bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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: 234

Article Word Count (inc. References & Fig Captions): 5011

Reference Count: 29

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/

Significance Statement: Category learning alters perceptual sensitivity by enhancing the discriminability of similar exemplars from different categories. These distortions could in part reflect changes in how sensory neural populations selective for category-defining features encode information. To test this possibility, we used multivariate analytical techniques to reconstruct and quantify representations of oriented stimuli after observers had learned to classify them into two discrete groups. Representations of orientation encoded by several early visual areas were systematically biased according to their category membership, with larger biases for orientations adjacent to the boundary that defined each category. This result suggests that categorizing a stimulus alters how that stimulus is represented at the earliest stages of the visual processing hierarchy.

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Categorization refers to the process of mapping sensory inputs onto discrete concepts. Humans and other animals can readily learn arbitrary categories defined by low-level visual features such as hue, and behavioral studies indicate that such learning distorts perceptual sensitivity for the category-defining feature 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 changes in how sensory neural populations selective for category-defining features encode information. Here, we tested this possibility by using noninvasive measurements of human brain activity (fMRI and EEG) to visualize and quantify population-level representations of oriented stimuli encoded by early visual cortical areas after participants had learned to classify these stimuli into discrete groups. Representations of orientation encoded by visual 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 within a few hundred milliseconds of stimulus onset. Collectively, these results suggest that categorizing a stimulus alters how that stimulus is represented at the earliest stages of the visual processing hierarchy, and may provide a physiological basis for distortions in perceptual sensitivity following category learning.

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 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 represented by sensory neural populations (*6*-*7*). Here, we tested this possibility by using noninvasive measurements of human brain activity (fMRI and EEG) to visualize and quantify population-level representations of oriented stimuli in early visual cortical areas after participants had learned to classify these stimuli into discrete groups. 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 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

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

stimulus onset, suggesting that the intent to categorize a stimulus modulates how neural populations in early visual areas represent sensory information.

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Experiment 1 - fMRI

Inspired by earlier work in non-human primates (*8-9*), 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 (*10-11*) 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.

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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). Similar, though less robust categorical biases were also evident in hV4v and V3a (mean shifts of 11.54° and 8.37°, 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. 3, insets). That is, trial-by-trial variability in participants' reports were more strongly determined by orientation channels near the center of each category rather than those near the physical orientation of the stimulus.

Before continuing, we considered the trivial possibility that the categorical biases shown in Fig 3 reflect intrinsic biases in stimulus selectivity in early visual areas (e.g., due to oblique

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

effects). This possibility is unlikely for two reasons. First, 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). Thus, the rule defining Categories A and B varied across participants independently of stimulus orientation. Second, 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

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (*6-7*, *12-13*), inferotemporal cortex (*14*), posterior parietal cortex (*8-9*), and lateral prefrontal cortex (*15*). 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., *16*). 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 categorical 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 tested this idea by examining categorical biases

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

over the first few hundred milliseconds of each category discrimination trial using EEG. 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 throughout each trial. To discriminate between these alternatives, 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, *17*). 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 (*18*). Thus, by using a relatively high flicker rate of 30Hz, most of the SSVEP is

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

likely generated locally in posterior regions of occipitoparietal 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 (*19*). 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 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 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 250 ms of each trial, including a temporal interval spanning 0-125 ms, suggesting that the intent to categorize a stimulus modulates how neural populations in early visual areas respond to incoming sensory signals.

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (*20*) and minimizing the discriminability of equally similar stimuli from within the same category (*6-7*). 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.

The categorical biases observed in V1-V3 (Fig. 3) could result from stimulus-invariant changes in the spectral preferences and/or response gain of sensory neural populations responsible for encoding to-be-categorized orientations (e.g., *21-24*). However, both alternatives are difficult to reconcile with our observation that the magnitudes of category selective biases are in part determined by the similarity between an exemplar and the boundary delineating the two categories (Fig. 4C). Alternatively, recent studies have identified time- and stimulus-dependent categorical biases in macaque inferotemporal cortex that are well-described by a recurrent dynamical model with discrete attractors (*25-26*). However, the recurrent nature of this model predicts that categorical biases should emerge gradually over time (on the order of several

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

hundred milliseconds), which 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 (*27-29*). In one example (*27*), 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

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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., *9*, *16*, *20*), 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 (*20*). More broadly, our findings add to a growing set of observations suggesting that information processing in early visual cortical areas is incredibly flexible and can be adapted to maximize performance on an observer's task.

