Category knowledge biases sensory representations 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 and other animals 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 and discrimination performance for equally similar exemplars from the same category is reduced. These distortions could result from systematic biases in neural representations that begin at the earlies 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 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 visual cortical areas. Reconstructed representations in several of these areas (V1-V3) were systematically biased by category membership, as indicated by shifts in the representation away from the physical stimulus' orientation and towards the center of the appropriate category. These shifts were strongest for orientations near the category boundary where they would be most beneficial for behavioral performance, predicted participants' overt category judgments, and emerged rapidly after stimulus onset. Collectively, these results indicate that category information can alter information processing at very early stages of the visual stream.

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 reflect changes in how information is represented by sensory neural populations (*9-10*). Here, we sought to test this possibility. We modeled 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 been trained to classify these stimuli into discrete groups. In Experiment 1, 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 category discrimination performance. In Experiment 2, 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 (Fig. 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 1-2 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 (Fig. 1C-D). During each scanning session, participants performed the category discrimination task and an orientation mapping 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).

To evaluate whether category learning alters representations of orientation, we used an inverted encoding model (*11*) to reconstruct a representation of stimulus orientation from activation patterns measured in early visual cortical areas during the category discrimination task. 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 (Fig. 2A). Note that stimulus orientation was irrelevant during this task. We therefore reasoned that voxel-by-voxel responses evoked by each oriented stimulus would be largely uncontaminated by the category membership of each oriented stimulus. 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. Conversely, 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. 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 representations of orientation for each ROI (Fig 2B). Finally, we sorted trial-by-trial reconstructions of stimulus orientation according to category membership such that any bias would manifest as a clockwise (rightward) shift of the reconstructed representation towards the center of Category 2 and quantified biases towards this category using a curve-fitting analysis (Supplementary Materials).

Reconstructed representations of orientation in visual areas V1, V2, and V3 exhibited reliable category biases of 22.13°, 26.65°, and 34.57°, respectively (Fig. 3; $P < 0.05$, bootstrap test, false-discovery-rate [FDR] corrected for multiple comparisons across regions; see Fig S1 for separate reconstructions of Category 1 and Category 2 orientations and Fig S2 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 (Fig S3). 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. We considered the possibility that the categorical biases shown in Fig 3 reflect intrinsic biases in stimulus selectivity in early visual areas (e.g., due to oblique effects; *12*). This possibility is unlikely, as the location of the boundary separating Categories 1 and 2 was randomly selected from the set of 15 possible orientations for each participant (Fig. 1C). That is, there was no consistent relationship between the category boundary and either the horizontal or vertical meridian across participants.

The category biases shown in Fig 3 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 (Fig. S4A). 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; Fig. S4B). 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 (Fig. S4C). 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 (Fig. 4). 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; 12-14*). 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; Fig. 5 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 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; Fig S5). 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 Figs. 3 and 4 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 Figs. 3 and 4 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; Fig 6A). 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*; Fig. 6B). 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. (Fig. 6C). 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 (Fig. 6D-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 method was similar to the modeling approach used in the neuroimaging experiment described above. 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 Fig. 6C). In the second phase of the analysis, we used these weights and responses measured during each trial of the category discrimination task across all electrodes to compute a time-resolved representation of stimulus orientation (Fig. 7A-B).

We reasoned that if the categorical biases shown in Figs 3 and 4 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 (Fig. 7B). 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 (Fig. S6), 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. Ultimately, targeted experiments will be needed to identify the mechanisms responsible for the category biases we have reported here. Nevertheless, to our knowledge these data represent the first demonstration of category biases in population-level representations of stimuli in "sensory" cortical areas.

We have shown that activation patterns in early visual areas reliably signal the category of a to-be-classified orientation (Fig. 5) and that representations of orientation are biased by category membership (Fig. 3). Both observations appear to conflict with results from nonhuman primate research which suggests that sensory cortical areas do not encode categorical information. There are at least two explanations for this disparity. First, 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 (*22-26*). In one example (*22*), extensive training was associated with a functional substitution of human visual area V3a for MT+ in discriminating noisy motion patches. Insofar as monkeys require tens or hundreds of thousands of trials to reach asymptotic performance on a given task, similar changes may explain why category selective signals are found in areas of prefrontal and posterior parietal cortex but not sensory cortex. Second, 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

Full methodological details can be found in the supporting online materials. Below, we provide an overview of the inverted encoding model used to reconstruct and quantify category biases in orientation-selective responses measured using fMRI and EEG.

