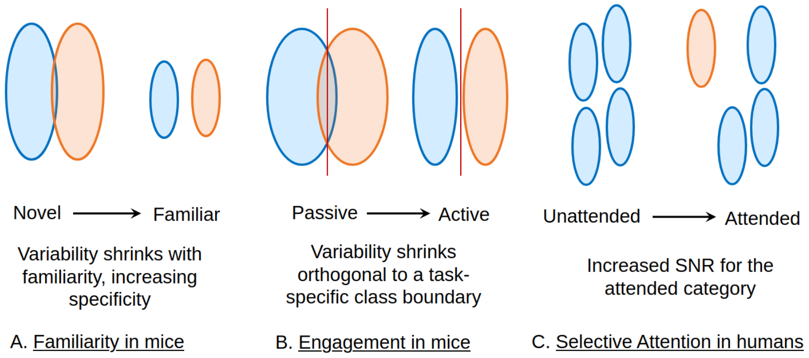
The Variability of Representations in Mice and Humans Changes with Learning, Engagement, and Attention

Praveen Venkatesh[*,1,2], Corbett Bennett[1], Sam Gale[1], Juri Minxha[3,4], Hristos Courellis[3,4], Greggory R. Heller[5], Tamina K. Ramirez[6], Séverine Durand[1], Ueli Rutishauser[3,4], Shawn Olsen[1], Stefan Mihalas[†,1]

Abstract

In responding to a visual stimulus, cortical neurons exhibit a high degree of *variability*, and this variability can be correlated across neurons. In this study, we use recordings from both mice and humans to systematically characterize how the variability in the representation of visual stimuli changes with learning, engagement and attention. We observe that in mice, familiarization with a set of images over many weeks reduces the variability of responses, but does not change its shape. Further, switching from passive to active task engagement changes the overall shape by shrinking the neural variability only along the task-relevant direction, leading to a higher signal-to-noise ratio. In a selective attention task in humans wherein multiple distributions are compared, a higher signal-to-noise ratio is obtained via a different mechanism, by mainly increasing the signal of the attended category. These findings show that representation variability can be adjusted with task needs. A potential speculative role for variability, consistent with these findings, is that it helps generalization.


Novel → Familiar
Variability shrinks with familiarity, increasing specificity
Passive → Active
Variability shrinks orthogonal to a task-specific class boundary
Unattended → Attended
Increased SNR for the attended category

A. Familiarity in mice B. Engagement in mice C. Selective Attention in humans

1. Introduction

The activity of cortical neurons in response to a stimulus can be extremely variable even in highly standardized recordings (de Vries et al., 2020; Siegle et al., 2021), despite the fact that neurons *have the capacity* to be very reliable (e.g., peripheral neurons: Dong et al., 2013). This variability limits the fidelity of sensory cortical coding (Rumyantsev et al., 2020): as reviewed by Averbeck et al. (2006), this "noise" in neural activity across trials is not independent between neurons, and coding bounds are mainly limited by correlated noise (Kanitscheider et al., 2015; Moreno-Bote et al., 2014; Pitkow et al., 2015).

While the geometry of trial-to-trial variability has been studied in detail for simple discrimination tasks (Rumyantsev et al., 2020), from a computational perspective, the role of this variability is less clear. One proposed role is that variability encodes perceptual

[*] praveen.venkatesh@alleninstitute.org; [†]stefanm@alleninstitute.org; [1]Allen Institute; [2]University of Washington; [3]Division of Biology and Biological Engineering, California Institute of Technology; [4]Dept. of Neurosurgery, Cedars-Sinai Medical Center; [5]Massachusetts Institute of Technology; [6]Columbia University; J. Minxha is now with Apple Inc.

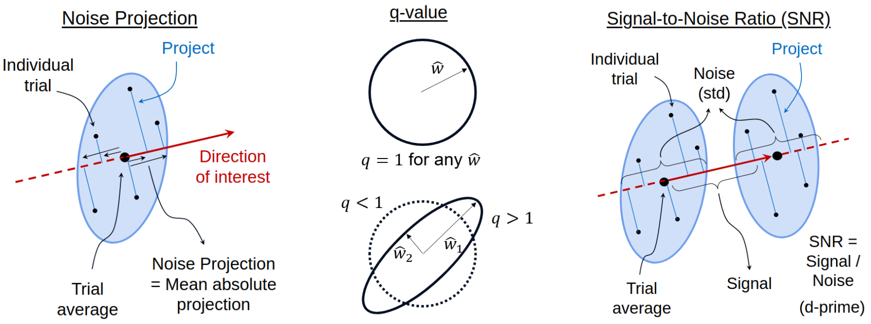


Figure 1: Depictions of metrics used to measure trial-to-trial variability. The noise projection measures the mean absolute deviation of variability in a direction of interest. The q-value computes the variance in a direction of interest, and normalizes it by the average variance over all directions. The signal-to-noise ratio is the distance between the trial-averages of two distributions, divided by the standard deviation along the same direction.