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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. 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)

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

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

9. 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)

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

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

12. 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)

13. 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)

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

15. 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)

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

17. 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)

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

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

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

20. 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)

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. 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).

23. Treue, S. & Martinez-Trujillo, J.C. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* **399** 575-579 (1999)

24. Navalpakkam, V. & Itti, L. Search goal tunes visual features optimally. *Neuron* **53** 605-617 (2007)

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

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

27. 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)

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

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

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

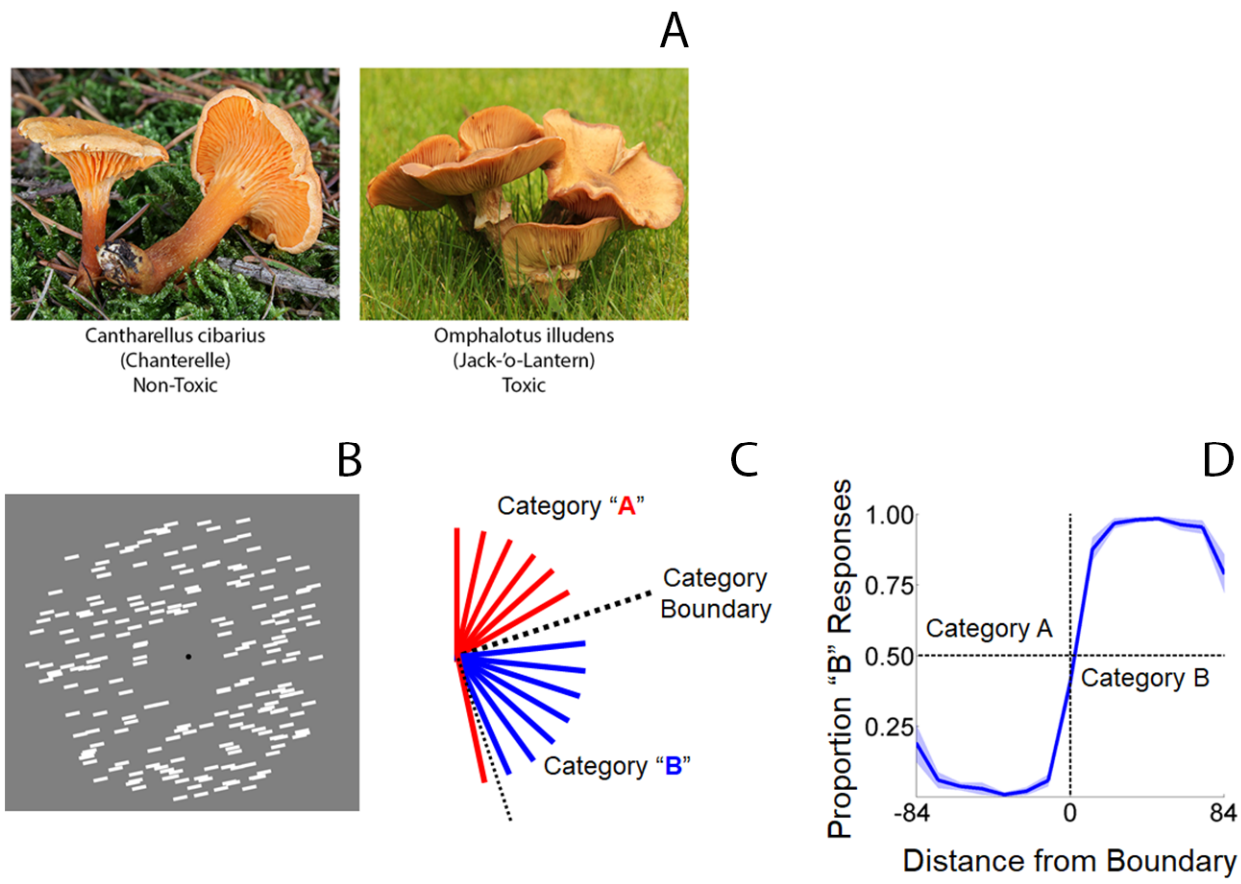

Fig. 1. Behavioral Task. (A) An example of physically similar yet categorically distinct stimuli. The edible and poisonous mushrooms are best distinguished by variations in low-level visual properties such as hue, and texture. (B) In the category discrimination task, participants viewed displays containing 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.).

