Category learning biases sensory representations in human visual cortex

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

[1]*Department of Psychology and FAU Brain Institute, Florida Atlantic University*
[2]*Department of Psychology, New York University*
[3]*Department of Psychology, Neurosciences Graduate Program, and Kavli Institute for Brain and Mind, University of California, San Diego*

Abstract Word Count: 126

Article Word Count: 3184

Reference Count: 33

Figure Count: 6

*Correspondence:

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

Acknowledgments: Funding provided by NIH R01 EY025872 (J.T.S.) and a James S. McDonnell Foundation award (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 declare no competing interests.

Data Availability: All data and analytic software are publicly available on the Open Sciences Framework at https://osf.io/xzay8/

Abstract

Category learning distorts perceptual space by enhancing the discriminability of physically similar yet categorically distinct exemplars. These distortions could in part reflect changes in how sensory neural populations selective for category-defining features encode information. Here, we tested this possibility by using fMRI and EEG to quantify the feature-selective information content of signals measured in early visual cortical areas after participants learned to classify a set of oriented stimuli into discrete categories. Reconstructed representations of orientation in early visual areas were systematically biased by category membership. These biases were strongest for orientations near the category boundary, predicted category discrimination performance, and emerged rapidly after stimulus onset, suggesting that category learning can produce significant changes in how neural populations in early visual areas respond to incoming sensory signals.

Categorization is 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 based on subtle differences in color, shape, size, and texture (Fig. 1A). 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*).

In principle, perceptual distortions following category learning could reflect changes in how information is "read out" from sensory neural populations (e.g., *6*) and/or changes in how information is represented by sensory neural populations (*7*-*8*). Here, we evaluated these alternatives by training participants to classify a set of oriented stimuli into two groups. After training, we used noninvasive measurements of human brain activity (fMRI and EEG) to visualize and quantify population-level representations of oriented stimuli while participants performed a category discrimination task. 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 are correlated with trial-by-trial variability in overt category judgments and are largest for orientations adjacent to the category boundary where they would be most helpful in facilitating task performance. In Experiment 2, we used

EEG to generate time-resolved representations of to-be-categorized orientations and show that categorical biases manifest as early as 125 ms after stimulus onset, suggesting that the intent to categorize a stimulus modulates how neural populations in early visual areas respond to incoming sensory signals.

Results

Experiment 1 - fMRI

Inspired by earlier work in non-human primates (*9-10*), we trained eight human volunteers to categorize a set of orientations into two groups, A and B. The stimulus space comprised a set of 15 oriented stimuli, spanning 0-168° in 12° increments (Fig 1B). For each participant, we randomly selected one of these 15 orientations as a category boundary such that the seven orientations anticlockwise to the boundary were assigned membership in Category A and the seven orientations clockwise to the boundary were assigned membership in Category B (Fig 1B-1C). After a one-hour training session, participants could categorize the stimuli with near-perfect accuracy (Fig. 1D). Each participant then completed two separate two-hour fMRI scanning sessions. During each 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 (Fig S1A). 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). Each participant's category boundary was kept constant across all testing sessions (behavioral training and scanning).

To evaluate whether category learning alters representations of orientation, we used an inverted encoding model approach (*11-12*) to reconstruct a representation of the stimulus' orientation during each trial of the category discrimination task from visual cortical fMRI activation 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 (Fig.

2A). Note that stimulus orientation was irrelevant in the orientation mapping task, so the orientation weights estimated using data from this task should be largely unaffected by extraneous factors such as stimulus category membership and/or mechanisms of selective attention. In the second phase of the analysis, we reconstructed trial-by-trial representations of stimulus orientation by combining information about these weights and the observed pattern of activation across voxels measured during each trial of the category discrimination task, resulting in single-trial representations of visual orientation for each ROI, with peaks closely tracking the presented orientation (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 representation towards the center of Category B (Fig. 2C) and quantified biases using a curve-fitting analysis (see Methods).

Reconstructed representations of orientation in visual areas V1, V2, and V3 exhibited reliable category biases of 22.33°, 26.81°, and 34.86°, 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 A and Category B trials). A logistic regression analysis established that these biases were strongly correlated with variability in overt category judgments (Fig. 3, insets). Similar, though less robust trends were also evident in hV4v and V3a (mean shifts of 11.54° and 8.37°, respectively; $p > 0.19$). One trivial possibility is that these effects reflect intrinsic biases in stimulus selectivity in early visual areas. This is unlikely as the location of the boundary separating Categories A and B was randomly selected from the set of 15 possible orientations for each participant (Fig. 1C). Moreover, no biases were observed in reconstructions of stimulus orientation computed from the orientation mapping task, as might be

expected if these biases are an intrinsic property of the visual system or an artifact of our analytical approach (Fig S2).

We propose that the biases shown in Fig. 3 reflect context-dependent changes in how visual areas process or represent sensory information during the orientation mapping and category discrimination tasks. We sought additional evidence for this alternative by reversing the IEM analysis shown in Fig 2. Specifically, we used data from the category discrimination task to estimate a set of orientation weights for each MRI voxel, then used these weights to reconstruct a representation of stimulus orientation on each trial of the orientation mapping task. We reasoned that if the categorical biases shown in Fig. 2 are caused by context-dependent changes in representations of sensory information during the orientation mapping and category discrimination tasks, then reconstructions of stimulus orientation during the orientation mapping task computed from weights estimated using data from the category discrimination task should exhibit a bias towards the incorrect category (Category A). This is precisely what we observed (Fig S3): reconstructions of stimulus orientation during the orientation mapping task in V1, V2, and V3 exhibited strong biases towards the center of Category A (average shifts of 55.09°, 46.56°, and 25.10° for V1-V3, respectively; all FDR-corrected p-values ≤ 0.05).

The 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. 4A). 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 (or equivalently, an acquired similarity between exemplars adjacent to the category boundary and exemplars near the center of each category) following training (Fig. 4B). 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. 4C). 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. 4D). As expected, reconstructed representations of orientations adjacent to the category boundary were strongly biased by category membership ($\mu = 43°$, $p < 0.05$, FDR-corrected for multiple-comparisons across exemplar-boundary distances), while reconstructed representations of orientations at the center of each category exhibited no signs of bias ($\mu = -4°$, $p > 0.56$). Reconstructed representations of orientations located between these extremes exhibited modest but reliable category biases (22° and 19° for exemplars two and three steps from the boundary, respectively; both $p < 0.05$), and reconstructed representations for orientations located one, two, or three steps from the category boundary all exhibited larger categorical biases relative to orientations located four steps from the category boundary (all FDR-corrected p-values < 0.005; see inset of Fig 4D). Biases were also larger for orientations located adjacent to the category boundary than those located two or three steps away from the category boundary (both FDR-corrected p-values <

0.02). Thus, categorical 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 (*6-8, 13-14*), inferotemporal cortex (*15*), posterior parietal cortex (*9-10*), and lateral prefrontal cortex (*16*). 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 A and Category B (independently of stimulus orientation; Fig. 5 and Methods). We successfully reconstructed 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 S4). This is perhaps unsurprising as representations in many mid-to-high order cortical areas are strongly task-dependent (e.g., *17*). 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

If category learning induces context-dependent shifts in how sensory neural populations encode low-level visual features, then these changes should be evident almost immediately after stimulus onset. In a second experiment, we tested this idea by examining categorical biases over the first few hundred milliseconds of each category discrimination trial using EEG. Specifically, we recorded EEG while a new group of 10 volunteers performed variants of the orientation mapping and categorization tasks used in the fMRI experiment (Fig. 6A). 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, *18*). 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 and longer transmission times along axonal fibers (*19*). Thus, by using a relatively high flicker rate of 30Hz, most of the SSVEP is likely generated locally within occipital visual cortex. An analysis of the spatial distribution of 30Hz power across scalp electrode sites supports this assumption (Fig. S5).

