Invariance of neural representations cannot be validly inferred from neuroimaging decoding studies: Empirical and computational evidence

Sanjay Narasiwodeyar & Fabian A. Soto

Department of Psychology, Florida International University, Modesto A. Maidique Campus, 11200 SW 8th St, Miami, FL 33199

Abstract

Many research questions in sensory neuroscience involve determining whether the neural representation of a stimulus property is invariant to some irrelevant stimulus change (e.g., viewpoint-invariant face representation, or modality-invariant object representations). Most neuroimaging studies have studied invariance using operational tests that have only face validity, of which the most popular in recent years is the cross-classification test. A recently proposed theoretical framework suggests that operational tests of invariance commonly used in the neuroimaging literature, such as cross-classification, might lead to invalid conclusions. Here, we provide empirical and computational evidence supporting this theoretical insight. In our empirical study, we use encoding of orientation and spatial position in primary visual cortex as a case study, as previous research has established that these properties are not encoded in an invariant way. In a functional MRI study with human participants of both sexes, we show that the cross-classification test produces false positives, in many cases leading to the conclusion that orientation is encoded invariantly from spatial position, and that spatial position is encoded invariantly from orientation, in primary visual cortex. The results of two simulations further suggest that the test can lead to the conclusion of invariance when no sensible definition of invariance holds at the neural level, and that encoding strategies known to be used in cortex may easily lead to such false positives. On the other hand, we show that it is possible to provide evidence against invariance (i.e., context-dependent or configural encoding) through appropriate theory-driven decoding tests.

Many research questions in sensory neuroscience involve determining whether the neural representation of a stimulus property is invariant to some irrelevant stimulus change. Most human neuroimaging research has studied invariance using operational tests, which have face validity but are difficult to link to invariance at the level of neural representations. Probably the most widely used of such tests is cross-classification (or cross-decoding; Allefeld and Haynes, 2014; Anzellotti and Caramazza, 2014; Kaplan et al., 2015), illustrated in Figure 1. The first step in cross-classification is to train a classifier to decode a particular stimulus feature, such as whether a presented face is male or female, from patterns of fMRI activity observed across voxels. The second step is to test the trained classifier with new patterns of fMRI activity, this time obtained from presentation of the same stimuli, but changed in an irrelevant property, such as head orientation. If accuracy with these test data is higher than chance, then the conclusion is that the neural representation of the target feature is invariant to changes in the irrelevant property, within the area from which the fMRI activity was obtained.

The cross-classification test has been used to provide evidence for invariant (or "tolerant") encoding of face identity across viewpoint (Anzellotti et al., 2014), object category and viewpoint across spatial position (Ramirez et al., 2014), object category across shape (and vice-versa; Kaiser et al., 2016), motor actions across modalities (Etzel et al., 2008), place of speech articulation features across manner of articulation (Archila-Melendez et al., 2018), object category (Man et al., 2012) or face identity (Anzellotti and Caramazza, 2017) across stimulus modality, word semantic category across stimulus modality (Akama et al., 2012), learned category labels across categorization tasks (Soto et al., 2013), and semantic word representation across languages (Buchweitz et al., 2012), among others (for a review, see Kaplan et al., 2015).

Despite its wide use, recent theoretical work shows that cross-classification and similar operational tests of invariance can lead to invalid conclusions (Soto et al., 2018). In general, it can be shown that no neuroimaging decoding test can provide evidence in favor of invariance. Rather, such a test can only provide evidence against invariance in a given brain region (see section *Tests of Invariance* below). This theoretical work may seem too abstract and unconvinc-

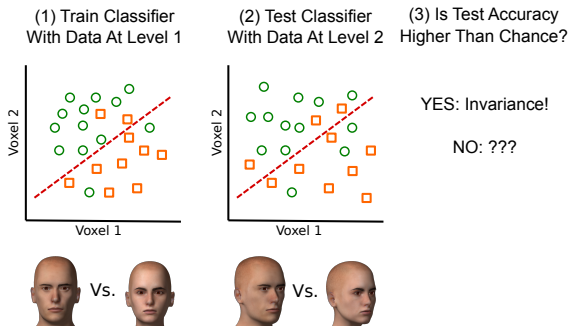

Fig. 1: Cross-classification test. (1) A classifier is trained to discriminate between levels of a target dimension (facial gender) in a given level of the irrelevant dimension (viewpoint). The axes represent activity in two exemplary voxels in response to stimuli. Each data point represents a single trial. (2) The same classifier is then used to classify data from a second level of the irrelevant dimension (different viewpoint). (3) If classification accuracy is above chance levels across viewpoints, then invariance is said to hold according to this test.

ing to the experimental neuroscientists who use the cross-classification test, and it assumes a strict definition of invariance that might differ from what most researchers are attempting to measure.

Here, we provide empirical and computational evidence supporting the theoretical insight that the cross-classification test, and neuroimaging decoding tests in general, cannot provide valid evidence to infer invariance of neural representations, for any sensible definition of invariance at the neural level. We use decoding of orientation and spatial position in V1 to show that the cross-classification test leads to conclusions of invariance when such invariance is known to not hold. The receptive fields of V1 neurons show narrow selectivity for both orientation and spatial position (Hubel and Wiesel, 1968; for reviews, see Issa et al., 2008; Ng et al., 2007). Such selectivity does not permit invariance across the scale of stimulus changes used in the present study. We show that the cross-classification test produces false positives, sometimes leading to the conclusion that, in V1, orientation is encoded invariantly from spatial position, and that spatial position is encoded invariantly from orientation. Simulations further suggest that the test can lead to the conclusion of invariance when neural representations do not satisfy any sensible definition of invariance, and that encoding schemes found in cortex may easily lead to such conclusions. On the other hand, as suggested by the theory, we show that it is possible to provide evidence against invariance (i.e., context-dependent or configural encoding) through appropriate theory-driven tests.

Materials and Methods

Participants

Five healthy volunteers (ages 19–27, three female) from Florida International University participated in the experiment; all had normal or corrected-to-normal vision. The study protocol was approved by Florida International University's Institutional Review Board and by the Center for Imaging Science Steering Committee. All subjects gave written consent to experimental procedures before participating in the experiment.

Stimuli

All stimuli were generated using Psychopy v.1.85.0 (Peirce, 2007). Images were displayed on a 40" Nordic Neurolab LCD InroomViewing Device, placed at the rear entrance of the scanner bore. Subjects viewed the screen via an angled mirror attached to the head coil. Visual stimuli were full-contrast square-wave gratings with a spatial frequency of 1.5 cycles per degree of visual angle (similar to Alink et al., 2013; Pratte et al., 2016; Sengupta et al., 2017), a frequency known to drive V1 responses strongly (Henriksson et al., 2008), shown through a wedge-shaped aperture window that spanned from 1.5° to 10° of eccentricity and 100° of polar angle (Figure 2). The aperture window had four possible locations, starting at 20°, 80°, 200°, and 260° of rotation. The square-wave gratings were oriented in one of four angles for each trial: 0°, 45°, 90°, 135°. The phase of the gratings was randomly changed every 250ms, to reduce retinal adaptation and afterimages.

Task and Procedures

To ensure that the data used to train a classifier in decoding analyses (see below) was independent from the data used to test the classifier and compute measures of performance, training trials and testing trials were presented on separate acquisition runs. Training and testing runs were identical in all aspects except one: the positions of the aperture window were restricted to 20° and 200° of rotation for the training runs, while testing runs included all four positions (Figure 2). During stimulus presentation, the phase of the grating was randomly shifted every 250 ms. The orientation of each grating was randomly chosen on each trial, while the spatial position of the window

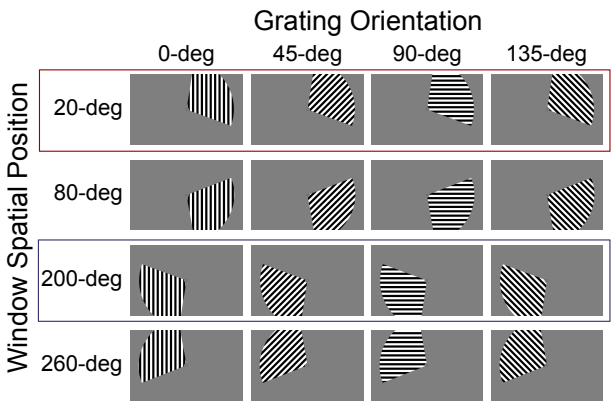

Fig. 2: Stimuli were composed of oriented gratings (dimension 1) presented in a windowed spatial position (dimension 2). Each trial consisted of a single combination of orientated gratings and spatial position. Training runs were composed of stimuli presented only in spatial positions 20° and 200° (highlighted through red and blue boxes). Testing runs included all sixteen stimulus combinations.

changed sequentially in a pre-determined manner. In training runs, the aperture window switched between 20° and 200° on every trial. In testing runs, the aperture window cycled through 20°, 200°, 80°, 260°, in that order. For both training and testing runs, each combination of spatial position (two or four levels) and orientation (four levels) was presented 35 times in a single acquisition session. Each subject went through 4 identical acquisition sessions to yield a total of 135 presentations of a given combination of orientation and spatial position (see all combinations in Figure 2) for both training and testing trials types. This large longitudinal sample size (3,240 trials total per participant) was chosen to focus our analyses on data at the level of individual participants (see *Statistical Analyses* below).