Overview

A linear inverted encoding model (IEM) was used to recover a quantifiable representation of stimulus orientation from multivoxel activation patterns measured in early visual areas (*10*). 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 14th power). The maximum response of each channel was set to unit amplitude; thus responses are quantified as BOLD z-score units (power in μV^2). 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:

$$B_1 = WC_1$$

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 predicted response of each orientation filter on each trial of the category discrimination task. Trial-by-trial reconstructions in C_2 were 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-score) 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 Reconstructed Representations. To quantify categorical biases in representations of orientation, reconstructions 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).

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 ni 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. Tajima C.I., Tajima S., Koida, K. Komatsu, H., Aihara, K. & Suzuki, H. Population code dynamics in categorical perception. *Sci Rep* **6** 22536 (2016)

23. Tajima, S., Koida, K., Tajima, C.I., Suzuki, H., Aihara, K. Komatsu, H. Task-dependent recurrent dynamics in visual cortex. *eLife* **6** e26868 (2017)

24. 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)

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

26. Itthipuripat, S., Cha, K., Byers, A. & Serences, J.T. Two different mechanisms support selective attention at different phases of training. *PLOS Biology* (in press)

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)

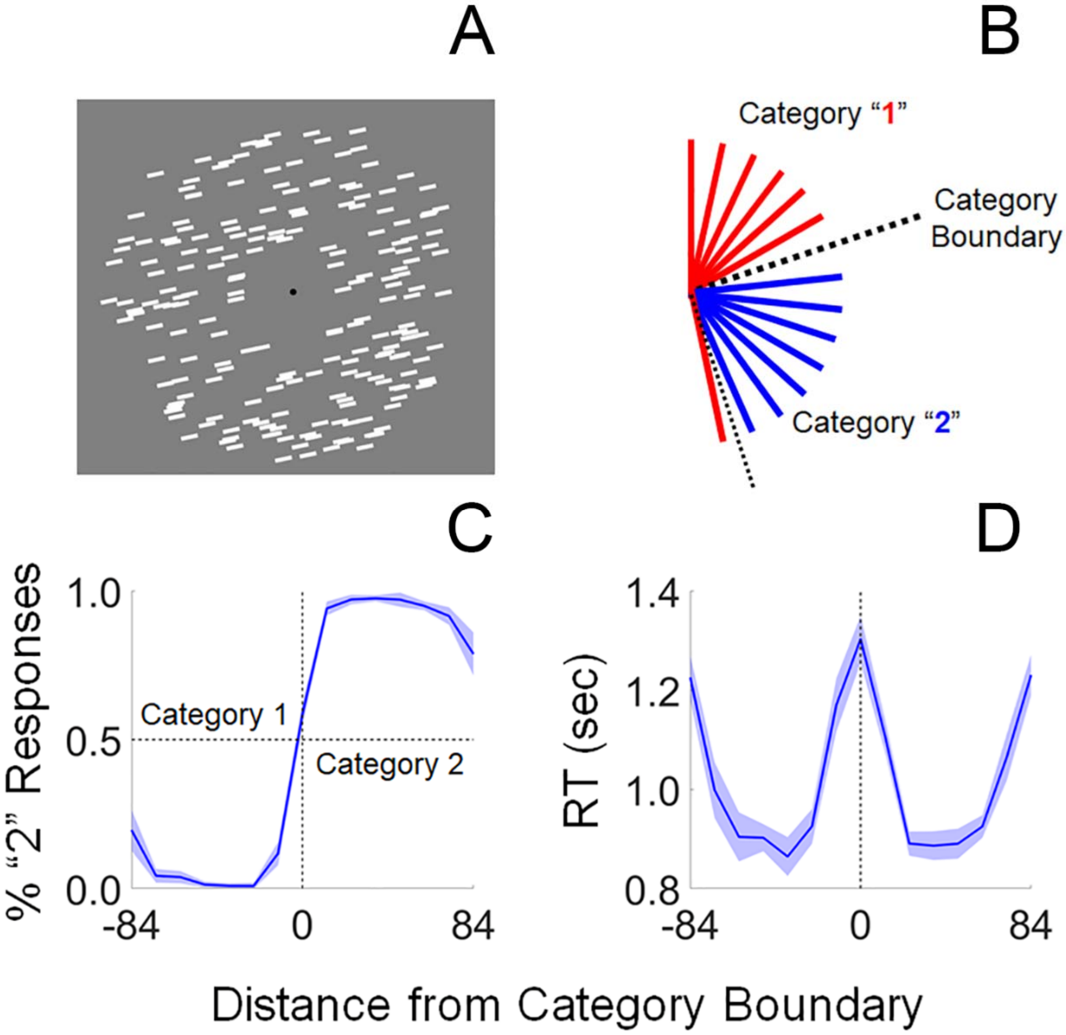

Fig. 1. Behavioral Task. (A) Participants viewed displays containing a circular aperture of iso-oriented bars. On each trial, the bars were assigned one of 15 unique orientations from 0-168°. (B) We randomly selected and designated one stimulus orientation as a category boundary (black dashed line), 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.).