We computed the power and phase of the 30Hz SSVEP response across each 3,000 msec trial (Fig. 6B; Methods) and then used these values to reconstruct a time-resolved representation of stimulus orientation (*20*). 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 (those with the highest power at 30 Hz; see Fig S5). 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. 6C). We reasoned that if categorical biases emerge gradually over the course of each trial, then reconstructed representations measured during the early part of each trial should resemble the physical orientation of the stimulus regardless of category boundary. To test this possibility, we computed a set of temporally averaged reconstructions from 0 to 250 ms after stimulus onset in 125 ms increments (Figure 6D) and estimated the center of each reconstruction using a curve fitting analysis. Categorical biases were observed across the first 500 ms of each trial, including a temporal interval spanning 0-125 ms (Fig. 4D), suggesting that the intent to categorize a stimulus modulates how neural populations in early visual areas respond to incoming sensory signals.

Discussion

Learning to categorize a set of stimuli based on a low-level feature property such as luminance or hue distorts perceptual representations of that property by increasing the discriminability of physically similar yet categorically distinct stimuli (*6*) and minimizing the discriminability of equally similar stimuli from within the same category (*7-8*). Critically, these distortions could reflect changes in how sensory neural populations selective for the task-relevant feature encode this information, changes in how information is "read out" from these populations, or some mixture of both. Collectively, the findings reported here provide strong support for the first of these alternatives. Using feature reconstruction techniques, we show 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 are correlated with overt category judgments and are adaptive insofar as they are largest for highly confusable exemplars adjacent to a category boundary and smaller for less confusable exemplars further from the boundary.

Several mechanisms could be responsible for the categorical biases reported here. One possibility is that category learning produces context-dependent changes in the spectral preferences of sensory neural populations, as suggested by reports indicating that single neurons and populations of neurons can alter their spectral tuning following changes in task demands (e.g., *21-22*). A second possibility is that biases emerge from recurrent interactions between top-down and bottom-up signals encoded by distinct neural populations. This alternative is supported by the results of one recent study (*23*) where a concurrent network model based on hierarchical Bayesian estimation successfully captured modest categorical biases in representations of hue computed from neural populations in inferotemporal cortex (*23-24*). However, the recursive

nature of this model predicts that categorical biases in representations of hue (or any other low-level visual feature) should emerge gradually over time (on the order of several hundred milliseconds). This prediction is difficult to reconcile with the results of Experiment 2, where robust categorical biases were observed within 125 ms of stimulus onset. 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 (*25*-*27*). In one example (*25*), 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 the so-called 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 A and Category B throughout the experiment. Depending on one's perspective, 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 strong motor signals would be present in early visual areas based on existing data. Nevertheless, to our knowledge this is the first demonstration that mechanisms of categorization modulate feature-selective representations at the earliest stages of the visual system.

Our results have several important implications. First, invasive electrophysiological studies in non-human primates have identified responses that discriminate between exemplars from different categories in prefrontal and posterior parietal cortex (e.g., *6*, *9*, *16*), but it is always unclear what effects (if any) these signals have on *representations* of to-be-categorized stimuli. Our results suggest that category learning is associated with context-dependent changes in how the brain represents sensory information, and that these effects reach as far back as the earliest stages of the visual cortical processing stream in humans. Second, our observation of categorical biases in visual areas V1-V3 is inconsistent with empirical findings and models indicating that category-selective signals manifest only at intermediate-to-late stages of the visual processing hierarchy (*6*). More broadly, our findings add to a growing set of observations

(e.g., *21*) suggesting that information processing in early visual cortical areas is incredibly flexible and can be adapted to maximize performance on an observer's task.

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. Engel, T.A., Chaisangmongkon, W. Freedman, D.J. & Wang, X-J. Choice-correlated activity fluctuations underlie learning of neuronal category representation. *Nat Commun* **6** 6454 (2015)

7. 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)

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

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

10. Sarma, A., Masse, N.Y., Wang, X-J & Freedman, D.J. Task specific versus generalized mnemonic representations in parietal and prefrontal cortices. *Nat Neurosci* **19** 143-149 (2016)

11. Brouwer, G.J. & Heeger, D.J. Decoding and reconstructing color from responses in human visual cortex. *J Neurosci* **29** 13992-14003 (2009)

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

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

16. 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)

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

18. Silberstein, R.B., Ciorciari, J., Pipingas, A. Steady-state visually evoked potential tomography during the Wisconsin card sorting test. *Electroencephalgor Clin Neurophysiol* **96** 24-35 (1995)

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

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

21. 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)

22. David, S.V., Fritz, J.B. & Shamma, S.A. Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proc Natl Acad Sci USA* **109** 2144-1249 (2011)

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

24. 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).