On each trial, a single grating was presented for 3s, followed by a 3s inter-trial interval. All runs began with a 10s fixation period and ended with a 1 min rest period. The training runs lasted for 5 mins and 43 secs, and the test runs lasted for 10 mins and 13s. Due to experimenter error during data acquisition, a portion of training trials were lost for participants 1 and 2. To compensate for the reduced number of training trials, we collected an additional session of data from subject 2, resulting in about 123 training trials and 112 testing trials per stimulus. For subject 1, we simply set aside half of the testing trials for training purposes and used the other half for testing; the number of testing trials for non-trained values of the irrelevant dimension remained the same as for all other participants.

The participants' task was to look at a small black ring presented in the center of the screen (similar to Alink et al., 2013). The black ring had a small gap that randomly switched position throughout the trial. Participants were asked to continuously report the side of the gap (left or right) by pressing the corresponding button. The task had the purpose of forcing participants to fixate at the center of the screen, and to draw attention away from the stimuli.

Functional Imaging

Imaging was performed with a Siemens Magnetom Prisma 3T whole-body MRI system located at the Center for Imaging Science, Florida International University. A volume RF coil (transmit) and a 32-channel receive array were used to acquire both functional and anatomical images. Each subject participated in four identical MRI sessions. During each session, a high-resolution 3D anatomical T1-weighted volume (MPRAGE; TR, 2.4s; TI, 1.1s; TE, 2.9 ms; flip angle, 7°; voxel size, 1×1×1 mm; FOV, 256 mm; 176 sagittal slices) was obtained, which served as the reference volume to align all functional images. During the main experiment, functional images were collected using a T2*-weighted EPI sequence (TR, 1.5 s; TE, 30 ms; flip angle, 52°; sensitivity encoding with acceleration factor of 4). We collected 60 transversal slices, with resolution of 2.4×2.4×2.4 mm, and FOV of 219mm. The first six volumes in each run were discarded to allow T1 magnetization to reach steady state.

Statistical Analyses

All data analyses, including multi-voxel decoding and tests of invariance, were performed on the individual data of each participant. In designing our experiment, we favored collection of a large amount of data per participant (3,240 trials, about 8 hours of scanning) rather than a large number of participants. Each separate analysis can be considered a replication of a single-subject experiment. With our sample sizes (n=135 per stimulus), our tests can detect a 6% difference from chance in classifier performance with 85% power, an 8% drop in classifier performance with >80% power, and kernel density estimate error is maximally reduced, according to simulation studies (Fortmann-Roe et al., 2012).

Region of Interest

The boundaries of V1 are commonly found using a functional localizer procedure. However, previous work has shown that the boundaries of V1 can be

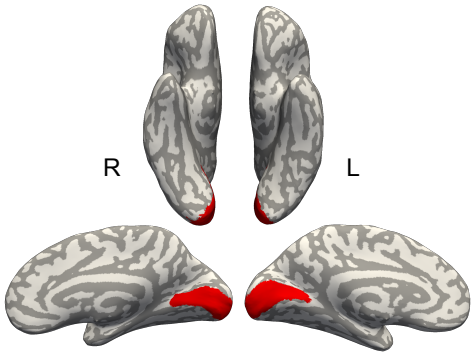

Fig. 3: Example V1 mask obtained from one participant in this study.

accurately estimated from cortical folds, without the need for a functional localizer (Hinds et al., 2008). Additionally, recent results show that the definition of V1 boundaries using the algorithm proposed by Hinds et al. (2008) has a precision that is equivalent to 10-25 minutes of functional mapping (Benson et al., 2012). Therefore, we applied the Hinds et al. (2008) algorithm, implemented in Freesurfer 6.0 (Fischl, 2012), to the anatomical T1-weighted images, to define the boundaries of V1 in each participant and obtain an ROI mask. The obtained V1 mask was then converted into a binary mask, and transformed to the individual's functional scan space (the averaged volume of the first functional run was used as a target) using linear registration with FLIRT. An example mask, obtained from one participant in the study, is displayed in Figure 3.

BOLD Data Preprocessing

Data were processed and analyzed using *nipype* Python wrappers for FSL (Gorgolewski et al., 2011; Jenkinson et al., 2012). Basic preprocessing of functional data included skull stripping, slice time correction, and head motion correction using MCFLIRT. All functional runs for a given subject were then aligned to an averaged volume of the first functional run for the same subject. This step ensured that the entire time-series for each subject lay in the same co-ordinate space. The aligned time-series was then concatenated into a single time-series file for further processing. The concatenated series for each subject was de-trended using a Savitzky-Golay filter with a polynomial order of 3 and a window length of 81 secs (Pedregosa et al., 2015).

Deconvolution

Using the obtained V1 mask, time-series from V1 voxels were extracted for further analysis. Single-trial activity estimates were obtained via a data-driven deconvolution technique in which deconvolved neural activation values and a model of the hemodynamic response function (HRF) are estimated together (Pedregosa et al., 2015). Unlike other methods that hold the shape of the HRF constant across voxels, this technique allows the shape of the HRF to be different in each voxel, resulting in more accurate activity estimates. The model is implemented via the *hrf_estimation* Python package v. 1.1 (`https://pypi.org/project/hrf_estimation/`). The *hrf_estimation* package presents 10 different options for HRF modeling, with varying options for the HRF basis function and for the General Linear Model estimation technique. To select the optimal combination of HRF and estimation method, we performed a cross-validated decoding analysis using data from the training runs of a single participant (data from the testing runs was not used in this pre-analysis). First, we generated activity estimates from all possible model combinations (estimation method and HRF). Then, for each model, we trained and tested an SVM classifier to decode orientations from a portion of the training set, and tested the classifier with the remaining data. We chose the Rank-1 General Linear Model with a 3-basis-functions HRF model, based on the fact that it yielded the highest testing accuracy score.

Decoding Analysis

To decode stimulus types based on voxel-wise activity patterns, we used a Nu-support vector machine (NuSVC) classifier with a linear basis function implemented via the Python package *scikit-learn* v. 0.19.1 (Pedregosa et al., 2011). We used the de-convolved activity patterns from V1 voxels as inputs to the classifier, while trial-specific stimulus values (either orientation or spatial position) were provided as labels.

To decode orientation, we employed two separate classifiers, corresponding to the two different spatial positions (irrelevant dimension) at which the oriented gratings were presented during the training runs of the experiment (see Figure 2). Each classifier was trained to decode grating orientation (0°, 45°, 90°, and 135°) using only trials in which a specific spatial position was presented. However, the classifier was then tested with data collected from independent test runs at all four spatial positions. This resulted in an accuracy estimate at the training position, as well as at the other three spatial positions. For exam-

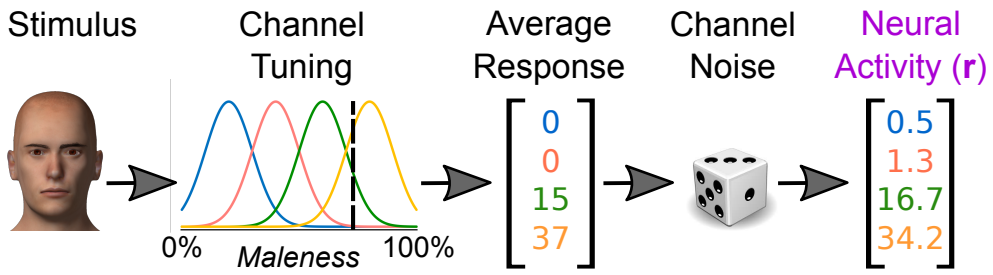

Fig. 4: Population encoding model. This figures illustrates how stimulus information can be encoded by a set of neurons in the brain. The encoding model consists of a set of channels that are tuned to specific stimulus values along a given dimension (e.g., maleness). When a stimulus with a particular value on the maleness dimension is presented, the channels respond according to their stimulus preferences. The channel responses are then perturbed by random channel noise. The final output represents a vector of noisy firing rates in response to a particular stimulus.

ple, to train the first classifier, we gathered all trials that were presented at spatial position 20°. After normalizing the data, the classifier was trained using leave-one-run-out cross-validation with data from the training runs. Cross-validation was used to optimized the *Nu* parameter of the classifier, to obtain the highest accuracies within the training set. A new classifier was then trained on all the training data using the chosen *Nu* parameter. This classifier was then tested with data from testing runs.