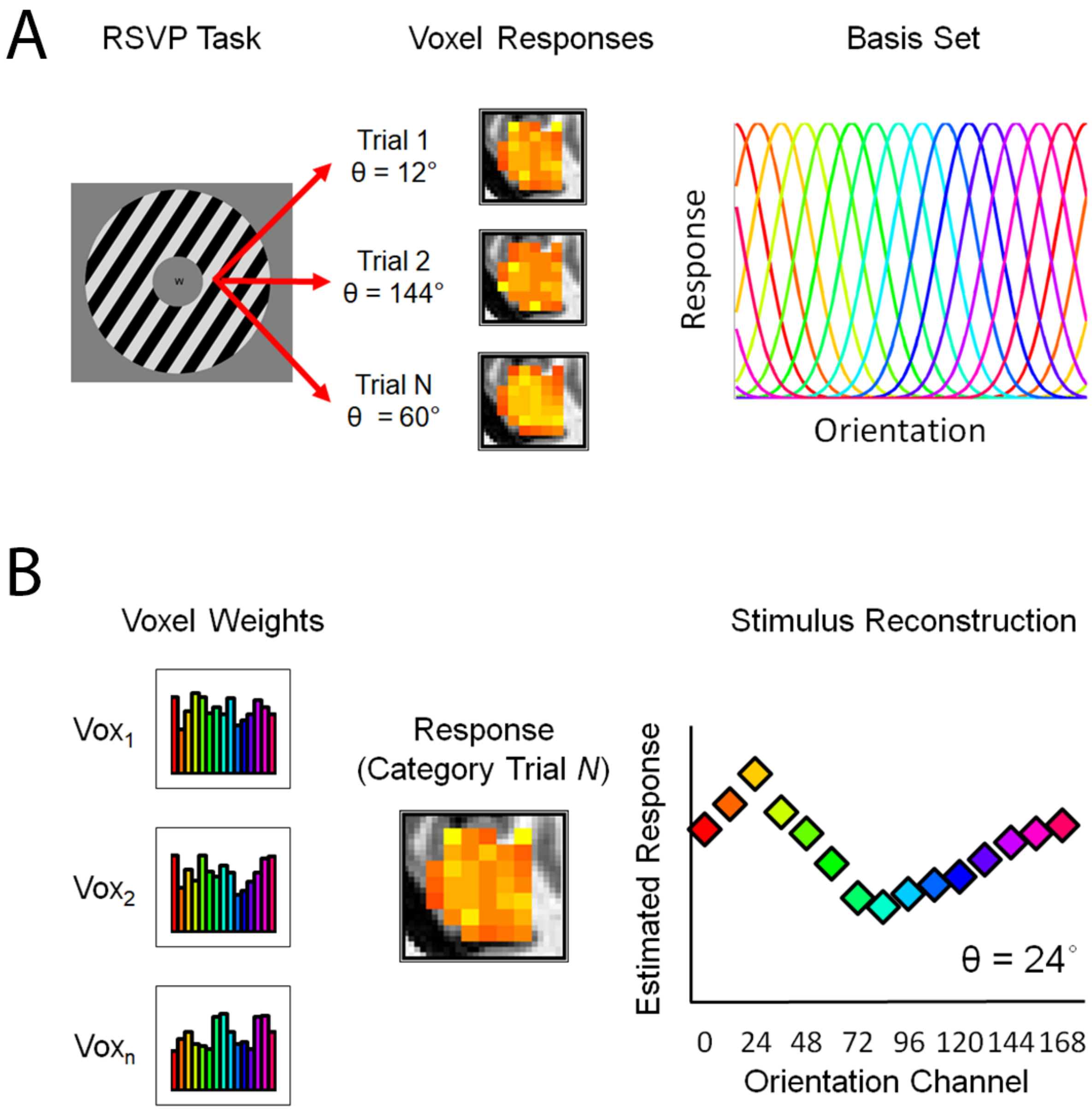

Fig. 2. Inverted Encoding Model. (A) In the first phase of the analysis, we estimated an orientation selectivity profile for each voxel retinotopically organized V1-hV4/V3a using data from an independent orientation mapping task. Specifically, we modeled the response of each voxel as a set of 15 hypothetical 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.