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

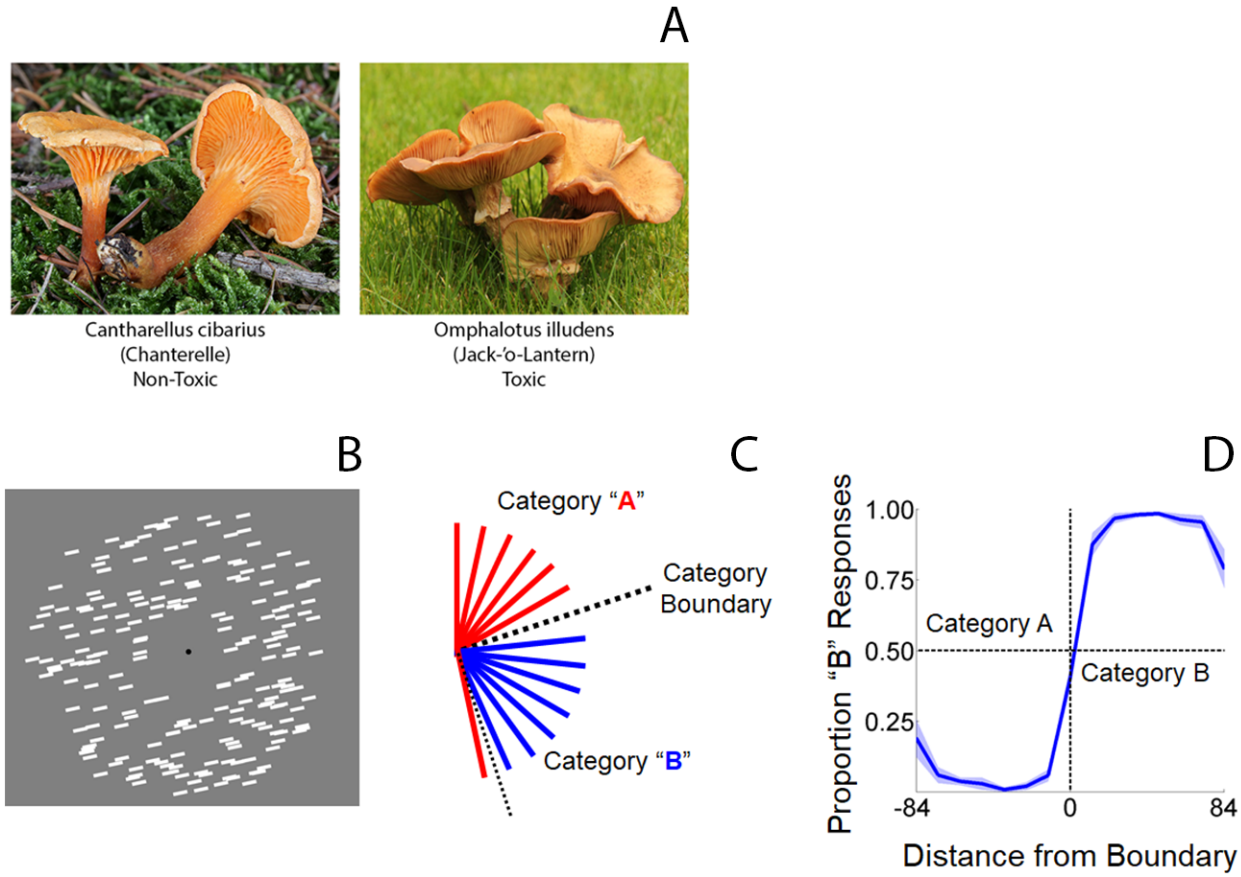

Fig. 1. Behavioral Task. (A) An example of physically similar yet categorically distinct stimuli. The edible and poisonous mushrooms are best distinguished by variations in low-level visual properties such as hue, and texture. (B) In the category discrimination task, participants viewed displays containing a circular aperture of iso-oriented bars. On each trial, the bars were assigned one of 15 unique orientations from 0-168°. (C) 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 A (red lines) and the seven orientations clockwise from this value were assigned to Category B (blue lines). (D) After training, participants rarely miscategorized orientations (shaded region ±1 within-participant S.E.M.).

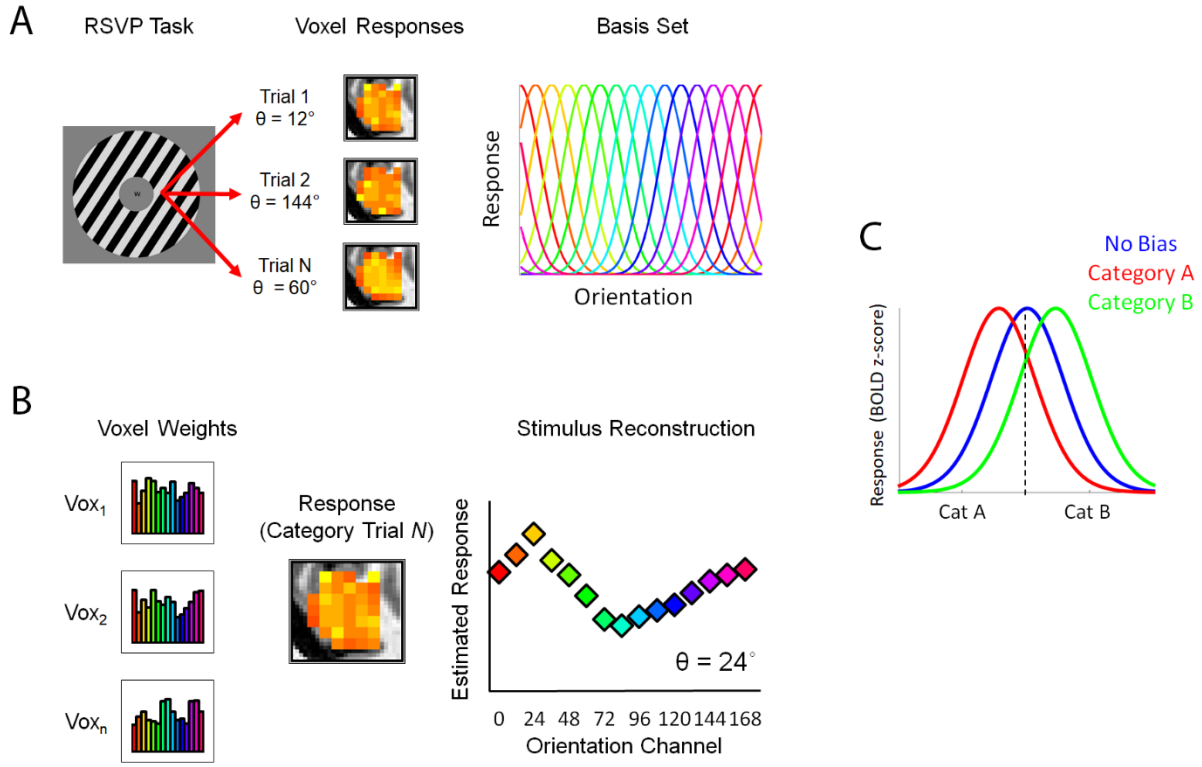

Fig. 2. Schematic Overview of the Inverted Encoding Analysis. (A) In the first phase of the analysis, we estimated an orientation selectivity profile for each voxel in each retinotopically organized visual and parietal region we examined. Participants performed a "rapid serial visual presentation" (RSVP) task where they were instructed to detect and reported the identity of target letters ("X" or "Y") that appeared in a rapid sequence in the center of the display. On each trial, a task-irrelevant, square, wave, phase-reversing grating with one of 15 orientations (0-168° in 12° increments) was presented in the periphery. We modeled the responses of each voxel to each orientation as a set of 15 hypothetical orientation channels, yielding a weight matrix that describes the orientation selectivity of each voxel. Note that stimulus orientation was irrelevant to the participants' task. This was done to minimize the influence of factors such as category learning and selective attention on multivoxel activation patterns measured during this task. (B) Using the weights estimated in (A), we inverted the analysis and computed the response of each

orientation channel from multivoxel activation patterns measured during each trial of the category discrimination task. (C) We hypothesized that representations of stimulus orientation would be shifted according to their category membership. To evaluate this possibility, we circularly-shifted trial-by-trial reconstructions of stimulus orientation to a common center (0°), then aligned these centered reconstructions with the participant's category boundary such that Category A exemplars were located anticlockwise of 0° and Category B exemplars were located clockwise of 0°.