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

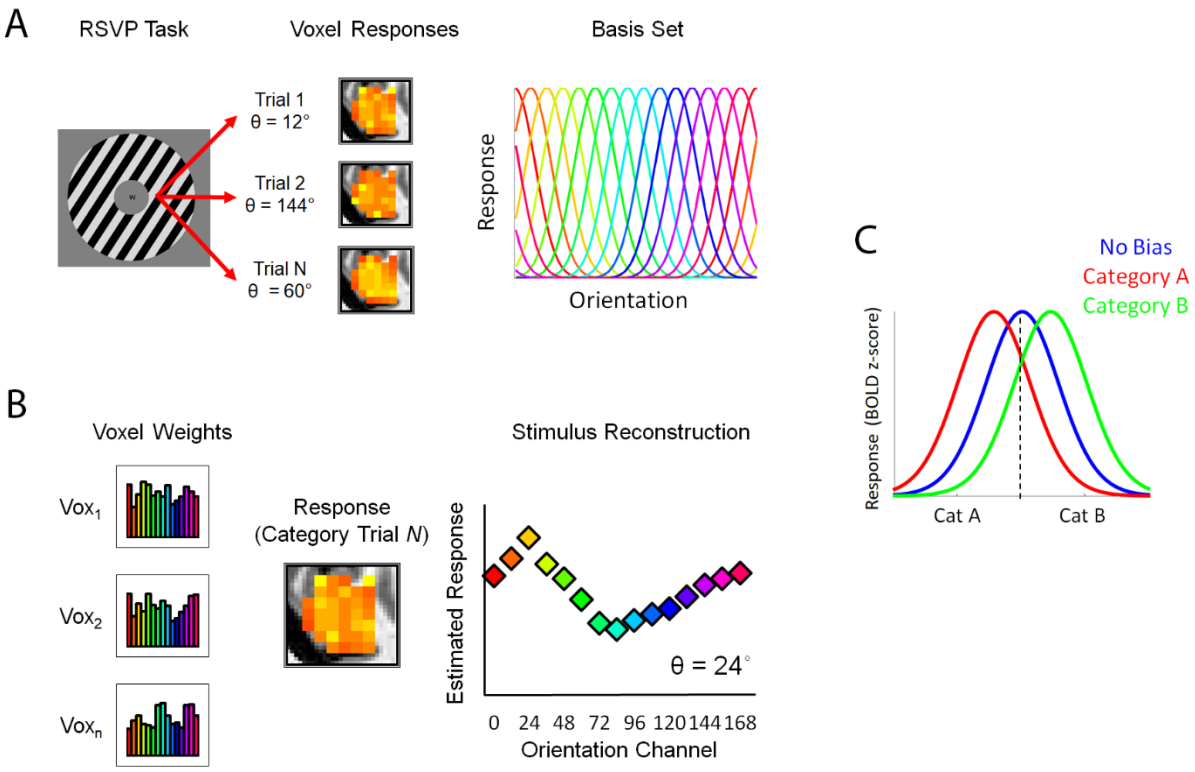

Fig. 2. Schematic Overview of the Inverted Encoding Analysis. (A) In the first phase of the analysis, we estimated an orientation selectivity profile for each voxel in each retinotopically organized visual and parietal region we examined. Participants performed a "rapid serial visual presentation" (RSVP) 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

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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°.