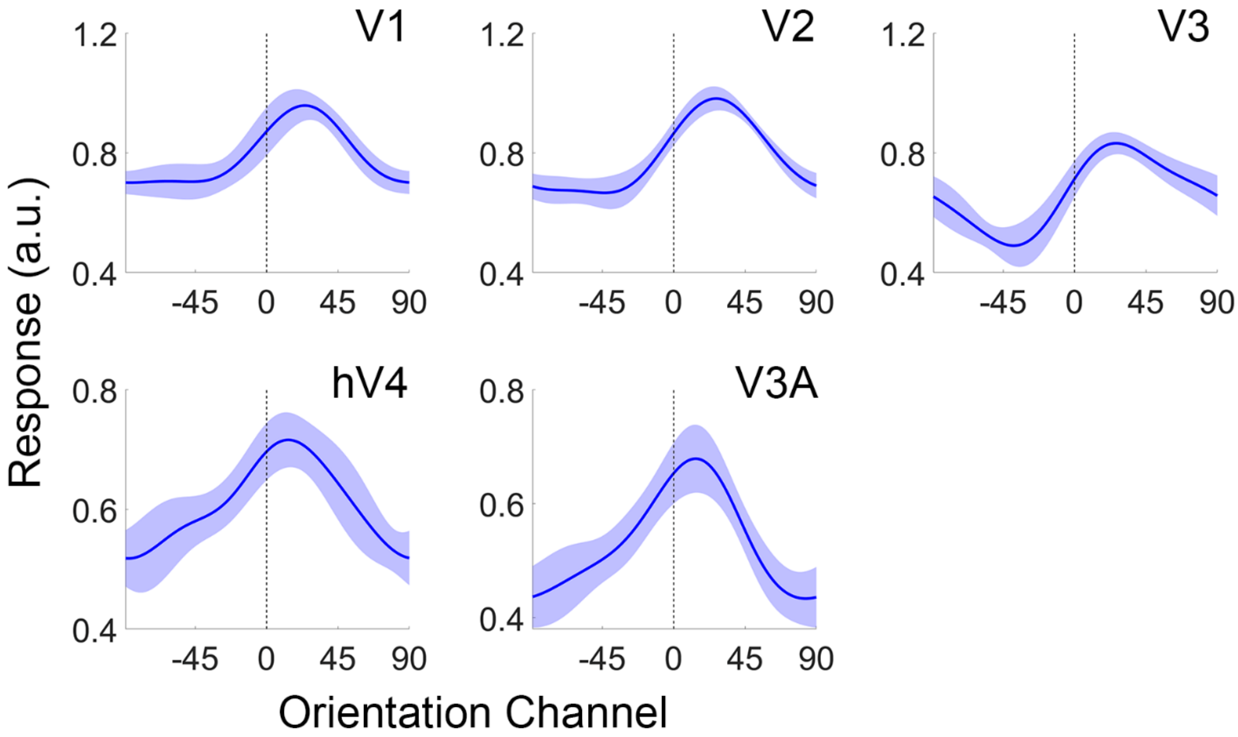

Fig. 3. Reconstructed representations of Orientation in Early Visual Cortex. The vertical bar at 0° indicates the actual stimulus orientation presented on each trial. Data from Category 1 and Category 2 trials have been arranged and averaged such that any categorical bias would manifest as a clockwise (rightward) shift 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.

