci* 35:8188-8199

34. Esterman M, Tamber-Rosenau BJ, Chiu Y-C, Yantis S. (2010) Avoiding non-independence in fMRI data analysis: Leave one subject out. *NeuroImage* 50:572-576

35. Cousineau D. (2005) Confidence intervals in within-subject design: A simpler solution to Loftus & Masson's method. *Quant Meth Psych* 1:42-45

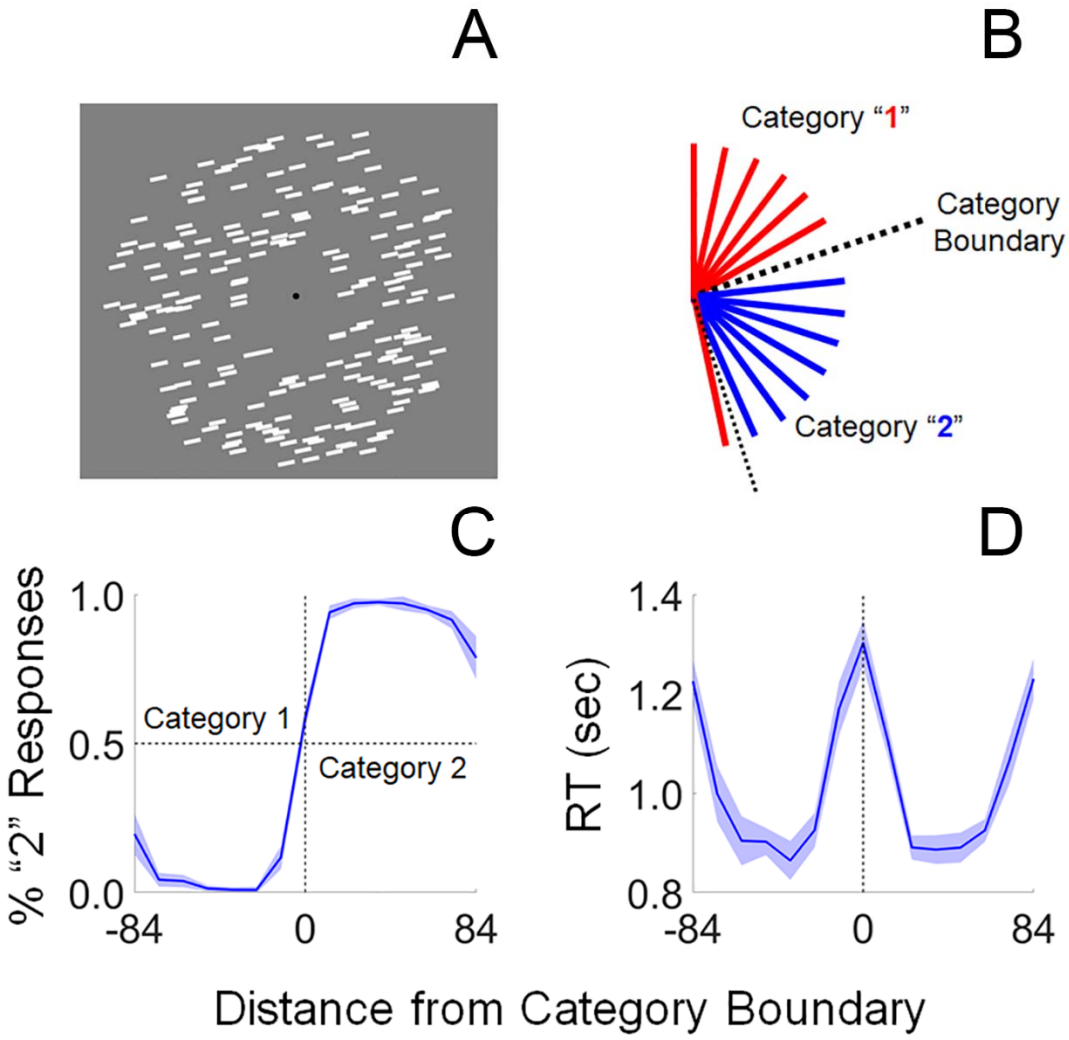

Figure 1. Behavioral Task. (A) Participants viewed displays containing a circular aperture of iso-oriented bars. On each trial, the bars were assigned one of 15 unique orientations from 0-168°. (B) We randomly selected and designated one stimulus orientation as a category boundary (black dashed line), such that the seven orientations counterclockwise from this value were assigned to Category 1 (red lines) and the seven orientations clockwise from this value were assigned to Category 2 (blue lines). (C) After training, participants rarely miscategorized orientations. (D) Response latencies are significantly longer for oriented exemplars near the category boundary (RT = response time; shaded regions in C-D are ±1 within-participant S.E.M.).

