Journal Club

Editor's Note: These short reviews of recent *JNeurosci* articles, written exclusively by students or postdoctoral fellows, summarize the important findings of the paper and provide additional insight and commentary. If the authors of the highlighted article have written a response to the Journal Club, the response can be found by viewing the Journal Club at www.jneurosci.org. For more information on the format, review process, and purpose of Journal Club articles, please see http://jneurosci.org/content/jneurosci-journal-club.

Complementary Brain Signals for Categorical Decisions

Paul S. Muhle-Karbe[1,2] and Timo Flesch[1]

[1]Department of Experimental Psychology, University of Oxford, Oxford OX2 6GG, United Kingdom, and [2]Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford OX3 9DU, United Kingdom

Review of Ester et al.

Apples come in various shapes, colors, and sizes, but humans can nonetheless easily distinguish them from other fruit (e.g., peaches) or other round objects (e.g., tennis balls). This capacity to assign complex and variable stimuli into discrete and meaningful categories is an essential cognitive function that dramatically reduces cognitive load, and enables agents to generalize their prior knowledge to novel exemplars. A long research tradition in Cognitive Science has studied how humans acquire such knowledge and learn to categorize unknown objects (Ashby and Maddox, 2011). One established finding from this literature is that category learning yields biases in perceptual sensitivity, whereby stimuli belonging to different categories can be better distinguished than stimuli belonging to the same category, even when stimuli vary along continuous features and the objective physical similarity is identical for pairs within and across categories (Goldstone, 1994; Fig. 1*A*).

Neuronal recordings in animals have revealed signatures of these biases in higher association areas, especially the posterior parietal and dorsolateral prefrontal cortices (Freedman and Assad, 2016). Many of these studies have used the delayed match-to-category (DMC) task, in which animals learn to categorize two successively presented stimuli, the sample and the probe, based on an arbitrary category boundary that was imposed by the experimenter. For example, two patches of coherently moving dots might be shown sequentially, and the animal needs to indicate at the end of the trial whether they belonged to the same or to different categories based on their direction of movement. Studies have typically focused on neural patterns evoked by the sample, as this permits dissociation of category-selective signals from motor signals. The sample must be assigned to the correct category to solve the task, but at this point no behavioral response can be planned, because observers cannot anticipate whether the probe will be a match or a mismatch. Using this procedure, numerous studies have shown that neurons in association areas exhibit shifts in their motion tuning that resemble perceptual biases of human observers: sample stimuli from the same categories evoke more similar responses than sample stimuli from distinct categories, even when the objective similarity of their movement direction is identical (Freedman and Assad, 2016). Changes in the position of the category boundary bring about corresponding changes in neural tuning, corroborating that tuning biases reflect the animal's acquired knowledge about the categorical structure of the task (Freedman and Assad, 2006; Cromer et al., 2010). Moreover, category signals in association areas contrast with more stable and stimulus-grounded responses in sensory areas (Freedman and Assad, 2006), suggesting that sensory neurons act as veridical feature detectors, whereas association neurons extract categorical information.

A recently published study by Ester et al. (2020) has challenged this view. In a multimethod neuroimaging study with human observers, category learning produced biases in the earliest stages of cortical visual processing. The study used a novel category discrimination (CD) task in which observers saw displays of oriented bars, and learned to categorize them based on their angle with respect to a boundary that was imposed by the experimenter and had to be inferred via trial-wise feedback (Ester et al., 2020, their Fig. 1*A*). To test whether learning of this boundary produced biases in perceptual sensitivity, the authors trained a pattern classifier to discriminate orientation signals using neural data from an independent orientation localizer task, in which observers searched for the letters "X" or "Y" within a stream of letters while oriented gratings were shown in the background. Orientation was irrelevant in the localizer task, so the classifier should extract neural patterns that reflect this stimulus feature truthfully. Critically, the same classifier was then tested on its ability to discriminate stimulus orientation using neural data from the CD task. If category learning does not affect the

Received Mar. 31, 2020; revised May 31, 2020; accepted June 7, 2020.

P.S.M.-K. is supported by a Sir Henry Wellcome Postdoctoral Fellowship funded by the Wellcome Trust (Award 210849/Z/18/Z) and a Junior Research Fellowship funded by Linacre College and the EPA Cephalosphorin Fund. T.F. was supported by a Medical Science Division graduate studentship funded by the Department of Experimental Psychology and the Medical Research Council. The Wellcome Center for Integrative Neuroimaging is supported by core funding from the Wellcome Trust (Award 203139/Z/16/Z).