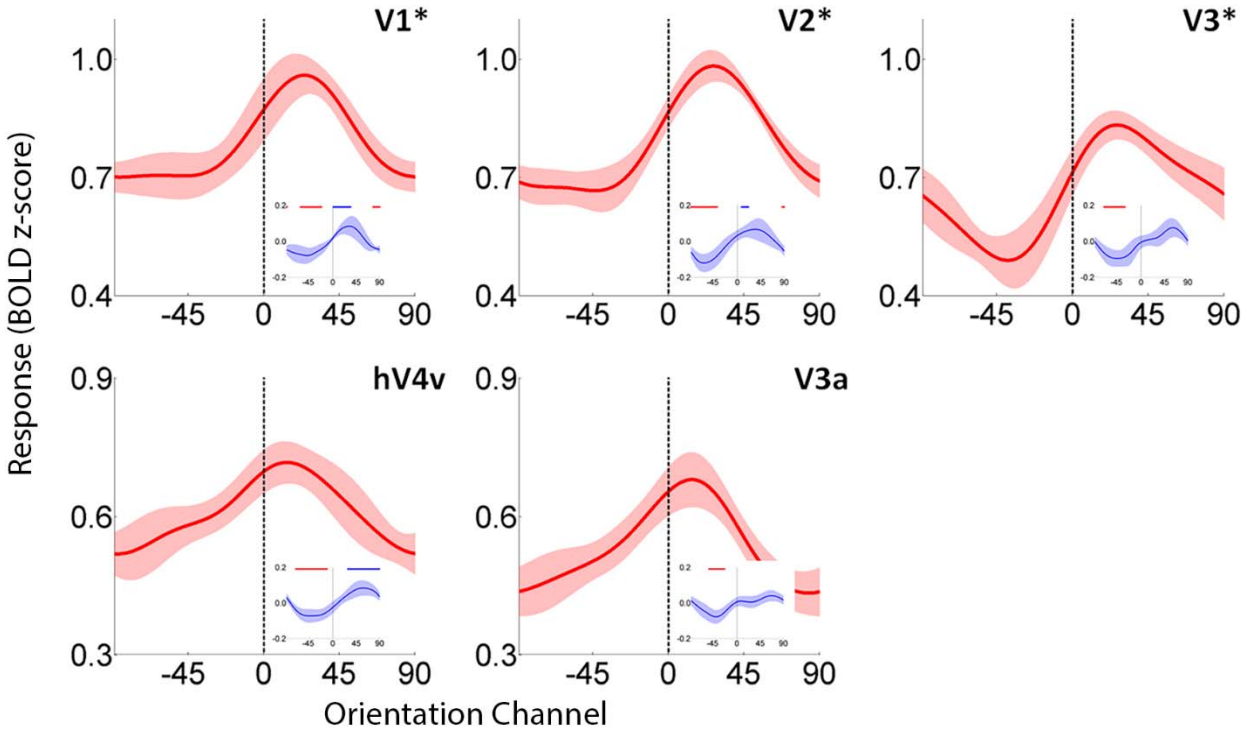

Fig. 3. Reconstructed representations of Orientation in Early Visual Cortex. We used an inverted encoding model to estimate a representation of stimulus orientation in visual areas V1-hV4v/V3a. Within each plot, the vertical dashed bar at 0° indicates the actual stimulus orientation presented on each trial. Data from Category A and Category B 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. S2). Asterisks next to each region-of-interest label indicate a shift towards the center of Category B (quantified via a curve-fitting analysis, $p < 0.05$, false-discovery-rate-corrected across regions). The inset of each plot shows a logistic regression of each orientation channel's response onto trial-by-trial variability in category judgments. A positive coefficient indicates a positive relationship between an orientation channel's response and the correct category judgment, while a negative coefficient indicates a positive relationship between an orientation channel's response and the correct category judgment. For all plots, shaded regions are ±1 within-participant S.E.M.

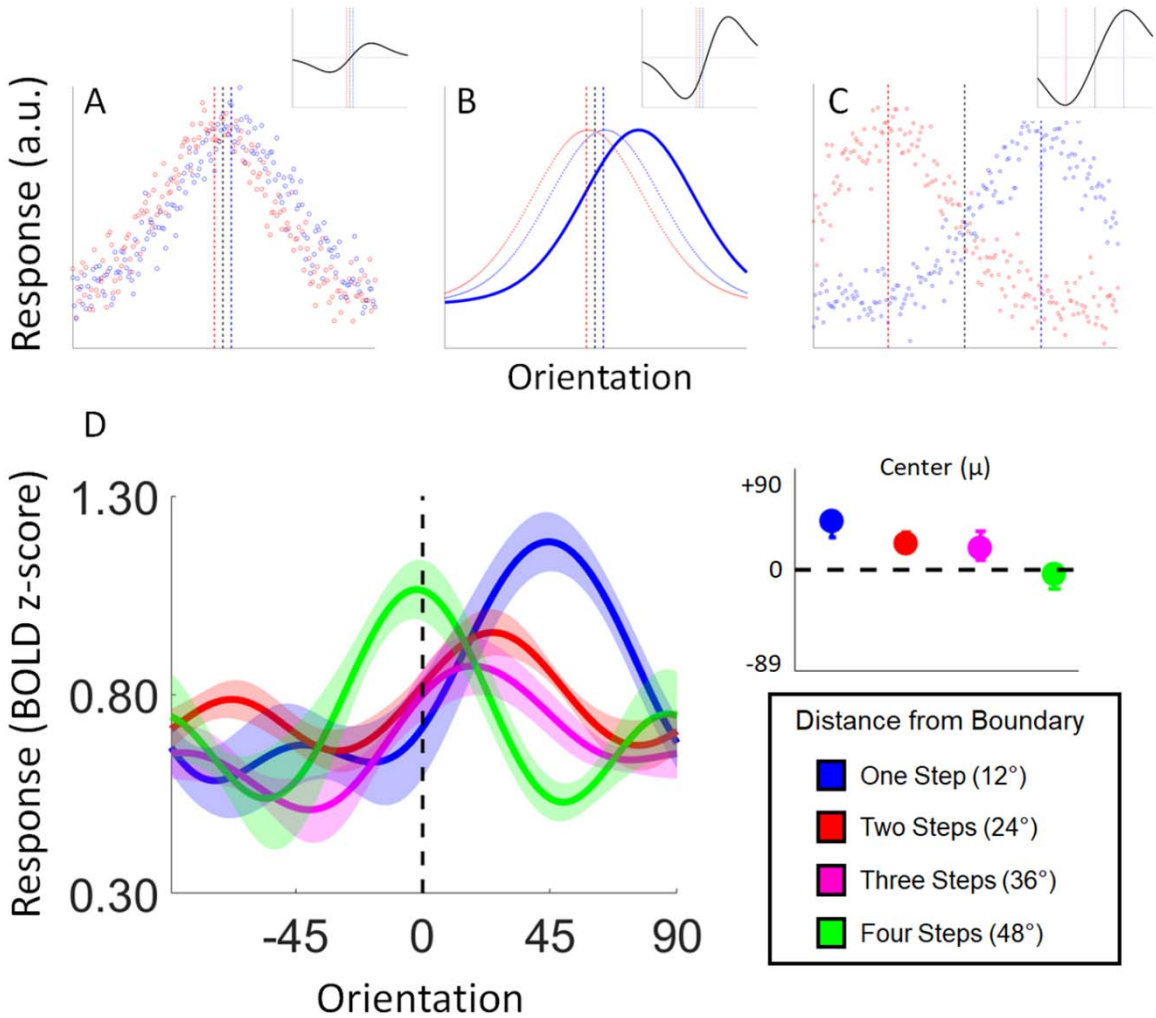

Fig. 4. Category Biases Scale Inversely with Distance from the Category Boundary. (A) Hypothetical example (synthetic data) depicting a scenario where an observer is shown an oriented exemplar from Category A (red dots and vertical dashed red line) or Category B (blue dots and vertical dashed blue line) that is adjacent to a category boundary (vertical black line). Each exemplar evokes a noisy response that is highly confusable with the other. Thus, the differential response evoked by each exemplar is relatively weak (inset). (B) The discriminability of the two signals could be enhanced by shifting or biasing the representation of each exemplar away from the category boundary. This would improve category discrimination performance by increasing the differential response evoked by the two exemplars (inset). Thus, categorical biases

should be particularly strong in these instances. (C) Categorical biases would be less helpful when participants are tasked with discriminating between two physically dissimilar exemplars, as the differential response across exemplars is already quite large. Thus, categorical biases should be small or absent in these instances. (D) To test this possibility, w sorted the reconstructions shown in Fig. 2 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°. Data have been pooled and averaged across visual areas V1-V3 as no differences were observed across these regions. Shaded regions are ±1 within-participant S.E.M. The inset 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.