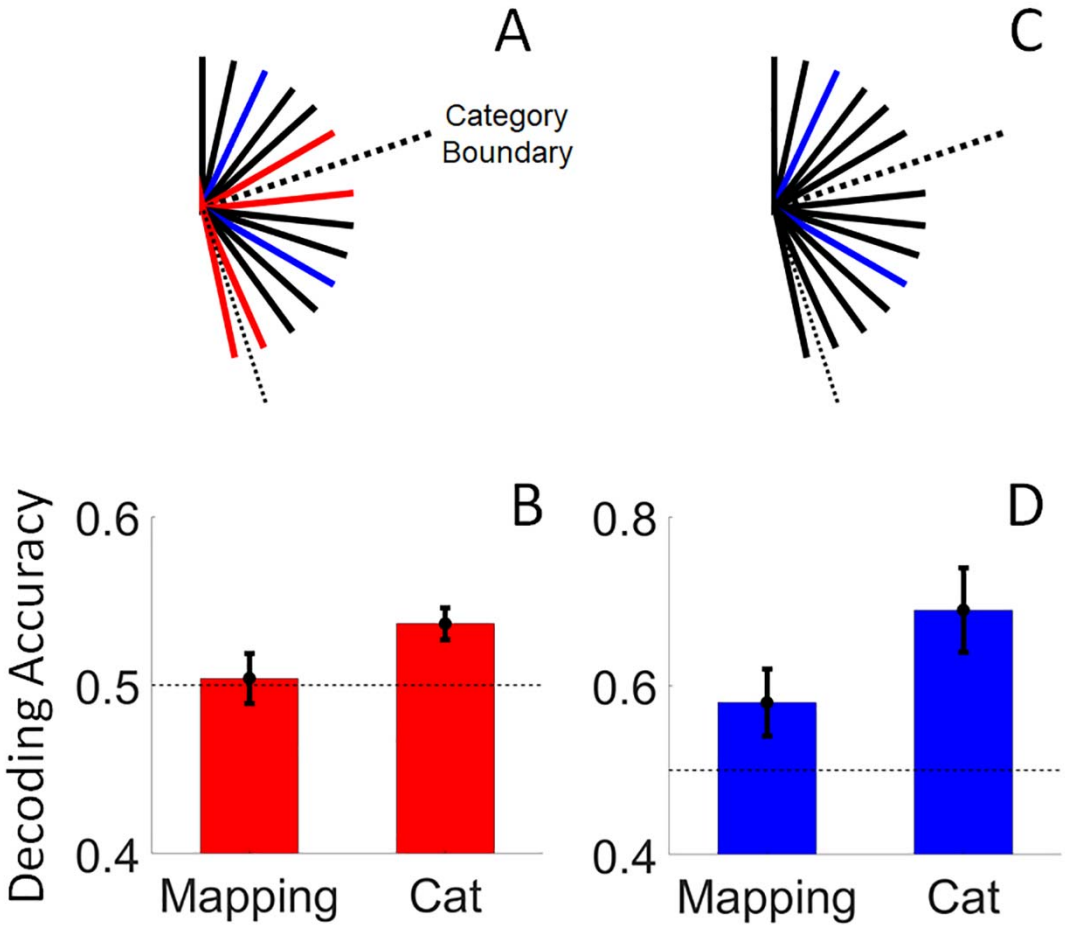

Figure 2. Category Decoding Performance. (A) We trained classifiers on activation patterns evoked by exemplars at the center of each category boundary during the orientation mapping and category discrimination task (blue lines), then used the trained classifier to predict the category membership of exemplars adjacent to the category boundary (red lines). (B) Decoding accuracy was significantly higher during the category discrimination task relative to the orientation mapping task ($p = 0.01$), suggesting that activation patterns evoked by exemplars adjacent to the category boundary became more similar to activation patterns evoked by exemplars at the center of each category during the categorization task. The absence of robust decoding performance during the orientation mapping task cannot be attributed to poor signal or a uniform enhancement of orientation representations by attention, as a decoder trained and tested on activation patterns associated with exemplars at the center of each category (C) yielded above-chance decoding during both behavioral tasks (D). Decoding performance was computed from activation patterns in V1. Error bars depict ±1 S.E.M.