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

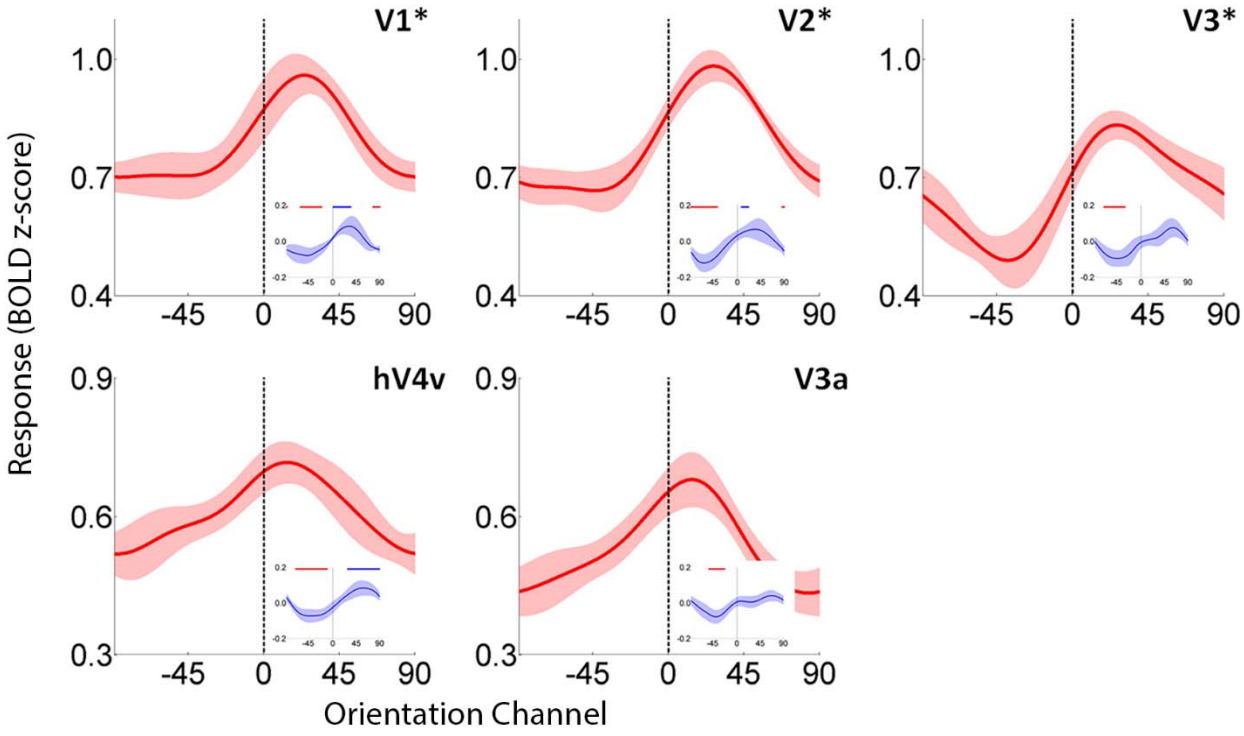

Fig. 3. Reconstructed representations of Orientation in Early Visual Cortex. We used an inverted encoding model to estimate a representation of stimulus orientation in visual areas V1-hV4v/V3a. Within each plot, the vertical dashed bar at 0° indicates the actual stimulus orientation presented on each 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.