Correspondence should be addressed to Paul S. Muhle-Karbe at paul.muhle-karbe@psy.ox.ac.uk.

https://doi.org/10.1523/JNEUROSCI.0785-20.2020

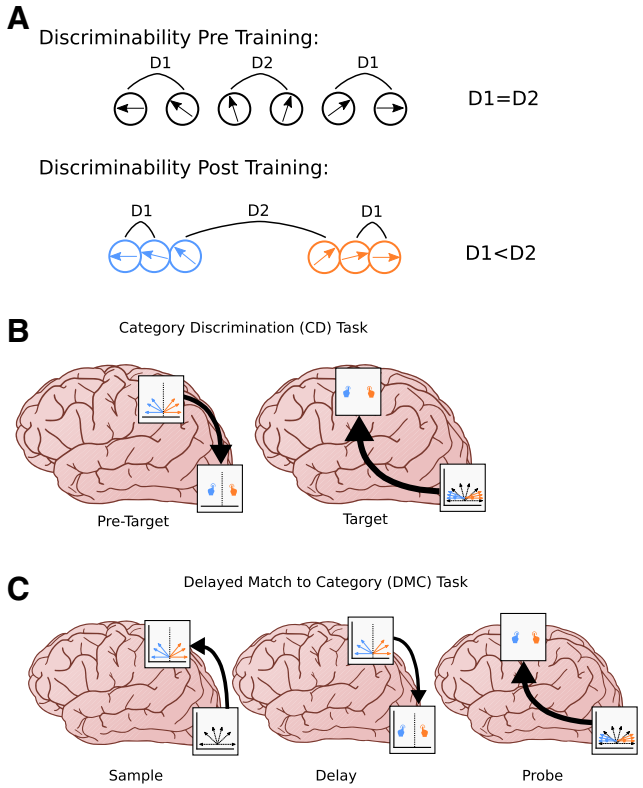

Figure 1. ***A***, Illustration of perceptual biases induced by category learning. Observers are trained to assign stimuli that vary along continuous features such as motion direction (represented via the direction of arrows) into two discrete categories (represented via blue and orange colors). Before training, stimulus discriminability truthfully reflects the evenly spaced physical similarity of stimuli. After training, stimulus representations are systematically biased away from the learned category boundary, such that pairs that were similarly discriminable before training are now more easily distinguishable when they belong to different categories (D2) than when they belong to the same category (D1). ***B***, Proposed information flow in the CD task. In this paradigm, the mapping between stimulus categories and responses is fixed, so observers can prepare a decision circuit in anticipation of a target stimulus (e.g., response 1 for category A, shown in orange; and response 2 for category B, shown in blue). Previous work suggests that this involves top-down signals from association areas toward sensory areas to enforce task-dependent stimulus–response mappings via connections between stimulus-selective and decision-relevant neurons (represented button presses on each side of the category boundary). This preactivated decision circuit enables rapid filtering of stimulus-evoked patterns, yielding a binary-like choice signal that can immediately guide behavior in downstream regions (Hayden and Gallant, 2013; Myers et al., 2015). ***C***, Proposed information flow in the DMC task. In this paradigm, responses are indicated at the end of the trial to signal whether sample and probe stimuli belonged to the same or to different categories. Consequently, both stimuli need to be processed to enable selection of a task-appropriate response. Under these conditions, sample processing likely follows a feedforward information flow, whereby sensory neurons detect stimulus features, and association neurons extract category membership. During the delay period, categorical information can be used to construct a temporary decision circuit, because the response contingencies for the probe are now explicit (e.g., response 1 for a match and response 2 for a mismatch). Probe processing may then resemble target processing in the CD task, though links between stimulus-selective and decision-relevant units are likely weaker, because the stimulus-response mappings change on a trial-by-trial basis in the DMC task, whereas they are fixed throughout the whole experiment in the CD task.

representation of orientation, the classifier should cross-generalize between tasks and estimate the orientation of stimuli from the CD task accurately. Alternatively, if category learning produces greater sensitivity to orientation changes around the learned category boundary, the classifier should exhibit systematic biases, and estimate stimuli from the CD task to be repelled from the boundary toward the center of the category to which they belong. Intriguingly, the latter is precisely what was found.