uncertainty through a *sampling-based probabilistic representation* (Orbán et al., 2016). Under this assumption, anisotropic variability helps with generalization, with higher variance along directions which require generalization and lower variance along directions which require specificity (Shang et al., 2021). If trial-to-trial variability in responses evolved to help generalization, we can make several predictions:

1. While it is important to generalize, especially for novel stimuli, more specialization is needed with increased familiarity. We expect a decrease in variability with familiarity.
2. Sometimes, a task requires the creation of a category boundary between items that are otherwise generalized over. Then, we expect a change in the signal to noise ratio along the task-relevant direction.
3. Sometimes, a task requires a dynamic change in the response category, which is not over-learned via experience, but rather instructed through an attentional cue. Then, we expect a change in the signal-to-noise ratio along the directions which are relevant at that moment.

To address predictions 1 and 2, we trained mice on a change-detection task, in which mice watched a sequence of flashes of the same image for a variable number of presentations, and had to respond with a lick when the image changed. For prediction 1, we compared the variability in representation for a set of eight images to which the mice were exposed for multiple weeks, against six novel images. For prediction 2, we compared the first hour of recording, which consisted of an active task component, with the following hour consisting of passive replay of the same sequence of stimuli (with the lick-port withdrawn, so no response could be made). To address prediction 3, we used recordings from humans shown images belonging to one of four categories, and instructed to provide yes/no responses for each stimulus, given a target category.

Metrics to Measure Variability. We quantify the extent of variability in different directions using three different metrics: (i) the noise projection, which measures the extent of variability in a given direction; (ii) the q-value, a measure of the *relative* variability in a given direction, compared to all other directions; and (iii) the signal-to-noise ratio (SNR),

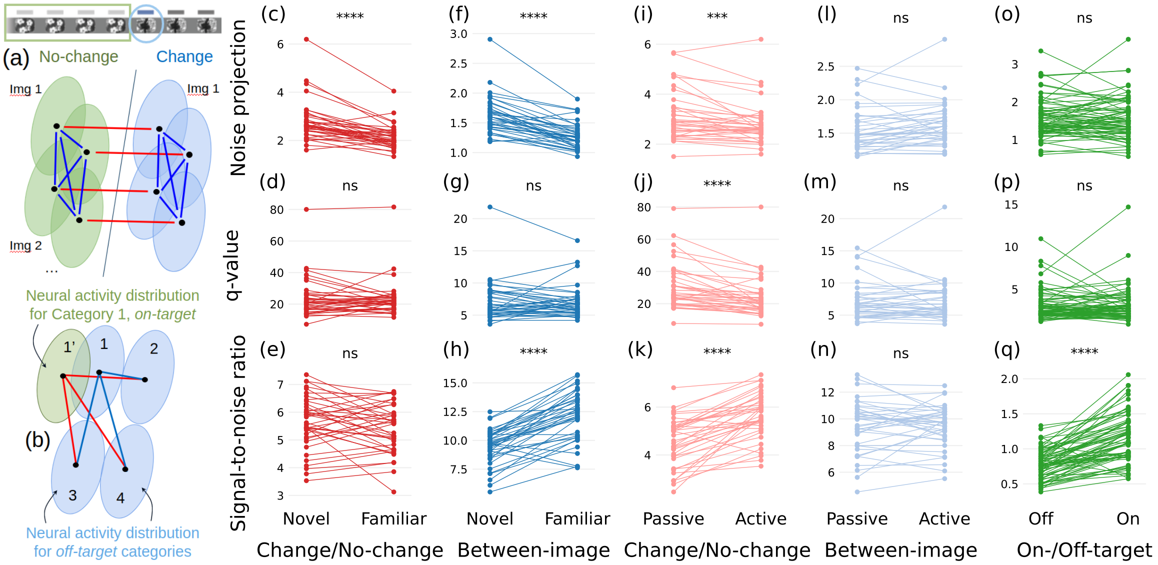


Figure 2: (a,b) Directions along which variability is measured in (a) the mouse and (b) the human dataset. (c-n) Different variability metrics (on rows) computed for (c-e,i-k) the change/no-change direction and (f-h,l-n) the between-image direction, for (c-h) novel vs. familiar stimuli and (i-n) active vs. passive sessions in mice. Lines correspond to different mice. (o-q) Metrics in two directions comparing unattended ("off-target") and attended ("on-target") categories in humans. Lines correspond to different sessions and target categories.