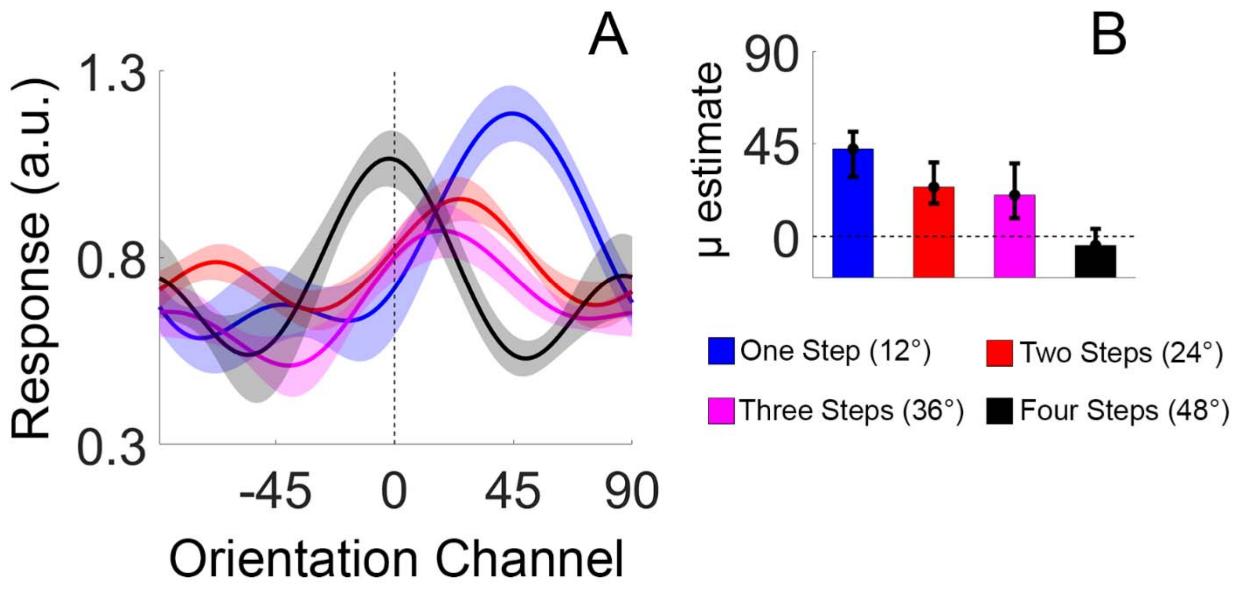

Fig. 4. Category Biases Scale Inversely with Distance from the Category Boundary. (A) The reconstructions shown in Fig. 3 by the absolute angular distance between each exemplar and the category boundary. In our case, the 15 orientations were bisected into two groups of 7, with the remaining orientation 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.