Initially, the task was combined with fMRI to identify brain areas exhibiting categorical biases. In contrast to previous work, biases were observed not only in frontoparietal association areas, but also in the primary visual cortex. To verify that these signals reflect a modulation of early perceptual processing, the task was subsequently combined with EEG. This demonstrated that tuning shifts emerged rapidly after stimulus onset. In both experiments, categorical biases exhibited several important properties: they were very large in magnitude (the classifier trained on stimuli from the localizer misjudged the orientation of stimuli from the CD task by 20–45° of visual angle), correlated with participants' behavioral judgments, and were most pronounced for stimuli that were close to the category boundary, and therefore required the greatest amount of perceptual disambiguation. Together, these results provide strong evidence for categorical tuning in sensory brain areas previously thought to be insensitive to categorical structure.

An important difference between the CD task by Ester et al. (2020) and the DMC task used in animal research is that the former uses a fixed mapping between categories and responses, whereas the latter explicitly decouples them by asking observers to match stimuli based on their category membership instead of reporting membership directly. Consequently, in the CD task, categories are confounded with the responses used to report them. To test whether the observed tuning shifts were caused by this confound, the authors conducted another experiment, combining a DMC task with EEG recordings (Ester et al., 2020, their Fig. 17*A*). Robust tuning shifts remained, leading the authors to conclude that category-selective signals in their studies did not reflect motor confounds. Notably, however, biases in the DMC task behaved quite differently from biases in the CD task: they were considerably smaller in magnitude, occurred much later in time, and unfolded gradually from initial signals coding for veridical stimulus features.

Collectively, the findings by Ester et al. (2020) indicate that different neural mechanisms might be responsible for tuning shifts in the two paradigms. Category-selective signals in the sample period of the DMC task reflect abstract decision variables that are detached from immediate action. The sample is only covertly categorized and used to set up a decision circuit for the forthcoming classification of the probe (see below). This form of categorization appears to rely on a feedforward information flow from sensory areas toward association areas. In contrast, the sudden category-selective signals in the CD task reflect "action-oriented perception" (Gibson, 1979), because the category membership of the stimulus immediately guides behavior. Here, the information flow likely differs, because observers can already establish a decision circuit before stimulus onset to link representations of anticipated inputs with representations of corresponding outputs via feedback signals from association areas toward sensory areas (Baldauf and Desimone, 2014; Muhle-Karbe et al., 2017). Enforcing task-dependent links between stimulus-selective and decision-relevant populations

enables stimulus-evoked patterns to be rapidly matched and transformed into discrete choices (Myers et al., 2015). Interestingly, in tasks with predictable stimulus–response mappings, a large proportion of visual cortical neurons indeed codes for behavioral choice in a binary-like manner that carries sufficient information to drive behavior in downstream regions (Mirabella et al., 2007; Sasaki and Uka, 2009; Hayden and Gallant, 2013; Xin et al., 2019). Similar signals have been measured in humans with fMRI (Woolgar et al., 2011).

Category biases in the CD task may thus reflect the coexistence of two neural representations in the visual cortex: a veridical stimulus representation, and a binary choice representation. To the extent that these representations are aligned (i.e., encoded along statistically nonindependent axes of the population firing space), this superimposition would reproduce the results by Ester et al. (Fig. 1). A choice signal in visual cortex could also help to explain the remarkable magnitude of tuning shifts in the CD task that appears unlikely to reflect changes in stimulus processing alone. Such extreme distortions would seem maladaptive for perception, and difficult to reconcile with previous studies on perceptual learning, attention, or predictive coding, all documenting far more rigid sensory responses, even after extensive training regimes (Law and Gold, 2008; Scolari et al., 2012).

Our proposal could be tested in multiple ways. For instance, in the DMC task, bias signals could be compared between sample and probe. Information about response contingencies is absent for the sample but is present for the probe. Accordingly, tuning shifts should be stronger and arise at earlier processing stages with the probe (though this would require splitting the data based on the category membership of the sample). Conversely, categorical perception could be decoupled from action in the CD task via delayed response-mapping cues (Bennur and Gold, 2011), which should yield more subtle and gradually unfolding bias signals that were similar to those observed during the sample period of the DMC task.