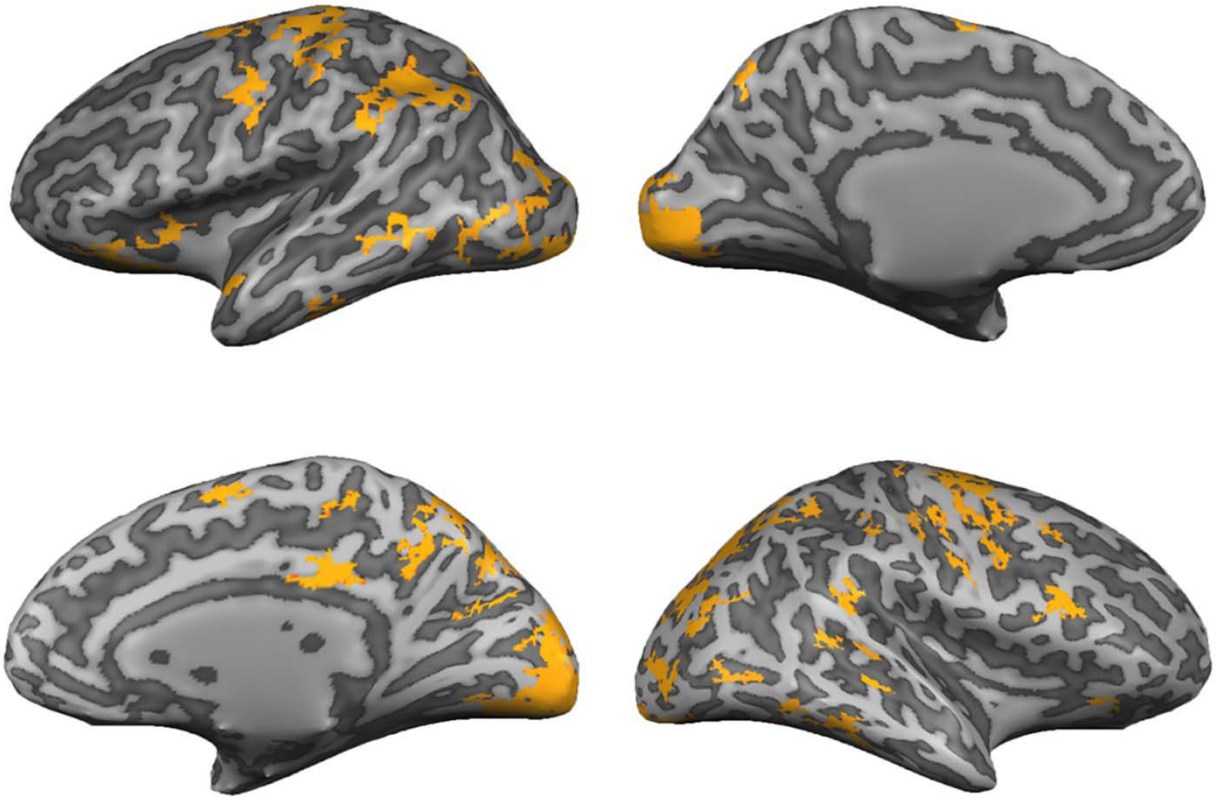

Fig. 5. Reconstructions of Stimulus Orientation in Cortical Areas Encoding Category Information. (A) We trained a linear support vector machine (LIB-SVM Implementation; *27*) to discriminate between activation patterns associated with Category A and Category B exemplars (independently of orientation; 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).

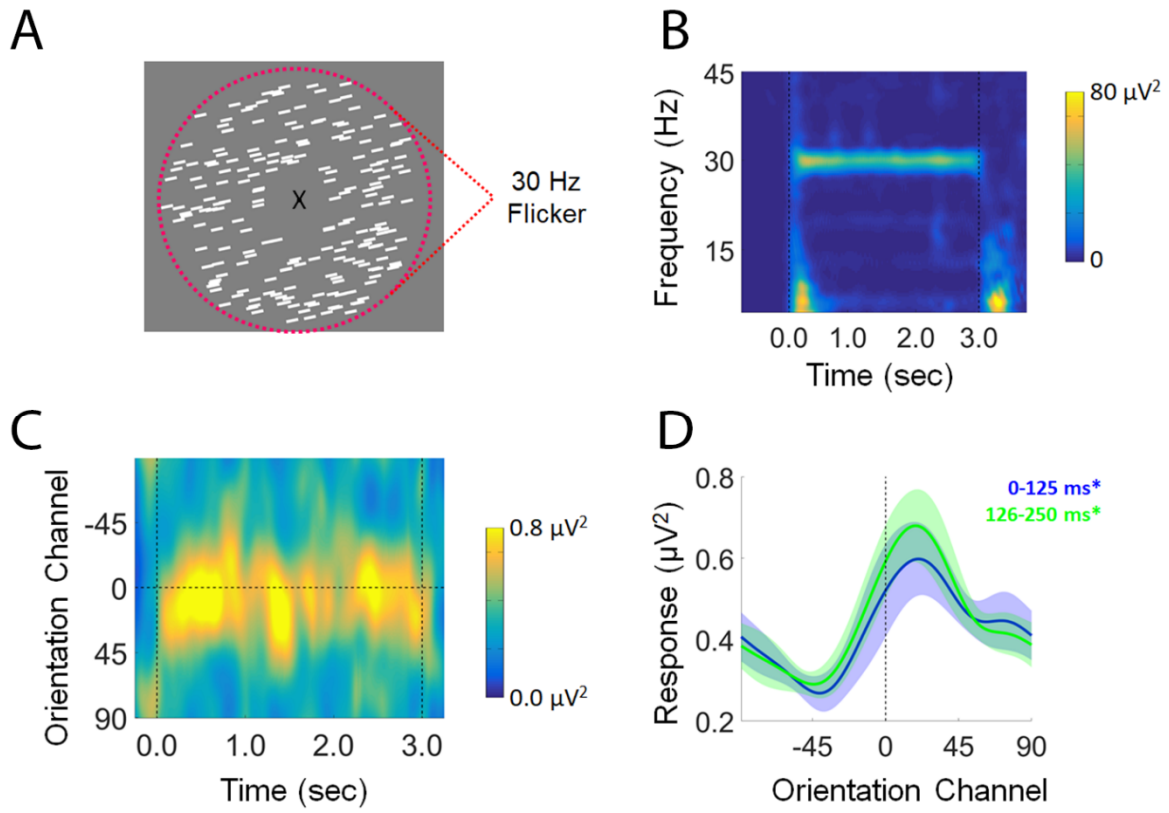

Fig. 6. Category Biases Emerge Shortly After Stimulus Onset. (A) On each trial participants viewed displays containing a large aperture of iso-oriented bars that flickered at 30 Hz while a rapid series of letters was presented at fixation. In separate blocks, participants detected the presence of a target in the letter stream while ignoring the oriented stimulus (orientation mapping task), or reported the category of the oriented stimulus (category discrimination task) while ignoring the letter stream. (B) The oriented stimulus drove a large frequency-specific response that was largest over posterior electrode sites (see Fig. S3). Dashed vertical lines at 0.0 and 3.0 sec mark stimulus onset and offset, respectively. (C) We used the power and phase of this frequency-specific response to generate a time-resolved representation of stimulus orientation. Dashed vertical lines at 0.0 and 3.0 sec mark stimulus onset and offset, respectively. (D) We examined the time course of category biases by averaging reconstructions from the first 500 ms

of each trial in 125 ms increments. Reliable category biases were present within 0-125 ms of stimulus onset and persisted throughout the 500 ms interval (asterisks $p < 0.05$, FDR-corrected across temporal epochs; shaded regions ±1 within-participant S.E.M.)