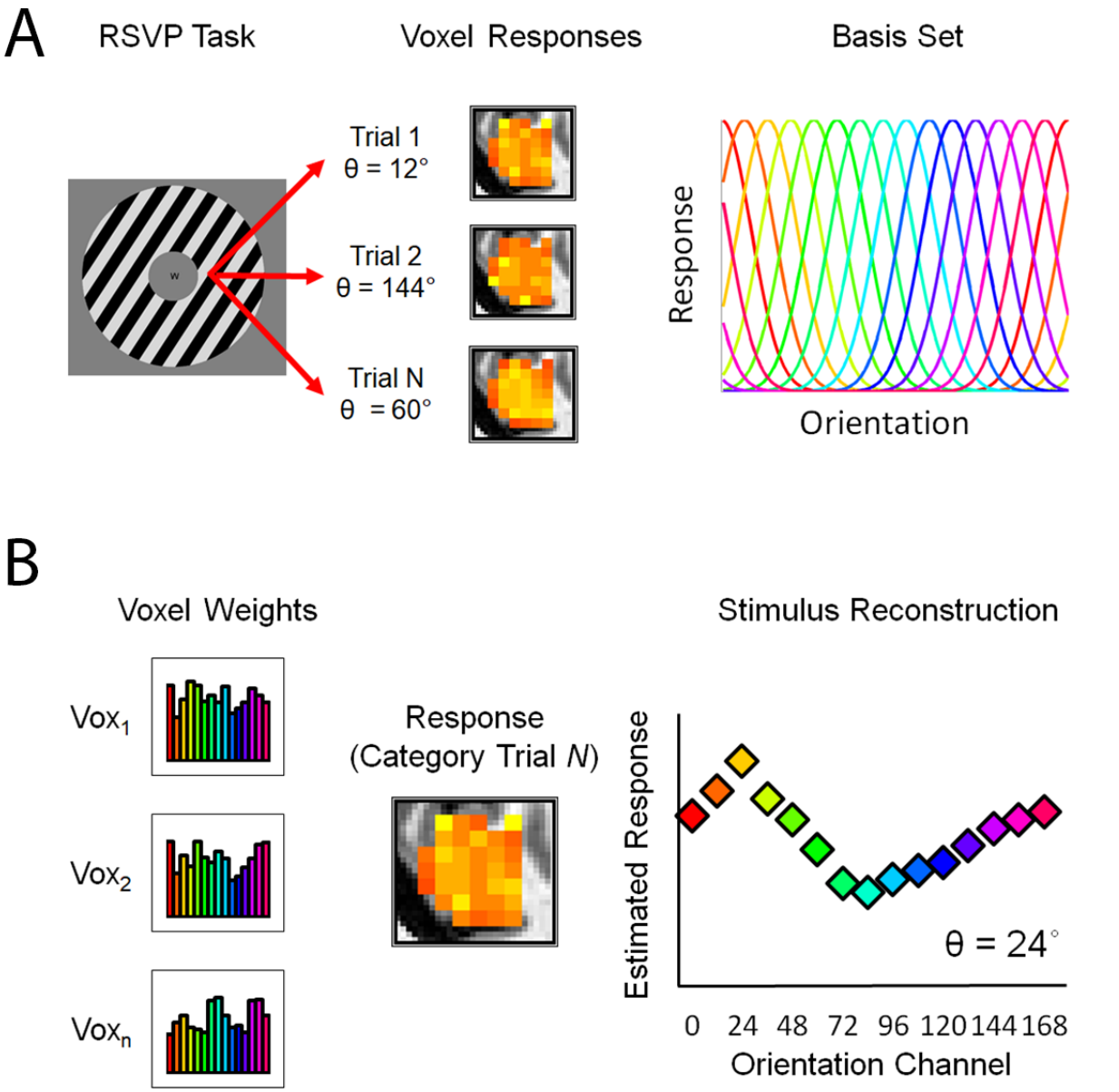

Figure 3. Inverted Encoding Model. (A) In the first phase of the analysis, we estimated an orientation selectivity profile for each voxel retinotopically organized V1-hV4/V3a using data from an independent orientation mapping task. Specifically, we modeled the response of each voxel as a set of 15 hypothetical orientation channels, each with an idealized response function. (B) In the second phase of the analysis, we computed the response of each orientation channel from the estimated orientation weights and the pattern of responses across voxels measured during each trial of the category discrimination task. The resulting reconstructed channel response function (CRF) contains a representation of the stimulus orientation (example; 24 deg), which we quantified via a curve-fitting procedure.