To decode spatial position, we employed four separate classifiers corresponding to the four levels of grating orientation (irrelevant dimension) that were presented during the training runs of the experiment. Each classifier was trained to decode spatial position (top-right vs bottom-left, see boxed stimuli in Figure 2) using only trials in which a specific grating orientation was presented. However, the classifier was then tested with data collected from independent test runs across all levels of grating orientation. This resulted in an accuracy estimate for the training grating orientation, as well as the other three levels of grating orientation. As in the orientation decoding procedure, we divided the data into independent training and test sets, performed normalization, and optimized the classifier's *Nu* parameter via leave-one-run-out cross-validation. One important difference is that spatial position decoding involved a two-class classification problem, where the classifiers had to discriminate between 20° and 200° spatial position of the stimulus window (the only two positions presented during training trials, see boxed stimuli in Figure 2). As the classifier was not trained to classify the 80° or 260° spatial positions, we dropped those trials from the testing data set in this analysis. This ensured that the model fitting and testing procedures remained consistent across both decoding analyses.

Tests of Invariance

Theoretical Definitions of Invariance at the Neural Level. In our previous work (Soto et al., 2018), we showed that population encoding models provide a formal theoretical framework within which it is possible to define several forms of invariance and independence of neural representations, and link such theoretical definitions to operational tests performed on neuroimaging or psychophysical data. Here we briefly summarize this framework to facilitate understanding of the invariance tests presented later.

Figure 4 shows a schematic representation of a population encoding model. In this model, the level of a stimulus dimension such as *Maleness* of a face is encoded through several neural populations, each represented in the figure by a curve with a different color. This curve indicates the average activity in the population as a function of different values of maleness. The output of the encoding model is a pattern of neural activity, which includes neural noise that is added during processing in the visual system, represented by the dice. Because of this neural noise, the pattern of responses is slightly different each time the same stimulus is presented. This means that the neural activity pattern is a random vector $\mathbf{r}$, which follows a particular *encoding distribution* conditional on the presented stimulus.

Strict Invariance. Formally, we define strict neural invariance as the case where the encoding distribution remains the same across any changes in an irrelevant dimension (Figure 5) (see Soto et al., 2018, where we refer to this property as encoding separability). For example, when strict invariance holds (see "YES" column in Figure 5), the neural encoding distribution for face maleness remains the same across changes in viewpoint. This property of strict invariance suggests

that a direct measurement of the channel responses to a face stimulus will reveal little differences when the viewpoint angle of the face is changed. On the other hand, strict invariance fails (see "NO" column in Figure 5) when the neural encoding distribution for face maleness changes with changes in viewpoint.

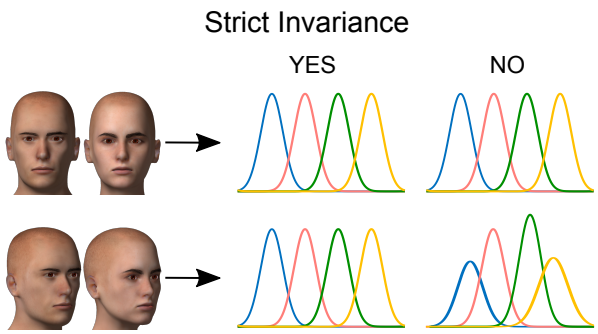

Fig. 5: Illustration of strict invariance at the neural level. In the example, the question is whether maleness is encoded invariantly from viewpoint. Left column (YES) represents a case where strict invariance holds. The tuning of channels encoding maleness remain the same across changes in viewpoint. Right column (NO) represents a case where strict neural invariance fails. The tuning of channels encoding maleness is altered in response to changes in viewpoint.

Partial Invariance. Many experimental neuroscientists may be dissatisfied by the definition of strict invariance presented above, as they are more interested in studying a form of partial invariance, or "tolerance", across changes in an irrelevant dimension. We define partial invariance as the case in which some properties of the tuning of neural populations are preserved across changes in an irrelevant dimension. Figure 6 shows two interesting cases. In *selectivity invariance*, the tuning of each population does not change with changes in the irrelevant dimension, although the overall response of the population is scaled. In *sub-population invariance*, strict invariance holds for only some neural populations, but not others. In the example presented in Figure 6, the red population shows strict invariance, but other populations change with changes in the irrelevant dimension of head orientation. Many other definitions of partial invariance are possible, and the conclusions based on the results of simulations presented below (particularly Simulation 1) apply to any sensible definition of invariance.

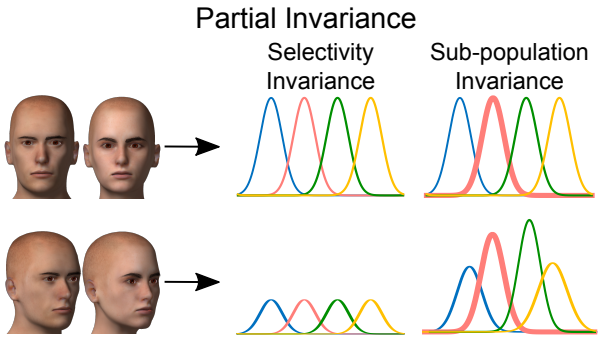

Fig. 6: Illustration of partial invariance at the neural level. Invariance is measured between two dimensions of interest – maleness and viewpoint. Selectivity invariance (left column) relates to cases where the tuning of encoding channels remains consistent across changes in viewpoint, but the response amplitudes are scaled. In sub-population invariance (right column), only some channels (red curve) retain their shapes across changes in viewpoint.

fMRI Decoding Tests In an fMRI experiment, we do not have direct access to neural activity. As shown in Figure 7, the neural activity is transformed into indirect activity measures, one per voxel, which themselves are contaminated by measurement error (again, represented by a dice in the figure). This transformation from the encoding distribution to a distribution of measurements is captured by what Figure 7 calls a measurement model. The result is a random vector $\mathbf{a}$, which again follows some distribution conditional on the presented stimulus.

Stimulus information can still be decoded from voxel-wise activity measures (Figure 7). Voxel-wise estimates of activity from an ROI can be used as inputs to a decoder that can be trained to predict stimulus values. Many possible operational tests of invariance can be proposed based on the results of a decoding study, but we applied three of them to our data: the cross-classification test, the classification accuracy invariance test, and the decoding invariance test. These tests are described below, and their theoretical relations to strict neural invariance are summarized in Figure 8 (for more details, see Soto et al. (2018)). The important conclusion that stems from prior theoretical work, as summarized in Figure 8, is that decoding tests cannot provide evidence *for* invariance, only evidence *against* it.

All the tests described below were implemented in Python expanded with *SciPy* v. 1.1.0 (`https://www.scipy.org/`) and *Statsmodels* v. 0.9.0 (`https://www.statsmodels.org/`). Plots were created using the *Matplotlib* library v. 2.2.2 (`https://matplotlib.org/`)

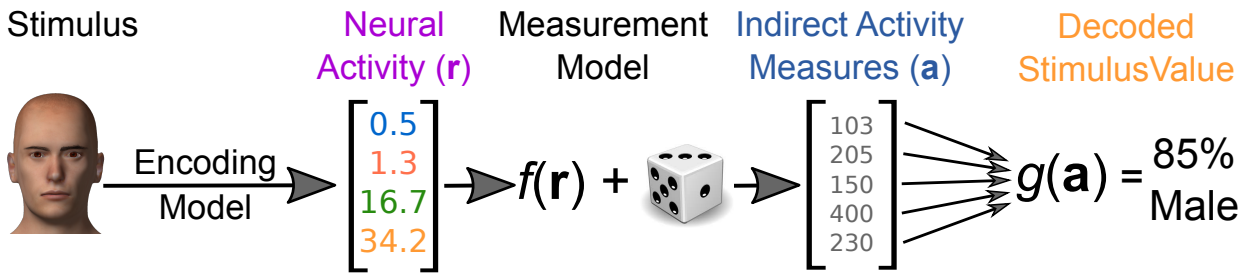

Fig. 7: Linking encoding models to neuroimaging decoding. The vector of neural activity generated by the encoding model is not directly accessible via neuroimaging methods. The neural activity is transformed via a measurement model into indirect measures of activity (e.g., BOLD response in a voxel). The measurement procedure adds noise to the observed response, contaminating the signal further. Stimulus values can be estimated from the indirect measures of activity by use of a decoder.