Supplementary Materials:

Materials and Methods

Figures S1-S5

References (*28-33*)

Supplementary Materials

Materials and 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 the neuroimaging experiment. 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) completed two 2-hour experimental scan sessions; an eighth participant (AH) 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 Psychophysics Toolbox software (*26*). 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 a rapid-serial-visual-presentation (RSVP) 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 RSVP task, participants attended a stream of letters presented at fixation (subtending 1˚ x 1˚ 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 (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 4˚ and 24˚ 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 (*9*), 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 A", while the seven orientations clockwise to this value were assigned membership in "Category B". 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 A or B (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. Whole-brain echo-planar images (EPIs) were acquired in 35 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). All images were acquired with a 32 channel Nova Medical head coil (Wilmington, MA). Functional scans 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). EPI images were unwarped using FSL (Oxford, UK), slice-time-corrected, motion-corrected, high-pass-filtered (to

remove first-, second- and third-order drift), transformed to Talairach space, and the timeseries of each voxel was normalized (z-scored) on a scan-by-scan basis.

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 a single retinotopic mapping run collected during each of two experimental scan sessions. 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 of 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.

Inverted Encoding Model. 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 (*11-12*). We modeled the responses of voxels measured during the RSVP 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; qualitatively similar results were obtained when we used 7 or 9 sinusoids raised to the 8th power). Let B_1 (m voxels x n_1 trials) be the response of each voxel 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 x k filters) be a weight matrix describing the mapping from "filter space" to "voxel space". The relationship between B_1, C_1, and W is given by a general linear model of the form:

$$B_1 = W C_1 \qquad \text{(Eq. 1)}$$

In the first phase of the analysis, in which we estimate an encoding model for each voxel, we use B_1 and C_1 to compute the weight matrix W via ordinary least-squares:

$$W = B_1 C_1^T (C_1 C_1^T)^{-1} \qquad \text{(Eq. 2)}$$

Next, we define a test data set B_2 (*m* voxels 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 inversion:

$$C_2 = (W^T W)^{-1} W^T B_2 \qquad \text{(Eq. 3)}$$

C_2 contains the predicted response of each orientation filter on each trial of the category discrimination task. We estimated 180-point reconstructions by repeating this analysis 12 times and shifting the centers of the basis functions of C_1 by 1° on each permutation. Finally, we circularly shifted each row (trial) of C_2 to be co-registered (aligned so that the presented orientation was labeled 0° by convention).

Trial-by-trial reconstructions were then sorted so that category biases would manifest as a clockwise or rightward shift (i.e., towards the center of category B). Consider a hypothetical example where the category boundary is 84°. Category A thus contains the orientations {0°, 12°, 24°, 36°, 48°, 60°, 72°} and Category B contains the orientations {96°, 108°, 120°, 132°, 144°, 156°, 168°}. In reality, there are two category boundaries: one located at 84°, and another located at 174°. To account for this, we circularly shifted reconstructions for trials where the stimulus' orientation was 0°, 12°, or 24° by -90°, then reversed (i.e., flipped from left-to-right)

the resulting reconstructions. This step aligns reconstructions for 0°, 12°, and 24° degree trials with reconstructions for 72°, 60°, and 48° trials. Next, we circularly shifted reconstructions for trials where the stimulus' orientation was 144°, 156°, and 168° by -24°, -48°, and -72°, respectively. We then reversed reconstructions from all Category A trials and averaged reconstructions across all trials. This procedure accounts for the fact that there are two category boundaries and ensures that any categorical bias should manifest in a rightward (clockwise) shift in representations of stimulus orientation.

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:

$$f(x) = \alpha\left(e^{k(\cos(\mu - x) - 1)}\right) + \beta \qquad \text{(Eq. 4)}$$

Here, 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 dispersion; a larger k corresponds to a "tighter" distribution) and the center of the function, respectively. Fitting was performed using a multidimensional nonlinear minimization algorithm (Nelder-Mead).

Putative changes in μ during the category discrimination task were evaluated with resampling tests. Specifically, we computed a "mean" reconstruction by randomly selecting (with replacement) and averaging reconstructions from the eight 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.

Searchlight Classification Analysis. We used a roving searchlight analysis (*27-29*) 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 Fig. 1) as members of Category A and the three orientations immediately clockwise of the boundary as members of Category B. 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 A and Category B orientations adjacent to the second (orthogonal) category boundary (see Fig. 1) and averaged the results across category boundaries.

We identified neighborhoods encoding stimulus category using a leave-one-participant-out cross validation approach (*31*). 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 Fig. S4). Finally, we extracted multivoxel activation patterns from each ROI and computed reconstructions of stimulus orientation 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 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.

Experiment 2 - EEG

Participants. Eleven neurologically intact volunteers (BA-BK; six females) from the UC San Diego community participated in the EEG experiment. Data from one Experiment (BD) were discarded due to an experimenter error (specifically, the wrong EEG montage was selected prior to recording and data from 64 of 128 channels were lost). The data reported here reflect the remaining 10 participants.

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.

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 RSVP and category tasks like in Experiment 1. Importantly, the same stimulus displays were used for 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 2.2˚ x 0.8˚) 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 RSVP 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 "A" or category "B". 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 "A" and the seven orientations clockwise to this value were assigned to category "B". 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. 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 (N = 9) or the right mastoid (N = 1; data from the left mastoid were highly unstable in this subject). 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. 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. = 1.99 ± 0.30).

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 S5).

To isolate stimulus-specific responses, the epoched EEG timeseries at each electrode was 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. Finally, we trimmed the first and last 750 ms of each trial as these intervals were only included as padding against edge artifacts that can manifest during filtering. Thus, the remaining data focus on a temporal interval from -250 to 3250 msec relative to stimulus onset.

To visualize and quantify orientation-selective signals from frequency-specific responses, We constructed a complex-valued training data set $B_1(t)$ (*m* electrodes x *n* trials) containing the complex-valued coefficients measured on trials 1:*n* and electrodes 1:*m* at time *t* during the RSVP task. 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 (see Equation 2 above). Finally, we estimated a complex-valued matrix of channel responses $C_2(t)$ (*m* channels x *n* trials) given W(t) and complex-valued test data set $B_2(t)$ (*m* electrodes x *n* trials) containing the complex Fourier coefficients measured on trials 1:*n* and electrodes 1:*m* at time *t* using data from the category discrimination task (Equation 3). 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. Categorical biases were quantified using the same curve fitting analysis described above.