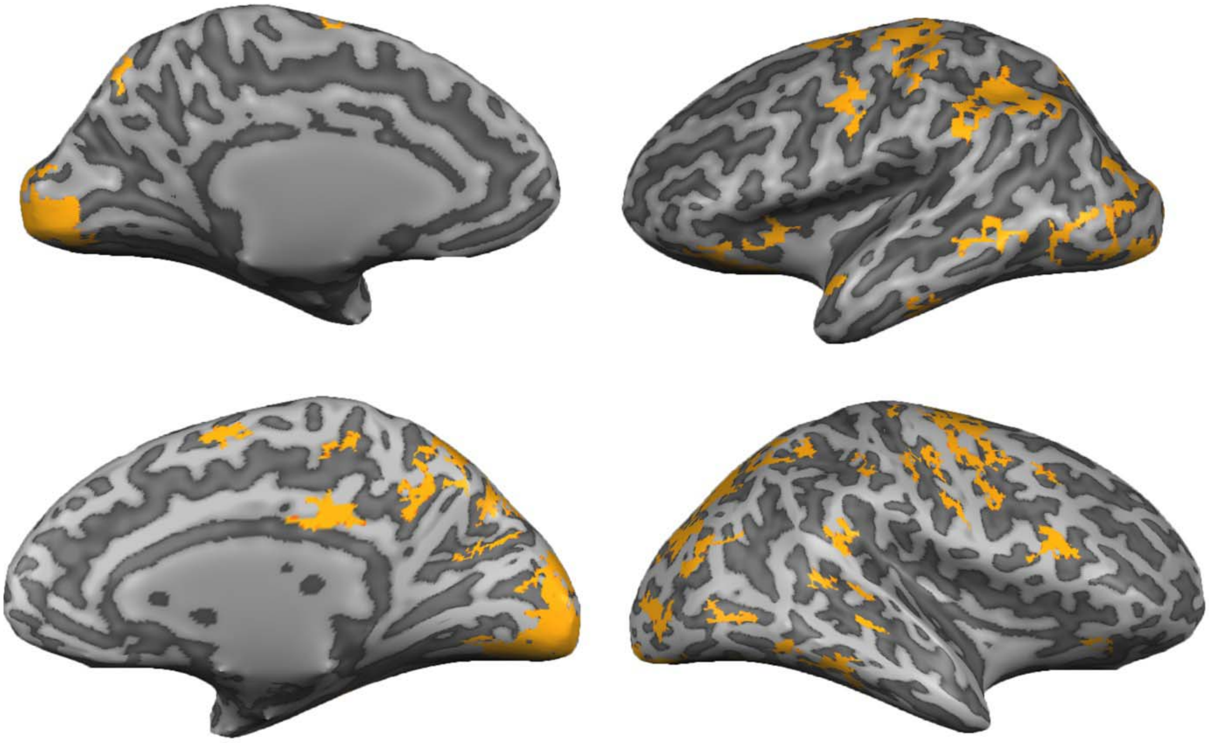

Fig. 5. Reconstructions of Stimulus Orientation in Cortical Areas Encoding Category Information. We trained a linear support vector machine to discriminate between activation patterns associated with Category A and Category B exemplars (independently of orientation; see *Searchlight Classification Analysis*; Methods).