which measures the (linear) distinguishability between two distributions (depicted in Fig. 1; see Appendix A for details).

2. Results

We explore how the geometry of variability changes with stimulus familiarity and task engagement in mice on a dataset we collected at the Allen Institute (2022), employing a visual change detection task. Mice were presented one of eight images in 250ms flashes separated by 500ms gray screens. The image *changed* after a variable number of flashes; mice had to lick to receive a reward when the image changed. Apart from the "Familiar" session with eight images that were seen during training, we also recorded an additional "Novel" session with six new images and two familiar images. In each recording session, after one hour of "active" task engagement, the mice were replayed the same sequence of stimuli for "passive" viewing. Neural activity was recorded using six neuropixels probes targeting visual cortical regions. We consider the variability in the representation defined by the spike counts of visual cortical neurons in a 50–125ms window after stimulus onset.

Evidence from stimulus familiarity in mice. To examine how variability changes with familiarity, we measure the variability of neural activity for familiar and novel images in the Allen Institute dataset. We consider two directions: (i) that between the trial-averages under change and non-change conditions for each image (red lines in Fig. 2a); (ii) the direction between every pair of images *within* change and non-change conditions (blue lines in Fig. 2a). The former is a task-relevant direction, while the latter is ethologically relevant.

We compute the noise projection, the q-value and the SNR, for each image (in the change/no-change direction) and for every pair of images (in the between-image direction), and average over all images or image pairs (details in Appendix B). Going from 'Novel' to 'Familiar', we observe a reduction in the noise projection in both the change/no-change direction (22.5% decrease in median across mice) and in the between-image direction (21.6% decrease in median across mice; see Fig. 2c,f). However, we are unable to detect a statistically significant decrease in the q-value (−3.5% median decrease in the change/no-change direction, −2.9% decrease in the between-image direction; one-sided signed-rank test; Fig. 2c–h), suggesting an overall shrinkage in neural variability. The shrinkage in variability is accompanied by increased SNR in the direction between images (28.0% increase in median across mice), suggesting that familiarity increases distinguishability of ethologically relevant stimuli (depicted in Fig. A in the graphical abstract).

Evidence from task engagement in mice. To examine the effect of task engagement on the geometry of neural variability, we compute the same metrics, in the same two directions, between active and passive sessions (details in Appendix B; results in Fig. 2i–n). We find a small but statistically significant reduction in variability in the change/no-change direction, going from passive to active task engagement (6.7% reduction in median across mice; Fig. 2i). Moreover, this decrease is highly specific, as indicated by a decrease in the q-value (17.7% decrease in median across mice; Fig. 2j). The decreased noise is also accompanied by increased distinguishability between non-change and change stimuli (22.4% increase; Fig. 2k; also depicted in Fig. B in the graphical abstract). In contrast, we are unable to detect statistically significant changes in any metric in the direction between images (median decreases of −8.7%, 5.8% and 2.5% in Figs. 2l–n respectively). These results suggest that the brain is able to dynamically decrease variability in a task-specific direction when moving from a passive to an active task context.

Evidence from selective attention in humans. To understand how variability changes when selective attention is paid to one of several stimulus categories, we examine a dataset that we previously analyzed (Minxha et al., 2020). Single unit data was collected from electrodes implanted in the medial temporal lobe (MTL) and medial frontal cortex (MFC) of patients performing a categorization task. Patients were given a target category and had to provide yes/no responses to stimulus images from one of four categories (fruits, cars, human faces and monkey faces). We study the variability in the representation given by spike counts of MTL and MFC neurons in a 100–700ms window after stimulus onset.