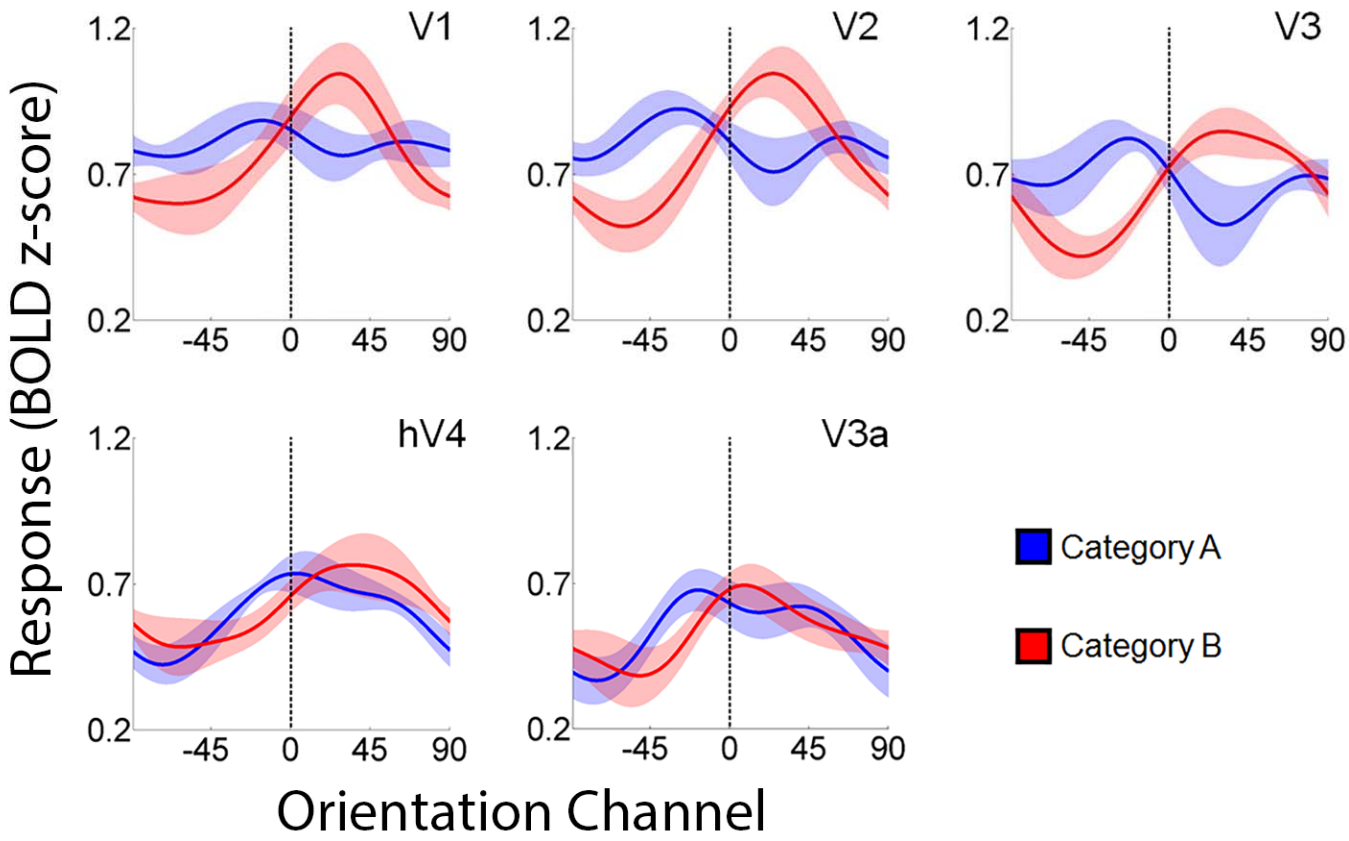

Fig. S1. Stimulus Reconstructions during Category A and Category B trials. Shaded regions are ±1 within-participant S.E.M.

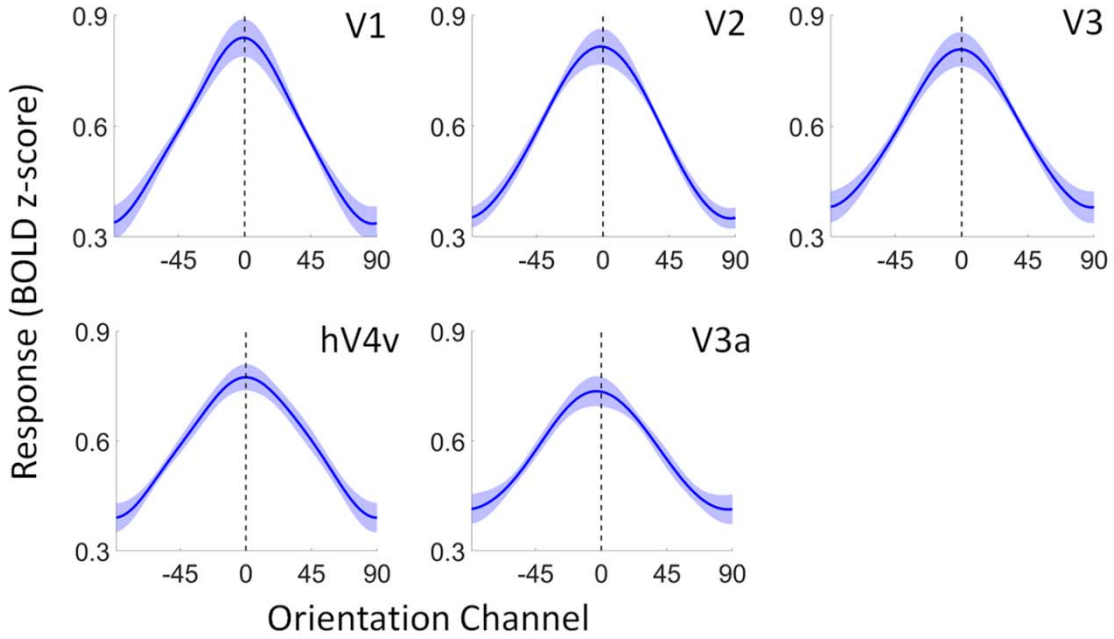

Fig. S2. Reconstructions of stimulus orientation during the orientation mapping task. For each participant, we defined a training data set containing responses from all but one block of the orientation mapping task. Data from the remaining block were used to define a test set. Data from the training set were used to estimate orientation weights for voxels in visual areas V1-V3 (Fig. 2); these weights were used to reconstruct a representation of stimulus orientation for each trial of the test set. This procedure was repeated until each orientation mapping block had served as the test set and the results were averaged across permutations. No categorical biases were observed in any of the visual areas we examined. Shaded regions are ±1 within-participant S.E.M.