Cross-classification Test. The cross-classification invariance test is a well-known test in the literature that is meant to provide evidence for invariant representations directly from voxel-level activity estimates (e.g., Allefeld and Haynes, 2014; Anzellotti and Caramazza, 2014; Kaplan et al., 2015). As shown in Figure 1, the first step in this test is to train a classifier or another procedure to decode a particular stimulus feature from patterns of fMRI activity observed across voxels. In the figure, the feature that is being decoded is whether a face presented from a frontal viewpoint is male or female. The second step is to test the same classifier with new patterns of fMRI activity, this time obtained from presentation of the same stimuli changed in an irrelevant property, such as head orientation. If accuracy with this test data is higher than chance, then the conclusion is that that an invariant neural representation of the target feature is stored in the area from which the fMRI activity was obtained.

We implement the cross-classification invariance test by training a linear SVM, as described above, to classify levels of the target dimension (grating orientation or spatial position) while holding the level of the irrelevant dimension constant. For example, to decode orientation we start by training the SVM classifier to predict orientation in a given spatial position. Then, we test the accuracy of the classifier with data from independent test sets at the training position, as well as three other spatial positions (i.e., different levels of the irrelevant dimension). We tested whether each of these accuracies was above the chance level of 25% correct using a binomial test, and corrected the resulting p-values for multiple comparisons using the Holm-Sidak method. If classification accuracy was above chance at any of the testing spatial positions, then the cross-classification invariance test concludes that there is evidence for invariance of orientation from spatial position.

Classification Accuracy Invariance Test. Classification accuracy invariance is defined as the case where the probability of correct classification is exactly the same across all levels of the irrelevant dimension. This test is theory-driven, as it was suggested as a valid way to obtain information about invariance by theory (Soto et al., 2018). However, unlike cross-classification is designed to provide evidence *against* strict invariance, in line with the relations shown in Figure 8. While the test was independently developed from theory, we are aware of at least one prior study using a version of this test to obtain evidence of position-dependent encoding of object category information in lateral occipital cortex, and of position-dependent encoding of face viewpoint information in right fusiform face area (Ramirez et al., 2014).

This test uses the same estimates of classification accuracy described for the cross-classification test, but uses them to check whether there is a significant drop in performance from the training to the testing spatial positions. We first performed an omnibus Chi-Square test of the null hypothesis that accuracy does not depend on level of the irrelevant dimension. In addition, we tested accuracy at each testing spatial position against the training spatial position using a pairwise z test for proportions, and corrected the resulting p-values for multiple comparisons using the Holm-Sidak method.

Decoding Invariance Test. Linear classifiers like the linear SVM used here perform classification of a new data point by computing a decision variable z, representing the distance of the data point from the classifier's hyperplane separating two classes. When the decision variable is larger than some criterion value (usually zero), the output is one class, whereas

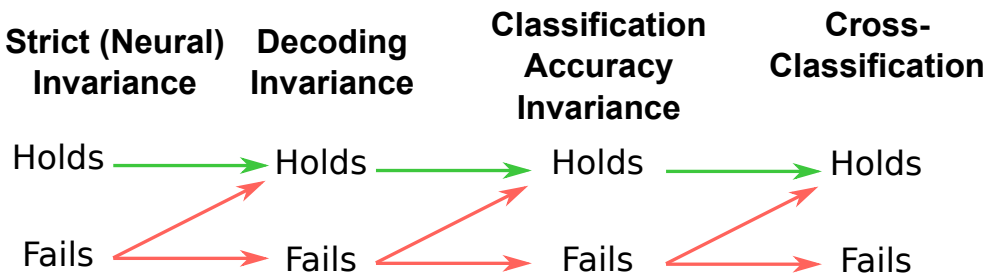

Fig. 8: Theoretical relations between different definitions of invariance, adapted from (Soto et al., 2018). Strict invariance (left column) is defined at a neural level within the framework of population encoding models (see Figure 5). All other columns represent tests of invariance that are based on decoding stimulus values from fMRI data. Each arrow should be read as representing an "if X, then Y" statement (e.g., "if strict invariance holds, then decoding invariance holds"). The set of green arrows suggest that if strict invariance holds at the neural level, invariance will hold in all decoding based tests of invariance. However, if strict invariance fails (red arrows), it is possible that invariance might hold or fail in any of the decoding based tests. Therefore, the theory suggests that positive evidence for the presence of strict invariance from any decoding based method is inconclusive. Instead, only failures of invariance can be reliably inferred from any decoding based test of invariance.

when the decision variable is smaller than the criterion the output is the other class.

This suggests that a better test of invariance should use not only classification accuracy scores, but rather the full distribution of such decision variables, or *decoding distribution*. Decoding invariance is defined as the case where the decoding distribution of a stimulus does not change across different levels of the irrelevant dimension (Figure 9). In Figure 9, the decoding distribution for a female face is shown for different viewpoints (irrelevant dimension). In this case, decoding invariance is said to fail when the estimated decoding distribution for a target stimulus feature (e.g., maleness) is significantly different across any two levels of the irrelevant dimension (viewpoint). The distance between the two distributions of interest provides a numerical estimate of deviations from decoding invariance. One way to measure this distance is through the $L1$ norm:

$$L1 = \int |p_1(z) - p_2(z)|\, dz, \tag{1}$$

where p_1 and p_2 represent the distributions of decoded values at levels 1 and 2 of the irrelevant dimension, respectively. The $L1$ distance between two distributions is represented in Figure 9 by the area highlighted in light yellow.

For each combination of values of the relevant and irrelevant dimensions, we obtained decision variables from the trained SVM linear classifier. These decision variables were used to estimate the decoding distribution using kernel density estimates (KDEs). A gaussian kernel and automatic bandwidth determination were used as implemented in the *SciPy* function *gaussian_kde*. Let $\hat{p}_{ij}(z)$ represent the KDE for a stimulus with value i on the target dimension and value j on the irrelevant dimension, evaluated at point z. Each $\hat{p}_{ij}(z)$ was evaluated at values of z going from -3 to 6, in 0.01 steps, indexed by k, which were confirmed to cover the range of observed decision variable values. Then an estimate of the summed $L1$ distances indicating deviations from decoding invariance was computed from all four KDEs obtained, according to the following equation:

$$L1_j^G = \sum_i \sum_k |\hat{p}_{i1}(z_k) - \hat{p}_{ij}(z_k)|\,. \tag{2}$$

, where $j = 1$ is the training level of the irrelevant dimension. The $L1_j^G$ (with G standing for *global*) simply takes an estimate of the $L1$ distance (obtained by discretizing the continuous decision variable z) defined in Equation 1 for each value of the relevant dimension, and then sums them together. We computed $L1_j^G$ separately for each value of the irrelevant dimension, or $j \neq 1$.

We used a permutation test to test whether each $L1_j^G$ statistic was significantly larger than expected by chance. In this test, the level of the irrelevant dimension j was randomly re-assigned to all data points, KDEs were estimated, and the $L1_j^G$ was computed according to Equation 2. This process was repeated 5,000 times, to obtain an empirical distribution for the statistic, from which accurate p-values were computed using the procedure proposed by (Knijnenburg et al., 2009). The resulting p-values were corrected for multiple comparisons using the Holm-Sidak method.

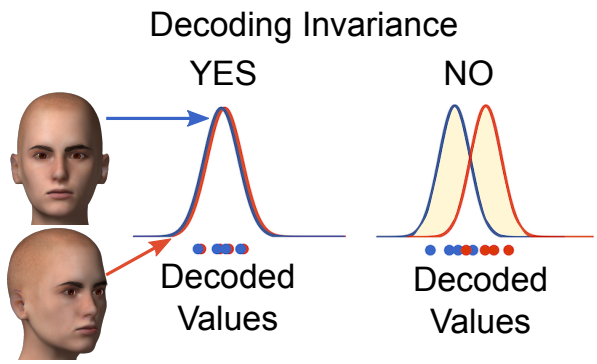

Fig. 9: The decoding invariance test. This is a theoretically-driven test of invariance that relies directly on the decoded stimulus distributions. Random error in the decoding of stimulus values gives rise to decoding probability distributions. In this example, decoding distributions are shown for some facial identity whose viewpoint changes. If the two decoding distributions corresponding to different viewpoints are the same (left), then decoding invariance is said to hold. If the decoding distributions are shifted for two levels of viewpoint (right), then decoding invariance is said to fail.