To understand the effect of attending to a particular category on the geometry of neural variability, we compared two different directions: (i) the direction between a given category when it was attended-to (i.e., it was the "on-target" category) and the other categories when they were unattended (refer red lines in Fig. 2b); and (ii) the direction between the same given category when it was unattended (i.e., "off-target") and the other categories when they were also unattended (refer blue lines Fig. 2b). This comparison specifically tells us about the change in neural representation variability due to selective attention paid to a given category[1] (details of the analysis are in Appendix C).

[1] Note that the distributions here are over multiple trials of a single *category*, which includes *several* distinct images. In contrast, the task on mice had distributions corresponding to multiple trials of a *single* image.

Fig. 2o–q present the above comparison for MTL neurons (similar results are observed on MFC neurons as well), with individual lines for each category and session. Figs. 2o,p do not show a statistically significant change in the noise projection or in the q-value between these conditions (−4.1% and −6.3% decrease in median across target categories and sessions). However, there is a highly significant increase in the SNR going from an unattended to an attended state (42.9% increase in median across target categories and sessions), suggesting that the brain dynamically changes the distinguishability of different categories based on task context (depicted in Fig. C in the graphical abstract).

3. Discussion

Our work presents a confirmation of three predictions for the normative theory that the role of variability is to promote generalization. To our surprise, the mechanisms of increased SNR in engagement and attention were different. We were only able to confirm these predictions *indirectly*, since we cannot experimentally manipulate the structure of variability directly to test generalization performance.

Our results complement prior observations in the attention literature (Cohen and Maunsell, 2009; Ni et al., 2018; Rabinowitz et al., 2015), which showed that attention decreased shared correlations in the V4 visual area of monkeys, providing a plausible explanation for observed improvements in behavioral performance. Our results on attention do not cover visual cortex, but we see increased SNR for the attended category in the higher order areas of MTL and MFC. We also see decreased variance and increased SNR in a task-specific direction, in visual cortex, when mice are engaged in a task.

This paper does not address what mechanisms can give rise to such changes. Mechanisms that can affect neural correlations in a state-dependent manner have been discussed in the literature (Doiron et al., 2016), but we leave further investigation of this to future work. We have also not considered how classical adaptation, in response to repeated stimulus presentation, interacts with the changes in variability we observe. Future work will also consider whether the variability across mice (or across humans) in Fig. 2 can explain behavioral differences observed between subjects.

An interesting avenue for follow-on studies is to introduce anisotropic trial-to-trial variability in artificial neural networks, and then examine the degree to which they generalize, and the directions in representation space they generalize over.

References

Bruno B Averbeck, Peter E Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature reviews neuroscience*, 7(5):358–366, 2006.

Marlene R Cohen and John HR Maunsell. Attention improves performance primarily by reducing interneuronal correlations. *Nature neuroscience*, 12(12):1594–1600, 2009.

Saskia EJ de Vries, Jerome A Lecoq, Michael A Buice, Peter A Groblewski, Gabriel K Ocker, Michael Oliver, David Feng, Nicholas Cain, Peter Ledochowitsch, Daniel Millman, et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature neuroscience*, 23(1):138–151, 2020.

Brent Doiron, Ashok Litwin-Kumar, Robert Rosenbaum, Gabriel K Ocker, and Krešimir Josić. The mechanics of state-dependent neural correlations. *Nature neuroscience*, 19(3):383–393, 2016.

Yi Dong, Stefan Mihalas, Sung Soo Kim, Takashi Yoshioka, Sliman Bensmaia, and Ernst Niebur. A simple model of mechanotransduction in primate glabrous skin. *Journal of neurophysiology*, 109 (5):1350–1359, 2013.

Allen Institute. Visual behavior neuropixels dataset overview, 2022. URL https://portal.brain-map.org/explore/circuits/visual-behavior-neuropixels.

Ingmar Kanitscheider, Ruben Coen-Cagli, and Alexandre Pouget. Origin of information-limiting noise correlations. *Proceedings of the National Academy of Sciences*, 112(50):E6973–E6982, 2015.

Juri Minxha, Ralph Adolphs, Stefano Fusi, Adam N Mamelak, and Ueli Rutishauser. Flexible recruitment of memory-based choice representations by the human medial frontal cortex. *Science*, 368(6498):eaba3313, 2020.

Rubén Moreno-Bote, Jeffrey Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and Alexandre Pouget. Information-limiting correlations. *Nature neuroscience*, 17(10):1410–1417, 2014.