On a larger scale, the distinction between abstract and action-oriented category signals prompts an important question; namely, to what extent do they reflect the computations of the brain during natural categorization? Abstract signals confer the great advantage of neural efficiency, because a single neuronal population can potentially be reused for different types of discriminations. In support of this idea, Fitzgerald et al. (2011) found that individual neurons in the lateral intraparietal area exhibited strong category selectivity for different object classes, suggestive of general purpose decision computations. Nonetheless, these signals may convey only part of the whole process, possibly indexing readout and/or maintenance of categorical judgments rather than the integration processes leading toward them. Choice-related signals may fill this gap and provide a window into neural dynamics that transform sensory evidence into categorical reference frames (Xin et al., 2019). Furthermore, the fact that different object classes afford unique actions likely plays a major role in the initial acquisition of category knowledge (Summerfield et al., 2020). Future research should therefore consider action not only as a confound that may contaminate category-selective signals, but also as a source of input that can sculpt continuous, high-dimensional stimulus spaces into discrete groups. Abstract and action-oriented signals may thus reflect complementary aspects of categorization, and a major contribution of the experiments by Ester et al. (2020) is that they have brought these questions into the realm of human cognitive neuroscience.

References

Ashby FG, Maddox WT (2011) Human category learning 2.0. Ann N Y Acad Sci 1224:147–161.

Baldauf D, Desimone R (2014) Neural mechanisms of object-based attention. Science 344:424–427.

Bennur S, Gold JI (2011) Distinct representations of a perceptual decision and the associated oculomotor plan in the monkey lateral intraparietal area. J Neurosci 31:913–921.

Cromer JA, Roy JE, Miller EK (2010) Representation of multiple, independent categories in the primate prefrontal cortex. Neuron 66:796–807.

Ester EF, Sprague TC, Serences JT (2020) Categorical biases in human occipitoparietal cortex. J Neurosci 40:917–931.

Fitzgerald JK, Freedman DJ, Assad JA (2011) Generalized associative representations in parietal cortex. Nat Neurosci 14:1075–1079.

Freedman DJ, Assad JA (2006) Experience-dependent representation of visual categories in parietal cortex. Nature 443:85–88.

Freedman DJ, Assad JA (2016) Neuronal mechanisms of visual categorization: an abstract view on decision making. Annu Rev Neurosci 39:129–147.

Gibson JJ (1979) The ecological approach to visual perception. Boston: Houghton-Mifflin.

Goldstone R (1994) Influences of categorization on perceptual discrimination. J Exp Psychol Gen 123:178–200.

Hayden BY, Gallant JL (2013) Working memory and decision processes in visual area V4. Front Neurosci 7:18.

Law C-T, Gold JI (2008) Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. Nat Neurosci 11:505–513.

Mirabella G, Bertini G, Samengo I, Kilavik BE, Frilli D, della Libera C, Chelazzi L (2007) Neurons in area V4 of the macaque translate attended visual features into behaviorally relevant categories. Neuron 54:303–318.

Muhle-Karbe PS, Duncan J, De Baene W, Mitchell DJ, Brass M (2017) Neural coding for instruction-based task sets in human frontoparietal and visual cortex. Cereb Cortex 27:1891–1905.

Myers NE, Rohenkohl G, Wyart V, Woolrich MW, Nobre AC, Stokes MG (2015) Testing sensory evidence against mnemonic templates. Elife 4:e09000.

Sasaki R, Uka T (2009) Dynamic Readout of Behaviorally Relevant Signals from Area MT during Task Switching. Neuron 62:147–157.

Scolari M, Byers A, Serences JT (2012) Optimal deployment of attentional gain during fine discriminations. J Neurosci 32:7723–7733.

Summerfield C, Luyckx F, Sheahan H (2020) Structure learning and the posterior parietal cortex. Prog Neurobiol 184:101717.

Woolgar A, Thompson R, Bor D, Duncan J (2011) Multi-voxel coding of stimuli, rules, and responses in human frontoparietal cortex. Neuroimage 56:744–752.

Xin Y, Zhong L, Zhang Y, Zhou T, Pan J, Xu N-L (2019) Sensory-to-category transformation via dynamic reorganization of ensemble structures in mouse auditory cortex. Neuron 103:909–921.