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

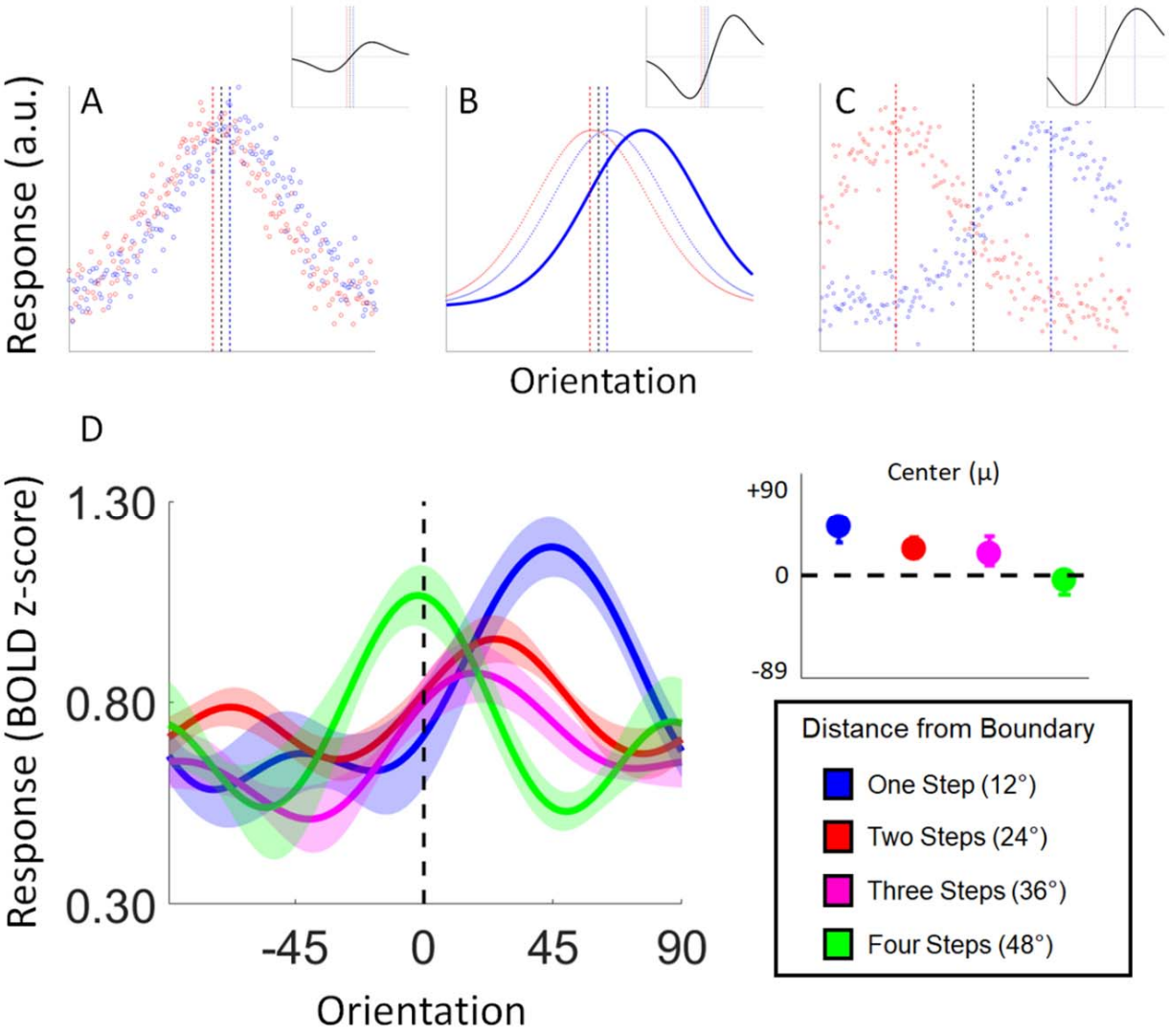

Fig. 4. Category Biases Scale Inversely with Distance from the Category Boundary. (A) Hypothetical example (synthetic data) depicting a scenario where an observer is shown an oriented exemplar from Category A (red dots and vertical dashed red line) or Category B (blue dots and vertical dashed blue line) 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