Simulations

The simulations described below were implemented in Python 2.6 extended with *Numpy* v. 1.16.2 (`https://numpy.org/`). The decoding analysis of simulated data was performed exactly as described for fMRI data in the sections *Decoding Analysis* and *Tests of Invariance* above, with the exception that the *Nu* parameter of the SVM was set to the default value of 0.5 rather than optimized based on cross-validation.

Model

In our simulations, we used a standard population encoding model (Pouget et al., 2003) and a linear measurement model. Both are common choices in encoding modeling of fMRI and EEG data, both in recent simulation work (e.g., Alink et al., 2018; Gardner and Liu, 2019; Liu et al., 2018), as well as in model-based data analysis (e.g., Brouwer and Heeger, 2009; Ester et al., 2013; Garcia et al., 2013; Van Bergen et al., 2015). We assumed a circular dimension with values ranging from -90 to 90, as is the case of grating orientation, but our conclusions apply to non-circular dimensions as well.

Encoding Model We used standard encoding models to represent the activity patterns of populations of neurons within a given voxel. Our encoding model was composed of several independent channels that acted as stimulus feature detectors. A channel could represent any number of neurons that have similar stimulus preferences. Each channel is highly tuned to a specific value along the target stimulus dimension, such that the channel's response becomes attenuated as we move away from the preferred value. The tuning function of a single channel is represented by a Gaussian function:

$$f_c(s) = r_c^{max} exp(-\frac{1}{2}(\frac{s - s_c}{\omega_c})^2), \tag{3}$$

, where r_c^{max} represents the maximum neural activity for channel c, the mean s_c represents the channel's preferred stimulus, and the standard deviation ω_c represents the width of the tuning function. The height of the tuning functions at any value along the stimulus dimension (i.e., $f_c(s)$) represents the average response of the channels to that particular stimuli.

We assume that the response of each channel r_c is a random variable with Poisson distribution:

$$P\left(r_c|s\right) = \frac{f_c\left(s\right)^{r_c} e^{-f_c(s)}}{r_c!}. \tag{4}$$

The full encoding model was composed of ten channels with activity described by Equations 3 and 4. Unless indicated otherwise below, we used a homogeneous population model, in which the parameters s_c were evenly distributed across all possible values of the dimension (i.e., from -90 to 90 degrees), and other parameters were fixed to the same values for all channels: $r_c^{max} = 10$, $\omega_c = 15$.

Figure 4 shows an example of the encoding process. When a stimulus with a value of 3 is presented to the model, the channel encoding distribution produces a vector of responses. Each element in this vector corresponds to the response of a particular channel. The channels with the strongest preference for the value 3 show the highest response in this vector. Since the response of neural populations are known to be noisy, channel noise is added to each element of the response vector. The final output is a noisy vector of channel responses that change slightly for repeated presentations of the same stimulus.

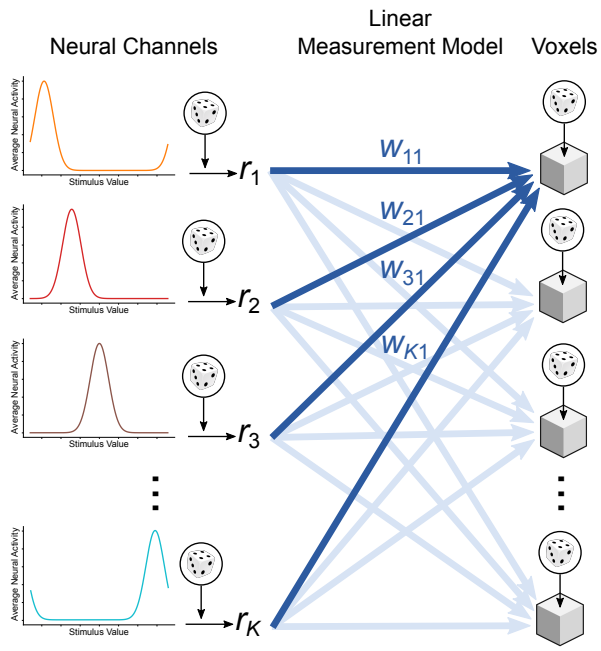

Fig. 10: Linear measurement model. The measurement model provides a link between neural encoding channels and voxel-wise activity measures. Activity in each voxel (represented by cubes) is a linear combination of neural channel responses ($\mathbf{r} = [r_1, r_2, r_3, ...r_K]$), plus independent random measurement noise (represented by the dice next to each cube)

Measurement Model Because neuroimaging studies produce only indirect measures of neural activity, a measurement model is required to link the neural responses of the encoding model with voxel-wise activity values. The measurement model is described by the following equation:

$$\mathbf{a} = \mathbf{r}\mathbf{W} + \epsilon, \tag{5}$$

where $\mathbf{a}$ is a row vector of voxel activity values, $\mathbf{r}$ is a row vector of neural responses obtained from the encoding model, $\mathbf{W}$ is a weight matrix were each column $\mathbf{w}_v$ represents the linear measurement model for a different voxel a_v, and ϵ is a random normal row vector with mean $\mathbf{0}$ and covariance matrix with σ in the diagonal and zeros elsewhere. The value of σ was varied in Simulation 1 and was fixed to 5 in Simulation 2 (see below).

Equation 5 indicates that the activity in each voxel is a linear combination of neural channel responses, plus some random measurement noise. As shown in Figure 10, the model for each voxel was composed of a finite number of encoding channels that independently contributed to the aggregate signal of the voxel according to a set of weights. The values of the weights were randomly and uniformly sampled from 0 to 1, and then normalized by column, so that weights in $\mathbf{w}_v$ would add up to one. This way, the weights can be interpreted as the relative contribution of each channel to a voxel's activity.

We simulated a total of 100 voxels. In each simulated trial, the encoding model was presented with a given stimulus and produced a random vector of neural responses $\mathbf{r}$ as explained in the previous section, which were then used as input to the measurement model to obtain a random vector of voxel activities $\mathbf{a}$.

Simulation 1: False Positive Invariance Can Result From Features Of The Measurement Model

Previous theoretical work shows that methods that rely on fMRI activity estimates in general, and the widely-used cross-classification test in particular, cannot provide evidence for strict neural invariance (Figure 5; Soto et al., 2018). However, as indicated above, many experimental neuroscientists may be more interested in studying a form of partial invariance rather than strict invariance, and they may be tempted to assume that cross-classification and similar tests allow one to provide evidence for such partial invariance. Here we propose that, even in those cases, it is impossible to determine whether positive evidence of invariance is due to real invariance of neural representations versus "false positive" invariance, produced by the transformation and mixing of neural responses from different populations that occurs at each voxel, as described by the measurement model.

The goal of the first simulation was to show that the cross-classification test can lead to a conclusion of invariance when neural representations do not satisfy any sensible definition of invariance, strict or partial. As shown in Figure 11, the model underlying the simulation was created so that the encoding of the target dimension (e.g., orientation) was completely different across levels of the irrelevant dimension (e.g., spatial position). That is, two separate encoding models were created to represents the levels of the irrelevant dimension. The first level model consisted of a homogeneous population encoding model. To make sure that there was no invariance across levels of the irrelevant dimension, the encoding model for the second level of the irrelevant dimension was composed of channels whose tuning parameters were completely randomized. For each channel, the position parameter s_c was randomly sampled from a uniform distribution covering all values in the dimension, r_c^{max} was similarly sampled from values between 5 and 20,

and ω_c from values between 5 and 25. The result was a completely randomized encoding model for the second level of the irrelevant dimension, which was extremely unlikely to share any properties with the encoding model for the first level of the irrelevant dimension (compare the top and bottom encoding models in Figure 12).

As shown in the middle part of Figure 11, the measurement weights of the first level model, $\mathbf{W}_1$, were randomly sampled to generate activity patterns in each voxel (as explained above). On the other hand, the measurement weights for the second level model, $\mathbf{W}_2$, were chosen so that the activity patterns generated by any stimulus presented to this second level model would be as similar as possible as those presented to the first level model. To do this, we presented the level 1 model with the preferred stimulus of each channel s_c 20 times, and each time sampled data from 100 voxels. We then presented the level 2 model with the same stimuli a single time, and recorded a vector of average responses from the encoding model using Equation 3 (i.e., neural channel responses without any noise). Finally, for each voxel, the vectors of weights in $\mathbf{W}_2$ were obtained via Lasso regression, where voxel-wise activity patterns produced by the first level model were used as outputs to be predicted from the average neural activities obtained from the second level encoding model. Using Lasso regression, as implemented in *sklearn*, allowed us to constrain the weights to be positive. The regularization parameter of the regression model was not optimized, but fixed to a value of 0.01.