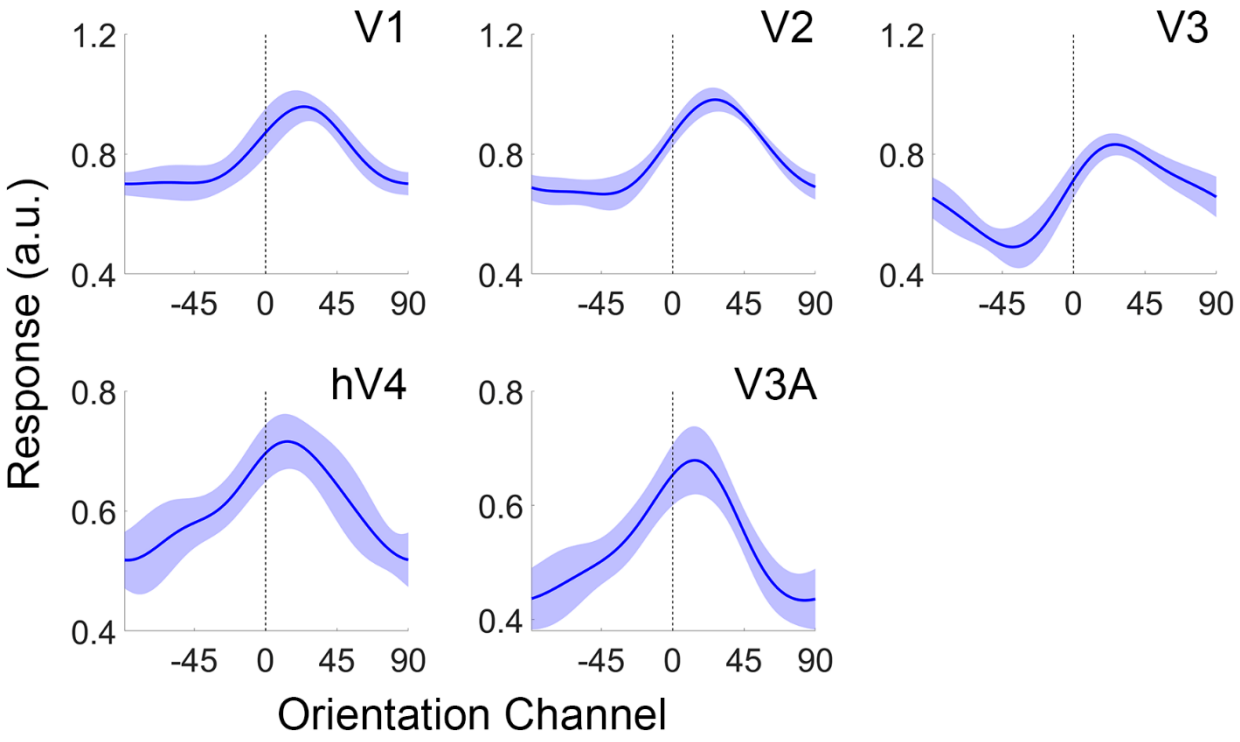

Figure 4. Reconstructed representations of Orientation in Early Visual Cortex. The vertical bar at 0° indicates the actual stimulus orientation presented on each trial. Channel response functions (CRFs) from Category 1 and Category 2 trials have been arranged and averaged such that any categorical bias would manifest as a clockwise (rightward) shift in the orientation representation towards the center of Category B (see Methods and Fig. S1). Shaded regions are ±1 within-participant S.E.M (see Methods). Note change in scale between visual areas V1-V3 and hV4-V3A. a.u., arbitrary units.