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

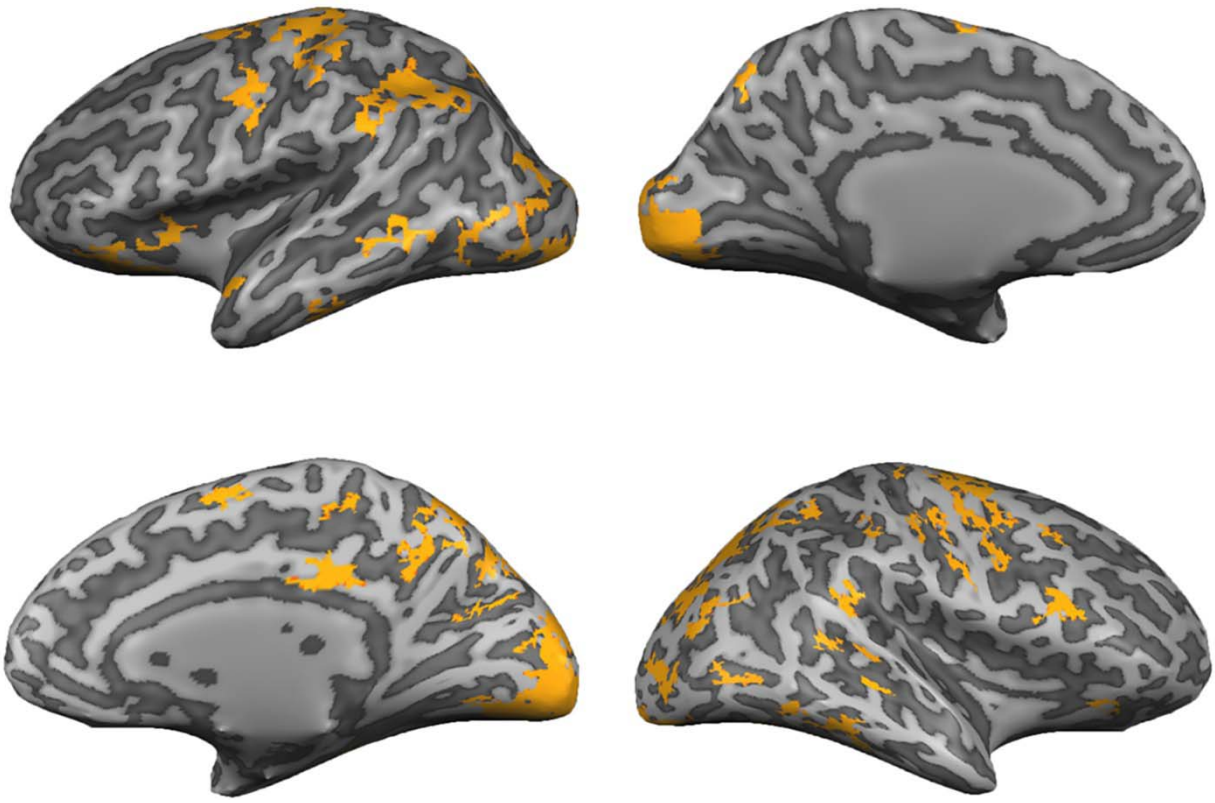

Fig. 5. Reconstructions of Stimulus Orientation in Cortical Areas Encoding Category Information. (A) We trained a linear support vector machine (LIB-SVM Implementation; *27*) to discriminate between activation patterns associated with Category A and Category B exemplars (independently of orientation;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

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

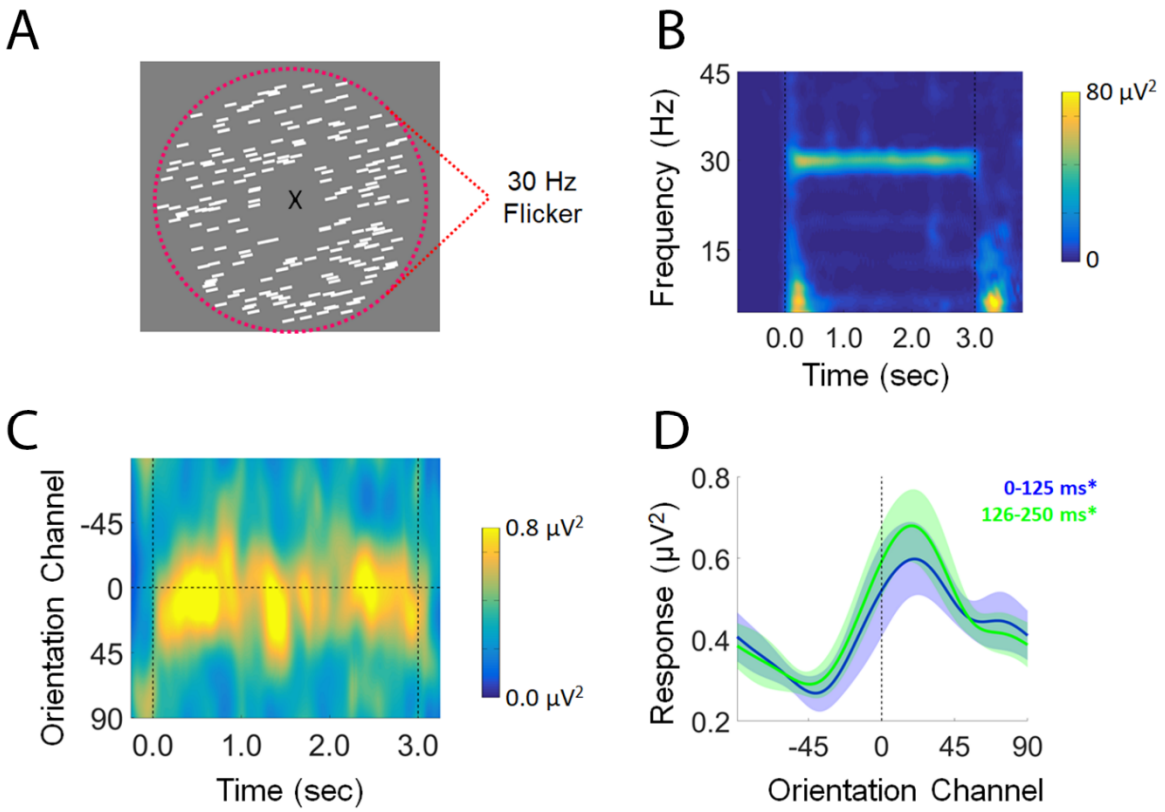

Fig. 6. Category Biases Emerge Shortly After Stimulus Onset. (A) On each trial participants viewed displays containing a large aperture of iso-oriented bars that flickered at 30 Hz while a rapid series of letters was presented at fixation. In separate blocks, participants detected the presence of a target 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 250 ms

bioRxiv preprint doi: https://doi.org/10.1101/170845; this version posted August 24, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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