Amy M Ni, Douglas A Ruff, Joshua J Alberts, Jen Symmonds, and Marlene R Cohen. Learning and attention reveal a general relationship between population activity and behavior. *Science*, 359 (6374):463–465, 2018.

Gergő Orbán, Pietro Berkes, József Fiser, and Máté Lengyel. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron*, 92(2):530–543, 2016.

Xaq Pitkow, Sheng Liu, Dora E Angelaki, Gregory C DeAngelis, and Alexandre Pouget. How can single sensory neurons predict behavior? *Neuron*, 87(2):411–423, 2015.

Neil C Rabinowitz, Robbe L Goris, Marlene Cohen, and Eero P Simoncelli. Attention stabilizes the shared gain of v4 populations. *Elife*, 4:e08998, 2015.

Oleg I Rumyantsev, Jérôme A Lecoq, Oscar Hernandez, Yanping Zhang, Joan Savall, Radosław Chrapkiewicz, Jane Li, Hongkui Zeng, Surya Ganguli, and Mark J Schnitzer. Fundamental bounds on the fidelity of sensory cortical coding. *Nature*, 580(7801):100–105, 2020.

Jiaqi Shang, Eric Shea-Brown, and Stefan Mihalas. Cortical representation variability aligns with in-class variances and can help one-shot learning. *bioRxiv*, pages 2021–01, 2021.

Joshua H Siegle, Xiaoxuan Jia, Séverine Durand, Sam Gale, Corbett Bennett, Nile Graddis, Greggory Heller, Tamina K Ramirez, Hannah Choi, Jennifer A Luviano, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592(7852):86–92, 2021.

Appendix A. Mathematical Definitions of the Variability Metrics

We measure the geometry of variability by considering the distribution of neural activity over multiple trials. We assume that the neural activity of a single neuron is given by a real number, e.g., its spike count within some temporal window. Then, in the datasets, the overall distribution of neural activity is an empirical distribution given by a set of vectors $x_i \in \mathbb{R}^N$, $i \in \{1, ..., T\}$, where N is the number of neurons and T is the number of trials.

Let $\{x_i\}$ represent the N-dimensional neural activity vector across T trials, and $\hat{w}$ be a unit vector in some direction of interest. Let $\bar{x} \in \mathbb{R}^N$ be the trial-averaged neural activity. Then, the three variability metrics—the noise projection, the q-value, and the signal-to-noise ratio—are given by:

$$\mathrm{NoiseProj}(\{x_i\}, \hat{w}) = \frac{1}{T} \sum_{i=1}^{T} |\hat{w}^{\mathsf{T}}(x_i - \bar{x})| \tag{1}$$

$$\mathrm{q}(\{x_i\}, \hat{w}) = \frac{\mathrm{Var}(\hat{w}^\mathsf{T}(x_i - \bar{x}))}{\frac{1}{N}\sum_{j=1}^{N} \mathrm{Var}(x_{ij} - \bar{x}_j)} \tag{2}$$

$$\mathrm{SNR}(\{x_i^{(1)}\}, \{x_i^{(2)}\}) = \frac{\|\bar{x}^{(1)} - \bar{x}^{(2)}\|}{\mathrm{Std}\big(\{\hat{v}^\mathsf{T}(x_i^{(1)} - \bar{x}^{(1)})\} \cup \{\hat{v}^\mathsf{T}(x_i^{(2)} - \bar{x}^{(2)})\}\big)}, \tag{3}$$

where $\hat{v} = (\bar{x}^{(1)} - \bar{x}^{(2)})/\|\bar{x}^{(1)} - \bar{x}^{(2)}\|$. In most cases in the results section, we refer to the noise projection and q-value between two distributions as well. In these cases, the noise projection and q-value are more precisely defined as:

$$\mathrm{NoiseProj}(\{x_i^{(1)}\}, \{x_i^{(2)}\}) = \frac{1}{2T}\sum_{i=1}^{T} |\hat{v}^\mathsf{T}(x_i^{(1)} - \bar{x}^{(1)})| + |\hat{v}^\mathsf{T}(x_i^{(2)} - \bar{x}^{(2)})| \tag{4}$$

$$\mathrm{q}(\{x_i^{(1)}\}, \{x_i^{(2)}\}) = \frac{\mathrm{Var}\big(\{\hat{v}^\mathsf{T}(x_i^{(1)} - \bar{x}^{(1)})\} \cup \{\hat{v}^\mathsf{T}(x_i^{(2)} - \bar{x}^{(2)})\}\big)}{\frac{1}{N}\sum_{j=1}^{N} \mathrm{Var}\big(\{x_{ij}^{(1)} - \bar{x}_j^{(1)}\} \cup \{x_{ij}^{(2)} - \bar{x}_j^{(2)}\}\big)}, \tag{5}$$

where $\hat{v}$ is as before, the unit vector along the line joining the centers of the two distributions.