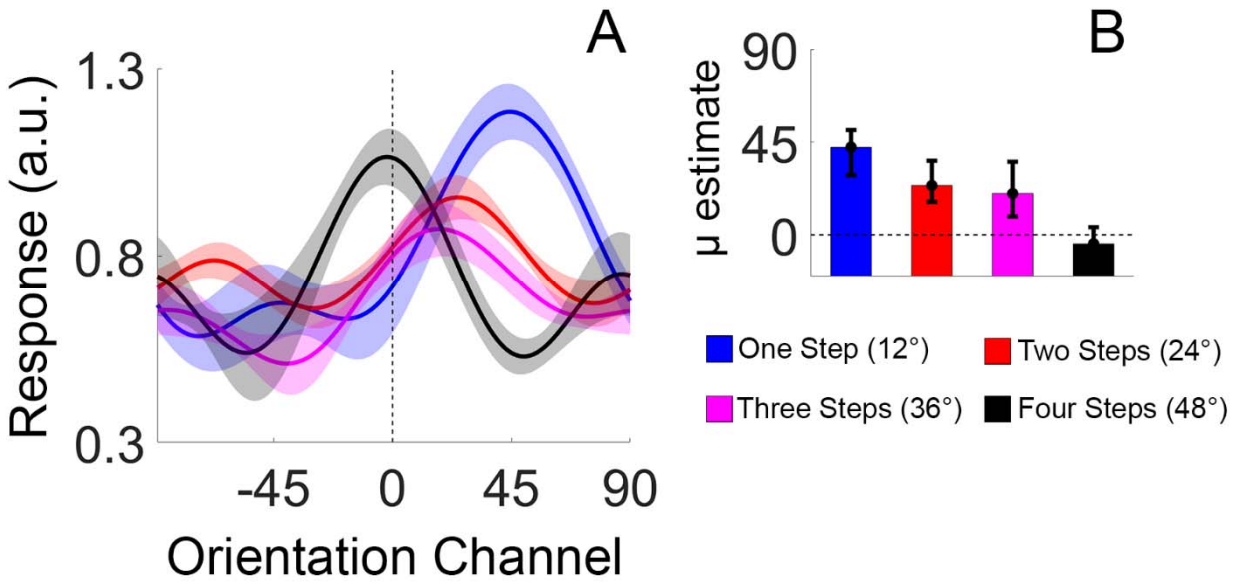

Figure 5. Category Biases Scale Inversely with Distance from the Category Boundary. (A) The reconstructions shown in Fig. 3 by the absolute angular distance between each exemplar and the category boundary. In our case, the 15 orientations were bisected into two groups of 7, with the remaining orientation serving as the category boundary. Thus, the maximum absolute angular distance between each orientation category and the category boundary was 48°. Participant-level reconstructions were pooled and averaged across visual areas V1, V2, and V3 as no differences were observed across these regions. Shaded regions are ±1 within-participant S.E.M. (B) shows the amount of bias for exemplars located 1, 2, 3, or 4 steps from the category boundary (quantified via a curve-fitting analysis). Error bars are 95% confidence intervals. a.u., arbitrary units.