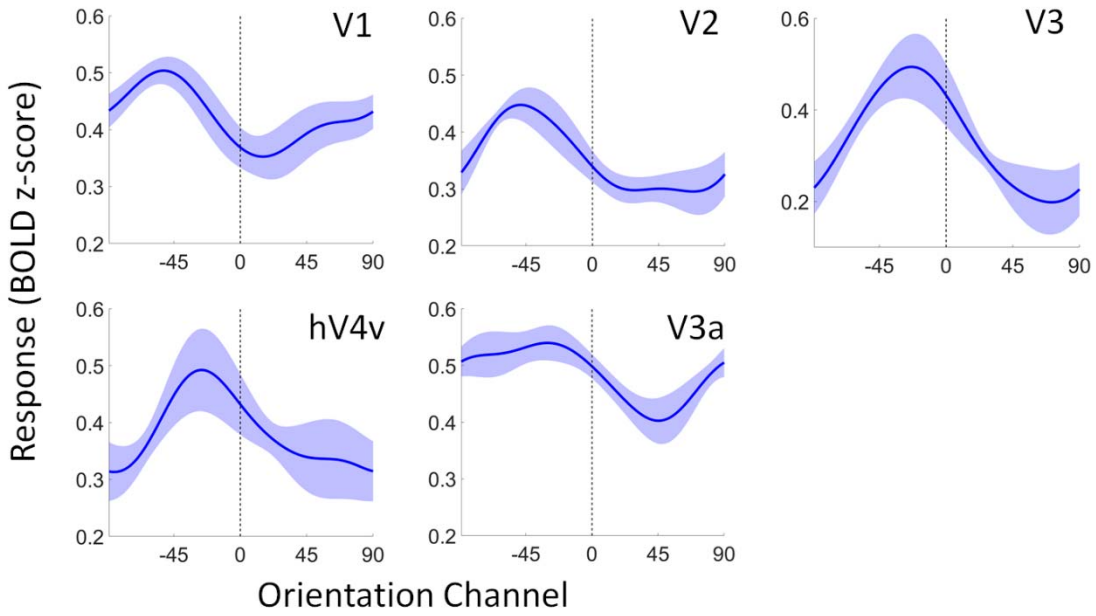

Fig. S3. Reversed Inverted Encoding Analysis. In a supplementary analysis, we used data from the category discrimination task to estimate orientation weights for each voxel in visual areas V1-hV4/V3a. These weights were used to compute a reconstruction of stimulus orientation on each trial of the orientation mapping task. We reasoned that if categorical biases in reconstructed representations of stimulus orientation during the category discrimination task (Fig. 3) are due to changes in the spectral response properties of feature-selective visual populations, then a model trained using data from the category discrimination task and applied to activation patterns during the orientation mapping task should produce biased reconstructions (even though orientation was wholly irrelevant to the participants' task during orientation mapping). Significant biases were manifest in V1-V3 (FDR-corrected p-values ≤ 0.05). Shaded regions are ±1 within-participant S.E.M.

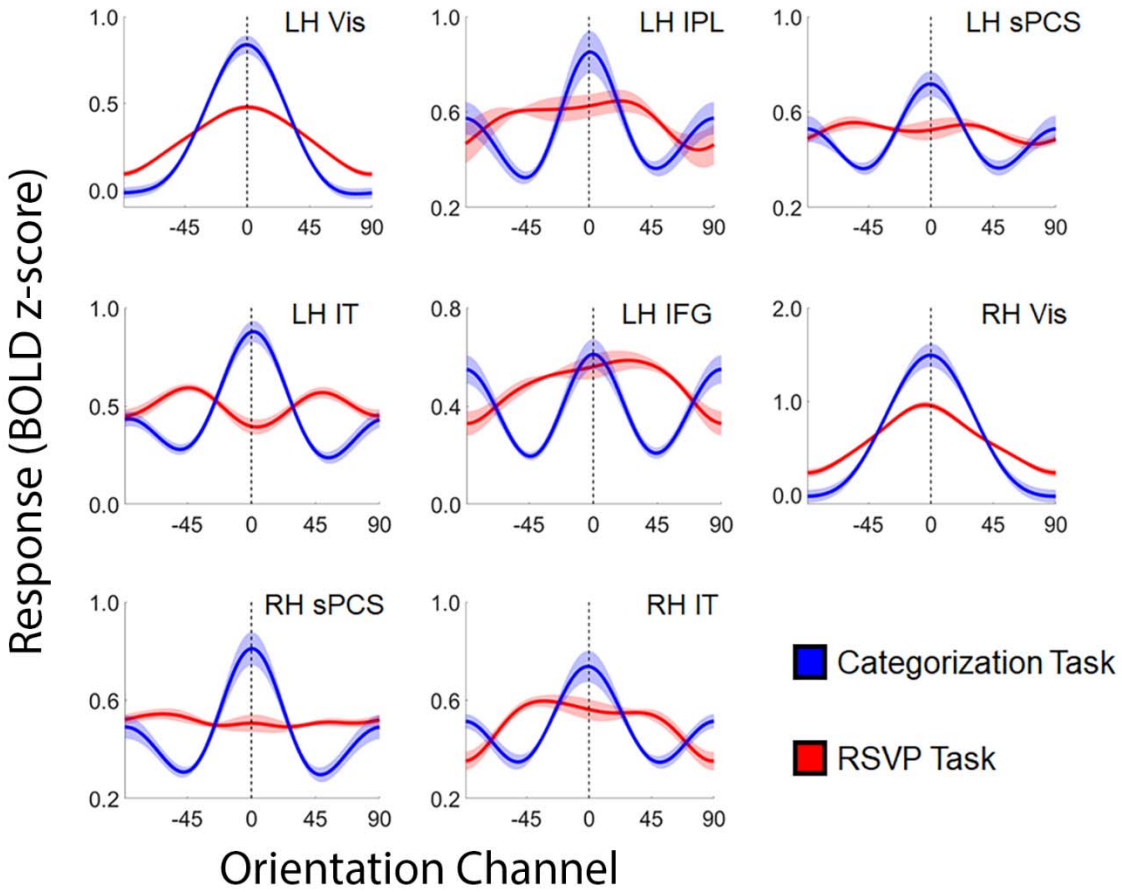

Fig. S4. Stimulus Reconstructions in Visual, Parietal, and Frontal cortical areas during the Orientation Mapping and Categorization Tasks. Shaded regions are ±1 within-participant S.E.M.

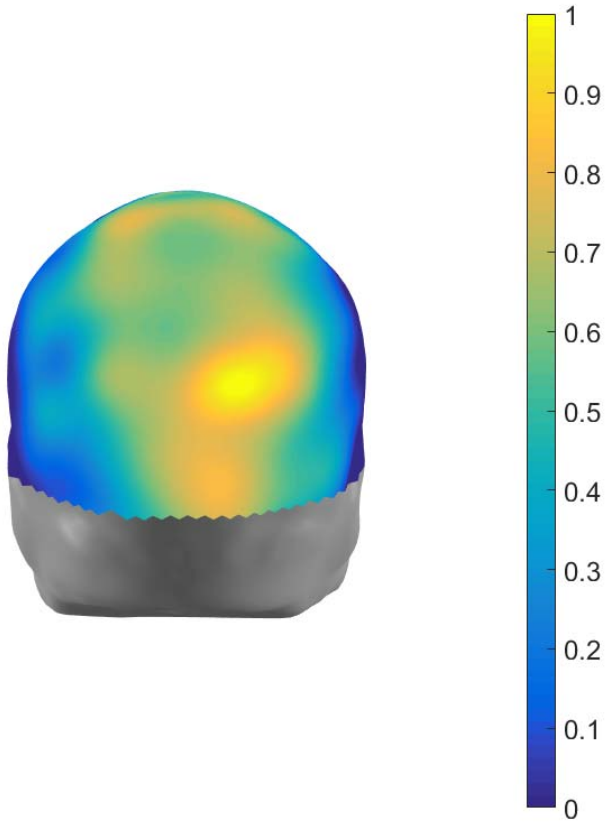

Fig. S5. Topographical plot showing the locations of electrodes with the highest power at the stimulus' flicker frequency in the EEG experiment. Posterior view of scalp. Scale bar refers to the proportion of all 10 subjects for which each electrode was selected. For example, a score of 1.0 would mean that the same electrode was included in the IEM analysis for all 10 subjects, while a score of 0.8 would mean that the same electrode was included in the IEM analysis for 8 of 10 subjects.

References

28. M. Kleiner, D. Brainard, D. Pelli, *Perception* **36** 14 (2007)

29. J.T. Serences, G.M. Boynton, *Neuron* **55**, 301-312 (2007)

30. E.F. Ester, T.C. Sprague, J.T. Serences, *Neuron* **87**, 893-905 (2015)

31. E.F. Ester, D.W. Sutterer, J.T. Serences, E. Awh, *J Neurosci* **35** 8188-8199 (2016)

32. C-C Chang, C-J Lin, *ACM Trans Intell Syst Technol* **2**(27), 1:27 (2011)

33. M. Esterman, B.J. Tamber-Rosenau, Y-C Chiu, S. Yantis, *NeuroImage* **50**, 572-576 (2010)