On the Stability of our Estimates. It should be noted that the expressions provided here involve estimating the variance or standard deviation of one-dimensional quantities. We avoid computing the covariance matrix itself, since the dimensionality of our data is larger than the number of available trials. A more careful analysis of the variance of our estimates of these metrics is left to future work.

On the Skewness of Variability. In constructing our metrics, we ignore the precise shape of the distribution of neural variability, and assume that it is approximately ellipsoidal. In practice, if the variability is highly anisotropic, with different degrees of skew along different directions, our metrics would not capture these effects, since they only consider up to the second moment. This could occur, for instance, with Poisson spike counts at low rates, wherein a Poisson distribution with a rate of λ has a positive skew of $1/\sqrt{\lambda}$. A more careful analysis of the impact of such effects are beyond the scope of the current study and could be taken up in future work.

Appendix B. Analysis Pipeline for the Mouse Data

1. We only considered mice with both familiar and novel sessions, and which had at least 20 neurons in each of the following visual cortical regions: VISp, VISl, VISal, VISam and VISpm. We sub-selected units that had a quality of 'good', with an SNR of at least 1, and with fewer than 1 inter-spike interval violations.

2. In each session, we computed the neural activity by counting the spikes of each unit in a 50–125ms time window after stimulus onset.

3. Stimulus flashes that corresponded to a 'change' were those trials in which the image changed (i.e., was different from the image in the preceding flash) and the mouse was engaged in the task (defined by having a rolling reward rate of at least 2 rewards/min).

4. Stimulus flashes that corresponded to a 'non-change' were those flashes that occurred between 4 and 10 flashes after the start of a behavioral trial and before the image

changed, which did not have an omission or follow an omission, on which the mouse did not lick, and while the mouse was engaged.

5. The three variability metrics in the change/no-change direction were computed separately for every image, between change and non-change distributions. The metrics were then averaged across all 8 images (for the familiar session) and across all 6 novel images (for the novel session; the two shared familiar images were ignored).
6. The variability metrics in the between-image direction were computed separately for every pair of images in the non-change class and for every pair in the change class. The variability metrics were then averaged across all images pairs across both classes.
7. The active-passive comparison was performed on the 6 novel images.
8. Each line in Fig. 2c-n corresponds to a different mouse. Statistical significance was assessed across mice using one-sided (paired) Wilcoxon signed-rank tests.

The selection criteria in Step 1 above yielded 39 mice with both familiar and novel sessions, with 525.15 ± 98.63 units in familiar sessions, and 423.97 ± 80.67 units in novel sessions (mean ± standard deviation). We also obtain 82.18 ± 19.57 trials of each image for the non-change condition and 23.78 ± 6.06 trials of each image for the change condition.

Appendix C. Analysis Pipeline for the Human Data

1. We only considered sessions with at least 10 recorded units in the medial temporal lobe (i.e., across both hippocampus and amygdala). We also removed all "control" sessions from the analysis.
2. We considered only the categorization blocks and ignored the memory blocks from the full experimental data (refer Minxha et al. 2020).
3. For each session, we computed the neural activity by counting spikes in a 100–700ms window after stimulus onset.
4. We sub-selected those trials on which the patient recorded a correct response to the categorization task.
5. For each category, we separated trials on which a stimulus image of that category appeared when the category was on-target and off-target.
6. For each category, we measured the three variability metrics between its on-target distribution and each of the three off-target distributions of the remaining categories, and averaged over them. We then compared these against the metrics that were computed between the selected category's off-target distribution and each of the three off-target distributions of the remaining categories, and averaged over.
7. The results in Fig. 2o-q collapse across all sessions (each patient could have more than one session), and across all target categories within each session. Statistical significance was assessed across all sessions and target categories using one-sided (paired) Wilcoxon signed-rank tests.

Using the selection criteria described above, we had a total of 20 sessions across 11 patients. Each session had 22.7 ± 8.76 units (mean ± standard deviation).