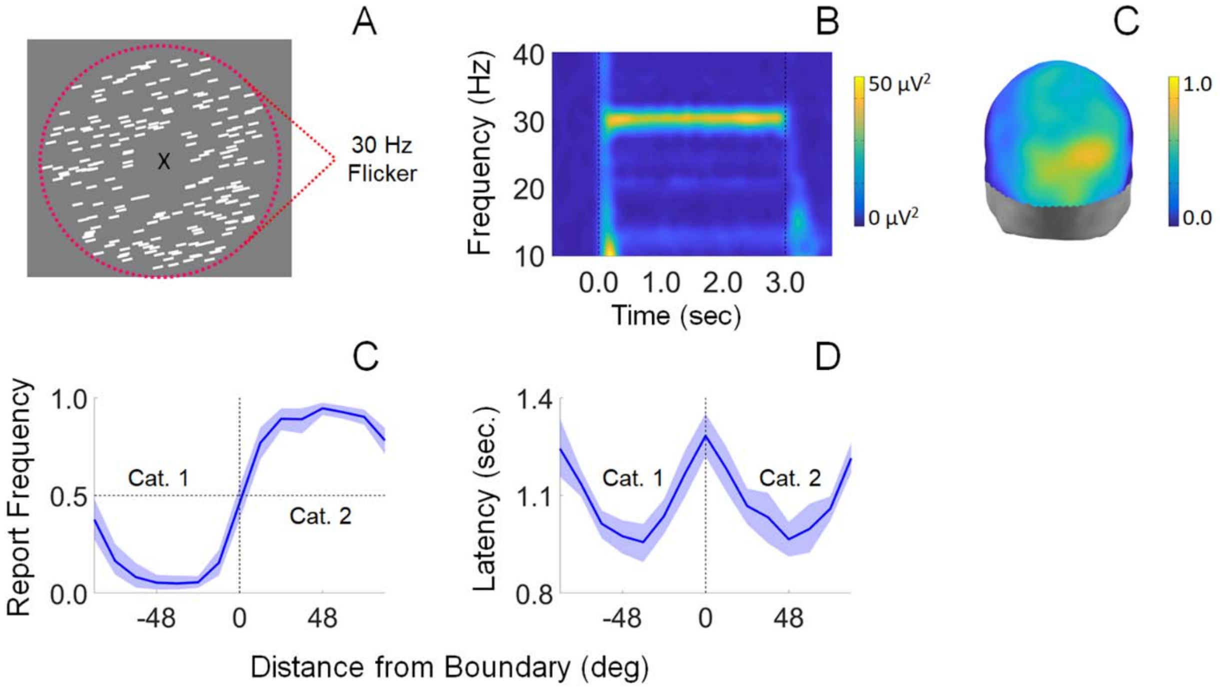

Fig. 6. Summary of Experiment 2. (A) Participants viewed displays containing an aperture of iso-oriented bars flickering at 30 Hz. (B) The 30 Hz flicker entrained a frequency-specific response known as a steady-state visually-evoked potential (SSVEP). (C) Evoked 30 Hz power was largest over occipitoparietal 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.

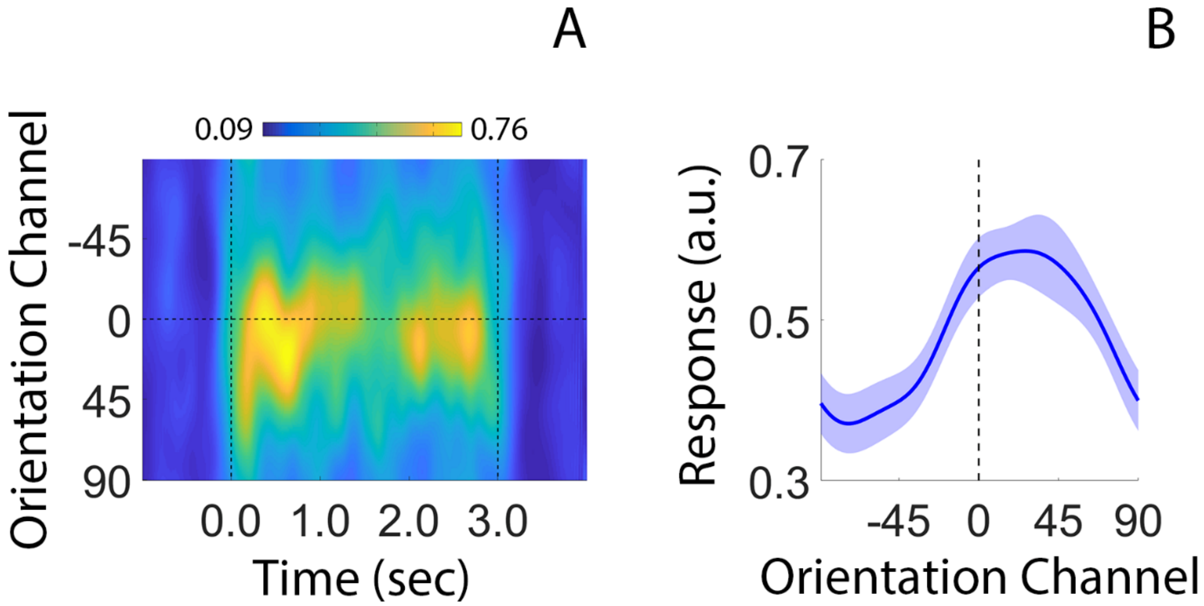

Figure 7. Category Biases Emerge Shortly after Stimulus Onset. (A) Time-resolved reconstruction of stimulus orientation. Dashed vertical lines at time 0.0 and 3.0 seconds mark stimulus on- and offset, respectively. (B) Average channel response function during the first 250 ms of each trial. The reconstructed representation exhibits a robust category bias ($p < 0.01$; bootstrap test). a.u., arbitrary units.