This procedure should result in a model in which no sensible definition of invariance holds at the level of neural responses, but that nonetheless should show some level of invariance at the level of indirect voxel activity measures. As shown in Figure 12, each simulation started by creating such a model (step 1), and continued by sampling data from it (step 2). To get that data, we presented the model with four stimuli, with values of -45°, 0°, 45°, and 90°, and sampled voxel activity patterns from it. Each stimulus presentation was repeated 20 times. We sampled data this way both from the first and second level models constructed as indicated above. Data was sampled twice from the first level model, to obtain training and testing data sets, and only once from the second level model, to obtain a testing data set only. We then performed a cross-classification test on the resulting data (steps 3 and 4 in Figure 12), following the same procedures as with the experimental data explained above, with the exception that the *Nu* parameter of the SVM was fixed to the default value of 0.5. Each simulation was repeated 200 times. The

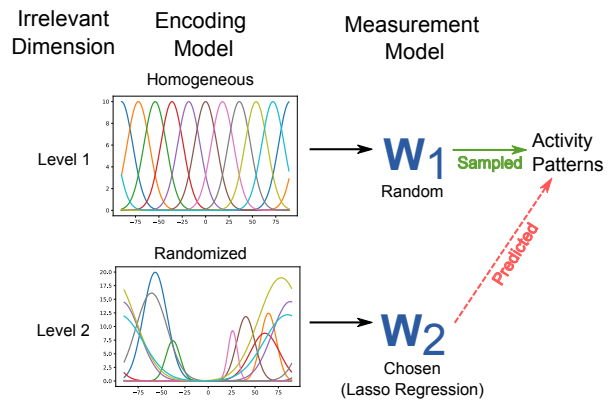

Fig. 11: Encoding models for simulation 1. Two separate encoding models were created to represent two levels of the irrelevant dimension. The first level model (top row) consisted of homogeneously tuned encoding channels. The measurement weights for this level were selected randomly. The second level model (bottom row) consisted of randomly tuned encoding channels, ensuring a complete lack of invariance between the first and second level models. The measurement weights for this level were then chosen via lasso regression to produce similar activity patterns to that of the first level model.

results presented below represent average statistics across all simulations, obtained from the testing data sets.

Finally, we repeated the group of simulations a total of 20 times, each time with a different value for the level of voxel measurement noise σ, going from 1 to 20.

Simulation 2: False Positive Invariance Can Result From Homogeneous Spatial Distribution Of Neural Populations

In the first simulation, the measurement model was chosen with the explicit goal of encouraging false positive invariance. Finding such false positive invariance would demonstrate, *at the very least*, that the cross-classification test is prone to a higher rate of false positives than the nominal 5% of standard statistical tests. However, a potential criticism to the simulation is that measurement models aligning to produce false positive invariance is an unlikely event, so the inflation of the false positive rates would exist, but would not be substantial. From this perspective, a simulation showing that false positive invariance results from a measurement model likely to be implemented in the brain would be more convincing. To understand how this might happen, it is useful to make a distinction between coding schemes that are spatially

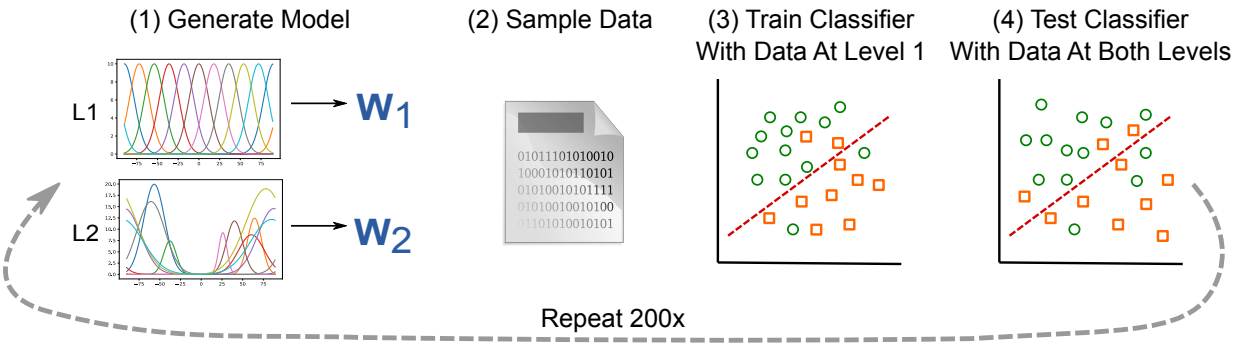

Fig. 12: Steps taken in each repetition of simulation 1. Data was sampled from both the first and second level models. A linear SVM classifier was fit to training data from the level 1 model. Then, the fitted model was tested with independent testing data from both the first and second level models.

smooth versus those that are fine-grained (Guest and Love, 2017).

Take the example shown in Figure 13. Each column in the figure represents a different voxel, which itself contains a large number of neurons (or neural populations), represented by small circles, with selectivity for some target stimulus property. In this simplified example, the neurons can show preference for one of two values of the target property, represented by the colors red and yellow. Different voxels have different proportions of the two types of neurons, so that despite of the spatial pooling of activity produced at each voxel, there is a distinctive pattern of activity produced across voxels by each stimulus property. This is a spatially smooth coding scheme.

On the other hand, note how within a voxel widely different spatial distributions of activity may produce the same value of global activity at the voxel level. For example, the same activity is obtained for voxel 1 in the case shown at the top (labeled Irrelevant Level 1) and in the case shown at the bottom (labeled Irrelevant Level 2), despite the fact that the fine-grained distribution of activities is widely different. The same is true for all other voxels. Thus, within each voxel one can see a fine-grained coding scheme that distinguishes levels of the irrelevant dimension.

More importantly for the issue of neural invariance, note how in the example shown in Figure 13 the neurons encoding the target dimension at level 1 of the irrelevant dimension (top) are completely different to those encoding the target dimension at level 2 of the irrelevant dimension (bottom). However, the spatial distribution of neurons specific to each value of the irrelevant dimension is spatially homogeneous, with about the same number of neurons of each kind at the top and bottom voxels.

The result of a spatially smooth encoding of the target dimension across voxels, together with a fine-grained spatial distribution of neurons specific to each value of the irrelevant dimension, produce as a result a case in which there is absolutely no invariance in the representation of the target dimension at the neural level, but invariance is perfect at the level of voxel activities.

Are there any examples of this type of coding in the brain? Indeed, a good example is the case under study in our main experiment: encoding of spatial position and orientation in V1. Encoding of spatial position is spatially smooth in V1, with the scale of retinotopic maps being similar to the voxel sizes typically used in neuroimaging, whereas encoding of orientation is much more spatially fine-grained (see Issa et al., 2008; Ng et al., 2007). Indeed, orientation maps are so fine-grained compared to the spatial scale of voxels in typical experiments that researchers have debated for years how it is that we are able to decode orientation from V1 using fMRI in the first place (e.g., Alink et al., 2013; Pratte et al., 2016; Maloney, 2014).

Thus, it is clear that the kind of encoding scheme exemplified by Figure 13 is plausible and should be cause of worry for researchers using decoding tests to offer evidence for invariance. The present simulation had the goal of exploring this issue in a systematic manner, and to determine its limits. Throughout the simulation, our main assumption is that a target dimension is encoded in a completely non-invariant manner, with one subpopulation of neurons responding whenever the irrelevant dimension is at level 1, and a different subpopulation of neurons responding whenever the irrelevant dimension is at level 2. Both subpopulations were modeled using a standard homogeneous encoding model (see above), but note that this similarity in tuning functions is not equivalent

to invariance, as each channel responded *only* at one of the levels of the irrelevant dimension.