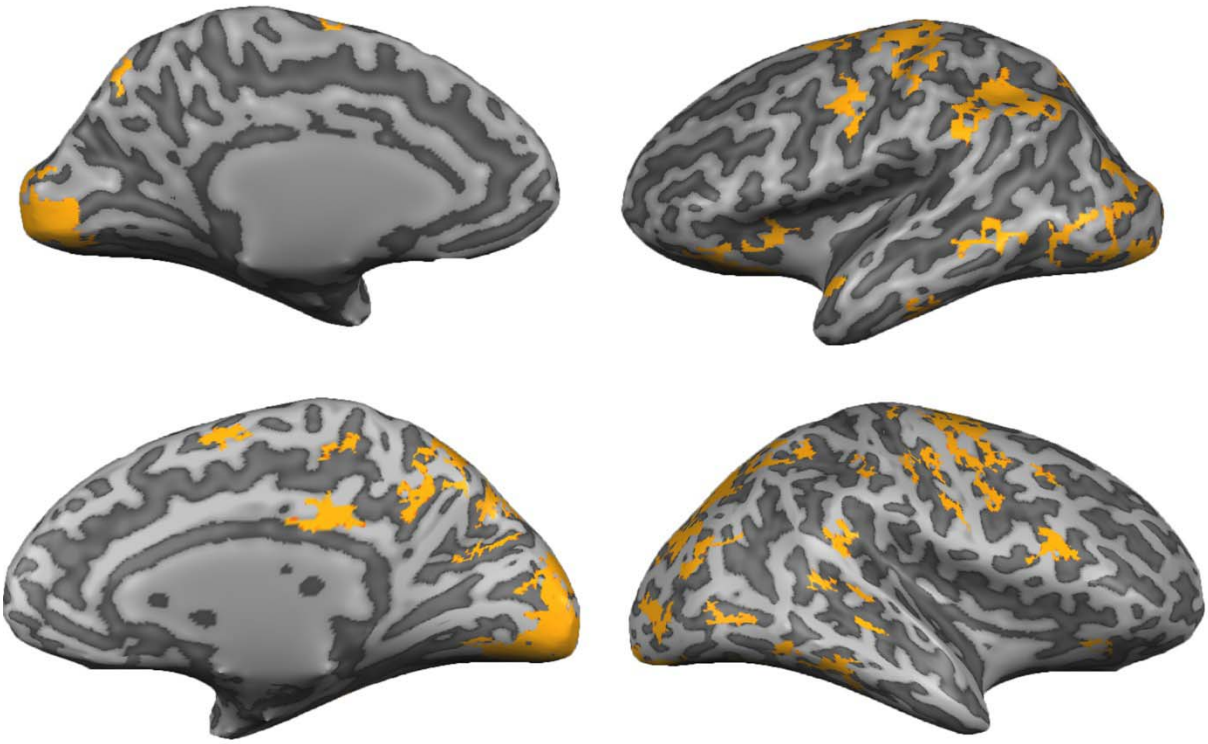

Figure 6. Cortical Areas Supporting Robust Decoding of Category Information. We trained a linear support vector machine to discriminate between activation patterns associated with Category A and Category B exemplars (independently of orientation; see *Searchlight Classification Analysis*; Methods). The trained classifier revealed robust category-specific information in multiple visual, parietal, temporal, and prefrontal cortical areas, including many regions previously associated with categorization (e.g., posterior parietal cortex and lateral prefrontal cortex).