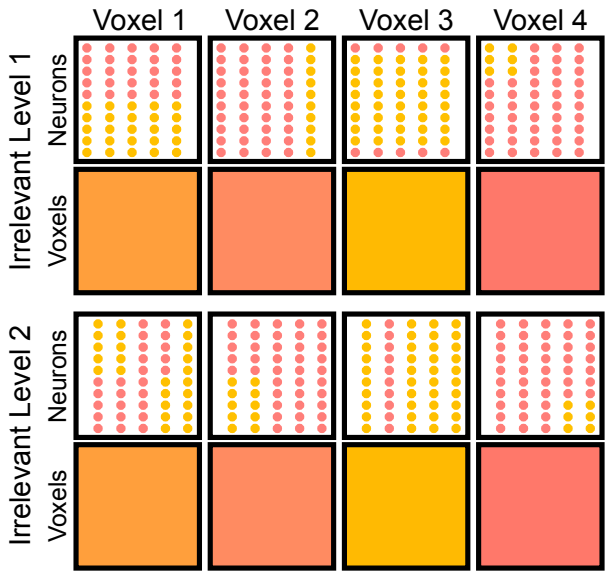

Fig. 13: Example highlighting the differences between spatially smooth versus fine-grained encoding schemes, and a particular combination of the two schemes that produces false-positive invariance in a voxelwise analysis. Each column in the figure represents a different voxel containing a large number of neurons represented by small circles, each with selectivity for one of two values of the target property, represented by the colors red and yellow. Different proportions of the two types of neurons are present across different voxels, so that despite of the spatial pooling of activity produced at each voxel, there is a distinctive pattern of activity produced across voxels by each stimulus property. This is a spatially smooth coding scheme. On the other hand, within a voxel widely different spatial distributions of activity may produce the same value of global activity at the voxel level, in a fine-grained coding scheme. Note how the multivoxel pattern of activity is the same for both levels of the irrelevant dimension, even though completely different populations of neurons encode each level. The combination of spatially smooth encoding of the target dimension and fine-grained encoding of the irrelevant dimension would produce false-positive invariance in a voxelwise analysis.

Note that in the example shown in Figure 13 the proportion of neurons encoding each feature is constant across levels of the irrelevant dimension. In that sense, one could say that at least this *macro* property of the code is invariant to changes in the irrelevant dimension. An important goal of this simulation was to determine to what extent disrupting this higher-level form of invariance would still lead to positive results in the cross-classification test. To do this, the measurement model of each voxel was obtained as shown in Figure 14. The weights $\mathbf{w}_v$ for level 1 of the irrelevant dimension were randomly generated, as explained above (step 1 in Figure 14). To create the measurement model for level 2 of the irrelevant dimension, we first obtained a vector $\mathbf{e}$ of random values sampled from a normal distribution with mean zero and standard deviation equal to $\sigma_{\mathbf{e}}$. This random vector was added to $\mathbf{w}_v$ (step 2), and then the values were made positive through rectification and normalized to add up to one (step 3), producing the final set of weights for level 2 of the irrelevant dimension.

Once the model was generated, the simulation was carried out following the same other steps as in Simulation 1, numbered 2 to 4 in Figure 12. The only difference was that $\sigma = 5$ in the measurement model, whereas the value of $\sigma_{\mathbf{e}}$ was varied from 0 to 0.5. Note that this implies that, at the highest values of $\sigma_{\mathbf{e}}$, the standard deviation of the changes in weights in the measurement model was 500% the average value of those weights (0.1). Thus, we made sure that even the higher-level *macro* invariance mentioned earlier was completely disrupted in the simulation. Any remaining false positive evidence of invariance would likely be the result of the homogeneous distribution of channels specific to each value of the irrelevant dimension.

Results

Experimental Results

The previous literature relies heavily on operational definitions of invariance, such as the cross-classification test (e.g., Allefeld and Haynes, 2014; Anzellotti and Caramazza, 2014; Kaplan et al., 2015). In contrast, our recent theoretical work formally defines invariance within the framework of population encoding models (Soto et al., 2018), and links such definitions with operational tests. We provide two clear definitions of invariance in neural encoding that could be the target of neuroimaging experiments (see subsection *Theoretical Definitions of Invariance at the Neural Level*). Strict invariance refers to the case where encoding of a dimension remains exactly the same across any changes in the irrelevant dimension (see Figure 5). Prior theoretical results show that fMRI decoding studies cannot provide evidence for the existence of strict invariance. Only failures of strict invariance can be reliably detected from decoding-based methods that rely on fMRI data (8).

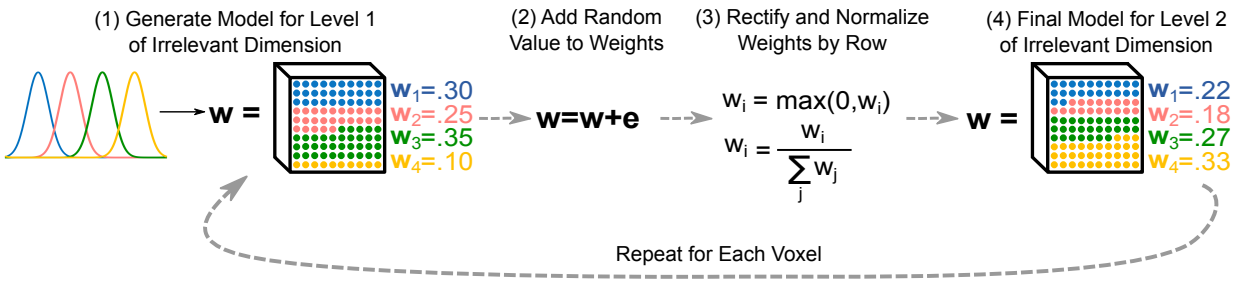

Fig. 14: Encoding model for simulation 2. The encoding model for the first level was generated in the same manner as in simulation 1 (see Figure 11). For the second level model, the channels weights within each voxel were perturbed by a vector of random values sampled from a normal distribution, and were then rectified and normalized to make sure that they were positive and added up to one for each voxel.

Many experimental neuroscientists may be dissatisfied by the definition of strict invariance, as they are more interested in studying a form of partial invariance, or "tolerance", across changes in an irrelevant dimension. We define partial invariance as the case in which some properties of the tuning of neural populations are preserved across changes in an irrelevant dimension. (Figure 6). Such a definition of invariance might seem more relevant, particularly to experimentalists who rely on voxel-wise fMRI data. Yet, even in the case of partial invariance, theoretical considerations suggest that fMRI decoding studies cannot provide reliable evidence for the existence of invariant representations. Instead, any meaningful test of invariance can only provide evidence against this property of neural encoding.

To validate these theoretical insights, we applied cross-classification and other decoding tests to the study of orientation and spatial position in primary visual cortex, as previous research has established that these properties are not encoded in an invariant way. Participants were presented with the stimuli in Figure 2 while they performed a task involving a stimulus presented at the center of the screen. Functional MRI data was acquired at the same time, with separate runs providing data for training and testing of an SVM classifier. Training runs were composed of stimuli presented only in spatial positions top-right and bottom-left (highlighted through red and blue boxes in Figure 2). Testing runs included all sixteen stimulus combinations. We trained a linear SVM classifier to decode a target dimension (e.g., orientation) while holding the irrelevant dimension (e.g., spatial position) constant. We then tested the classifier with data obtained at the trained value of the irrelevant dimension (e.g., top-right spatial position) as well as new values of the irrelevant dimension (e.g., bottom-right, bottom-left, and top-left spatial positions). The classifier provided decision variables and accuracy estimates used to perform cross-classification and other tests of invariance presented below (for more details, see *Materials and Methods*).

The Cross-Classification Test Produces False Positives. The popular cross-classification test of invariance (e.g., Allefeld and Haynes, 2014; Anzellotti and Caramazza, 2014; Kaplan et al., 2015) was specifically designed to provide evidence in favor of invariant encoding in a given brain region. Theoretical considerations suggest that such a test would be biased to generate false positives, as any method relying on indirect measures of neural activity may produce an observation of invariance that is due to properties of the measurement model (i.e., the model describing how neural activity is transformed to indirect measures, see Figure 10 and section *Measurement Model*) rather than properties of neural encoding. We performed a set of analyses using the cross-classification test to validate our theoretical prediction that this method should produce findings of false-positive invariance; that is, cases in which invariance is concluded even though our knowledge of the underlying neural code indicates that such invariance does not exist. The cross-classification test was conducted by assessing whether a linear decoder trained to classifying the target dimension at one level of the irrelevant dimension, could perform the same classification significantly above chance level across non-trained levels of the irrelevant dimension. A positive result in the cross-classification test is usually taken as evidence for the existence of invariant representations in the area of interest (Anzellotti and Caramazza, 2014; Kaplan et al., 2015).