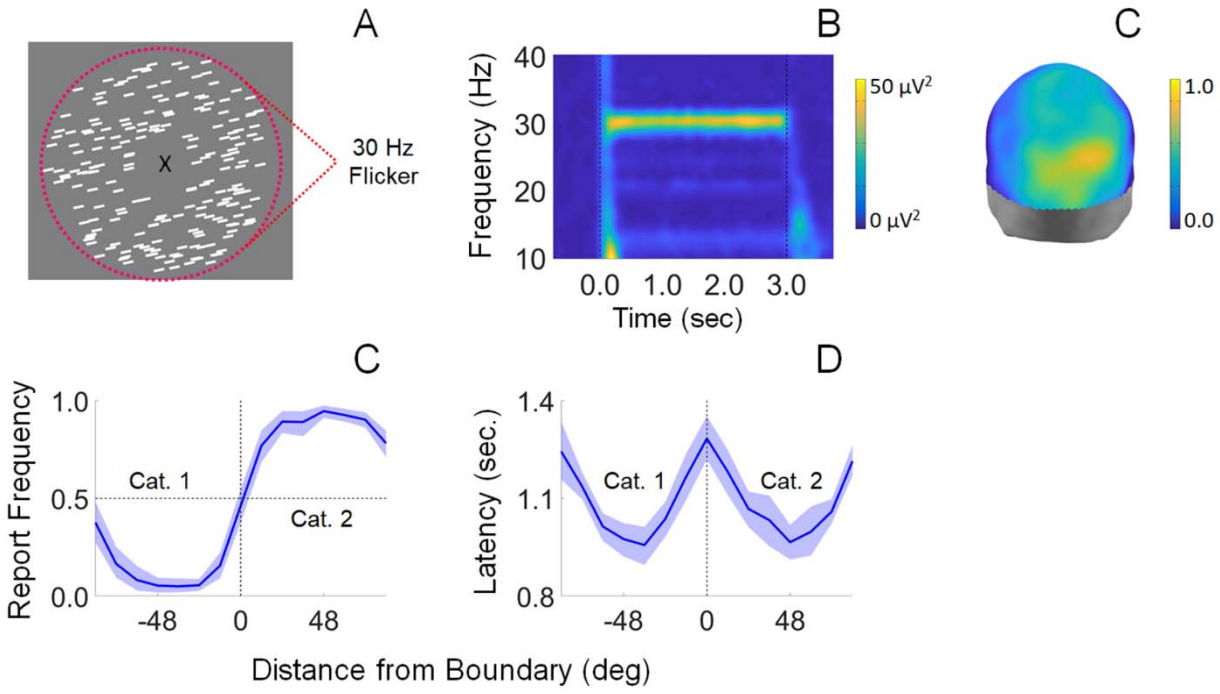

Figure 7. Summary of Experiment 2. (A) Participants viewed displays containing an aperture of iso-oriented bars flickering at 30 Hz. (B) The 30 Hz flicker entrained a frequency-specific response known as a steady-state visually-evoked potential (SSVEP). (C) Evoked 30 Hz power was largest over occipitoparietal electrode sites. We computed stimulus reconstructions (Fig. 7) using the 32 scalp electrodes with the highest power. The scale bar indicates the proportion of participants (out of 27) for which each electrode site was ranked in the top 32 of all 128 scalp electrodes. (D-E) Participants categorized stimuli with a high degree of accuracy; incorrect and slow responses were observed only for exemplars adjacent to a category boundary. Shaded regions are ±1 within-participant S.E.M.

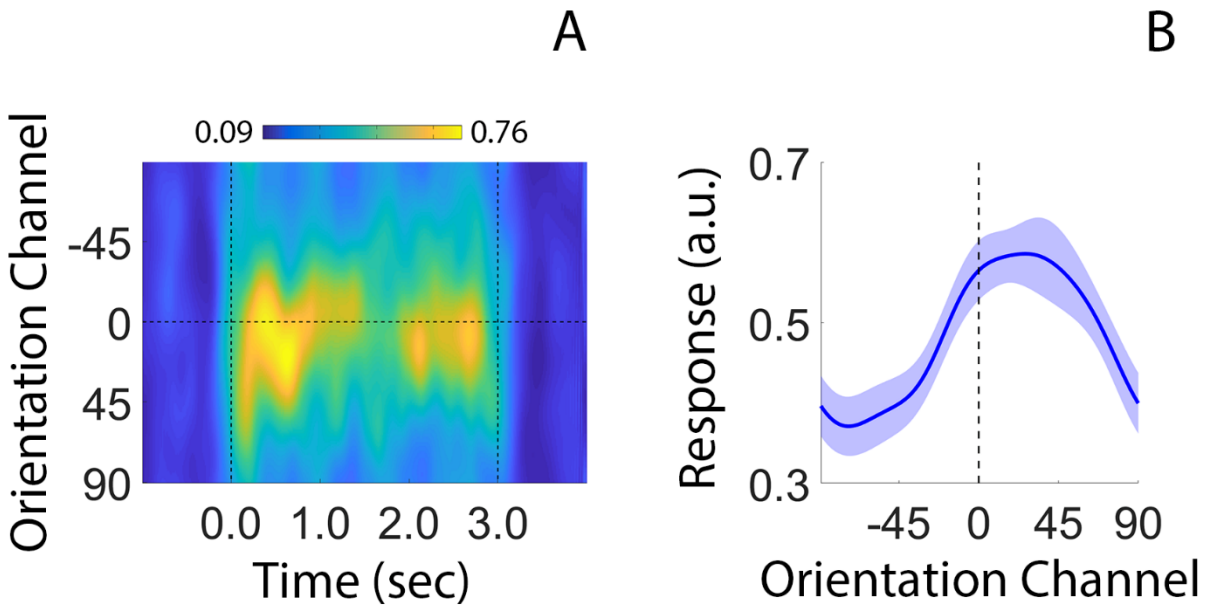

Figure 8. Category Biases Emerge Shortly after Stimulus Onset. (A) Time-resolved reconstruction of stimulus orientation. Dashed vertical lines at time 0.0 and 3.0 seconds mark stimulus on- and offset, respectively. (B) Average channel response function during the first 250 ms of each trial. The reconstructed representation exhibits a robust category bias ($p < 0.01$; bootstrap test). a.u., arbitrary units.