We conducted two separate analyses using the cross-classification test in which we switched the identities of the relevant and irrelevant dimensions. In

the first analysis, spatial position was treated as the target dimension to be decoded, while orientation remained as the irrelevant dimension. To obtain decoded stimulus values for spatial position, we used deconvolved single-trial estimates of activity in V1 voxels as input to our SVM linear decoder. We trained the decoder to classify trials based on spatial position labels (top-right vs bottom-left, see boxed stimuli in Figure 2) and holding constant the level of grating orientation (irrelevant dimension; for example, 0°) using leave-one-run-out cross-validation, and tested it with independent data sets across all levels of grating orientation (0°, 45°, 90°, and 135°). To test for cross-classification invariance, we performed a binomial test on the accuracy estimates from the testing data set, corrected for multiple comparisons using the Holm-Sidak method. (for more details, see *fMRI Decoding Tests*) If the accuracy score was significantly above chance, then the cross-classification test concludes that spatial position is encoded invariantly from orientation in V1, a conclusion known to be false.

For each participant, we repeated the analysis four times, once for each level of grating orientation that was held fixed in the classifier's training data. Based on theoretical considerations (see *fMRI Decoding Tests* and *Simulation 2: False Positive Invariance Can Result From Homogeneous Spatial Distribution Of Neural Populations*; Soto et al., 2018), we predicted that the cross-classification test would generate consistent false positives in the case where spatial position was used as the relevant dimension to be decoded. Since spatial position is encoded in a spatially smooth manner in V1, we expected strong performance of the classifier across all levels of orientation. In other words, we expected the accuracy scores of the classifier to remain above chance across different levels of the irrelevant dimension.

Figure 15 shows accuracy estimates from such a decoding procedure for all five subjects. The SVM linear decoder achieves extremely high levels of classification accuracy in test sets across all 5 subjects. As predicted, the test incorrectly finds evidence for invariance of spatial position from orientation in all participants and all tests (all $p<.001$). This result is unsurprising, in the sense that one would intuitively expect it given the properties of encoding in V1, where information about the spatial position of stimuli is spatially smooth, distributed at around the same scale as our voxel size, but information about orientation is fine-grained, distributed at a smaller scale than our voxel size. The point to be made from these results, however, is that in most applications of the cross-classification test researchers do not know much about the encoding in the area under study, and they could easily conclude in favor of invariance when the underlying code does not show such property.

We performed a second analysis in which orientation was treated as the target dimension to be decoded, while spatial position was treated as the irrelevant dimension. We trained the decoder to classify trials based on grating orientation (0°, 45°, 90°, and 135°, see boxed stimuli in Figure 2) and holding constant the position of the spatial window (irrelevant dimension; for example, top-right or 20° in Figure 2) using leave-one-run-out cross-validation, and tested it with independent data sets across all levels of spatial position (top-right, bottom-right, bottom-left, and top-left; or 20°, 80°, 200°, and 260° in Figure 2). All other procedures remained the same as in the first analysis. Figure (16) shows decoding accuracy results for the orientation analysis. The SVM classifier was able to successfully decode orientation information at the original training position in all subjects, but for subjects 1 and 4 this was restricted to a single training window (200° window). In contrast to spatial position classification, the classifier's accuracy scores drop significantly in untrained testing windows. In general, the classifier's accuracy drops in a monotonic fashion as the testing spatial position moves farther away from the training position. We applied the cross classification test to these set of results, while expecting false positives in cases where the classifier accuracy was above chance at the training window (which is considered to provide a "ceiling" of performance for the cross-classification accuracy; see Anzellotti and Caramazza, 2014; Kaplan et al., 2015). Sure enough, the cross-classification test generated a false positive in two separate instances (Figure 16). In the first instance, the cross-classification test was sensitive to an above-chance accuracy score for a spatial position that was presented on the same visual hemifield as the training position (Figure 16, sub#3, left panel; $p<0.01$). The second instance highlights an even more unexpected scenario. The cross-classification test was sensitive to a higher than chance accuracy score in a spatial position that was presented in the opposite hemifield to the training position (Figure 16, sub#2, right panel; $p<0.05$). As shown in our simulations results below (see *Simulation 2: False Positive Invariance Can Result From Homogeneous Spatial Distribution Of Neural Populations*), such false positive invariance can be expected as resulting from properties of the measurement models linking neural activity to voxelwise activity. By random chance, such measurement models will sometimes produce false positive invariance.

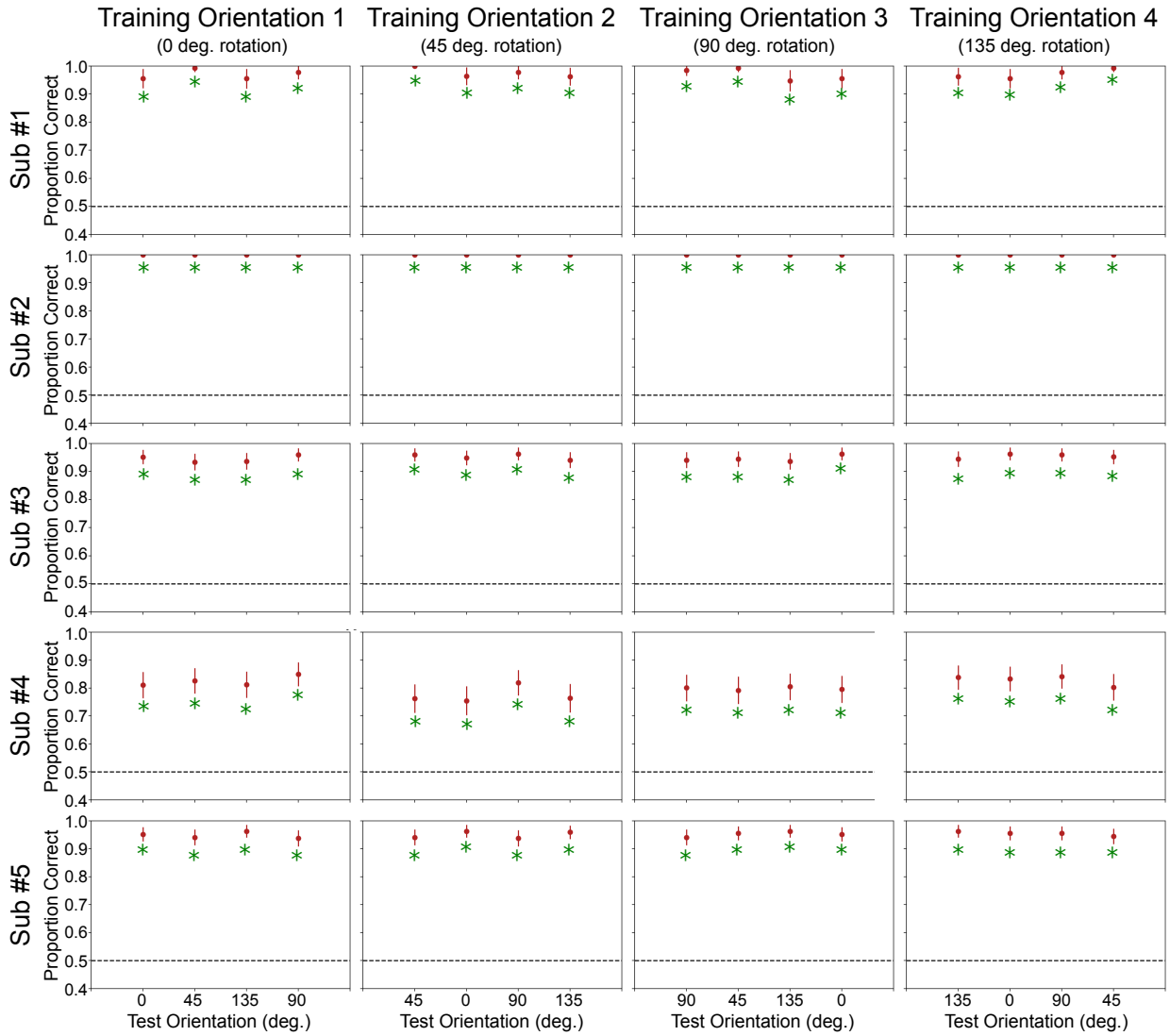

Fig. 15: Classification accuracy results with test data for the decoding of spatial position. Each row represents a complete analysis for a single subject. Columns represent different levels of the irrelevant dimension (orientation) held fixed during training. For example, the leftmost cell for subject 1 displays accuracy results from a decoder that was trained to discriminate spatial positions while the oriented grating was held constant at 0°. The same decoder was then tested with independent testing data in which the oriented grating was held at all orientations (0°, 45°, 90°, and 135°). A green asterisk represents accuracy values significantly above chance (dotted line), leading to the conclusion of invariance from the cross-classification test. As expected, the cross classification test consistently generates false positives and yields a conclusion that spatial position is encoded invariantly from orientation in V1.

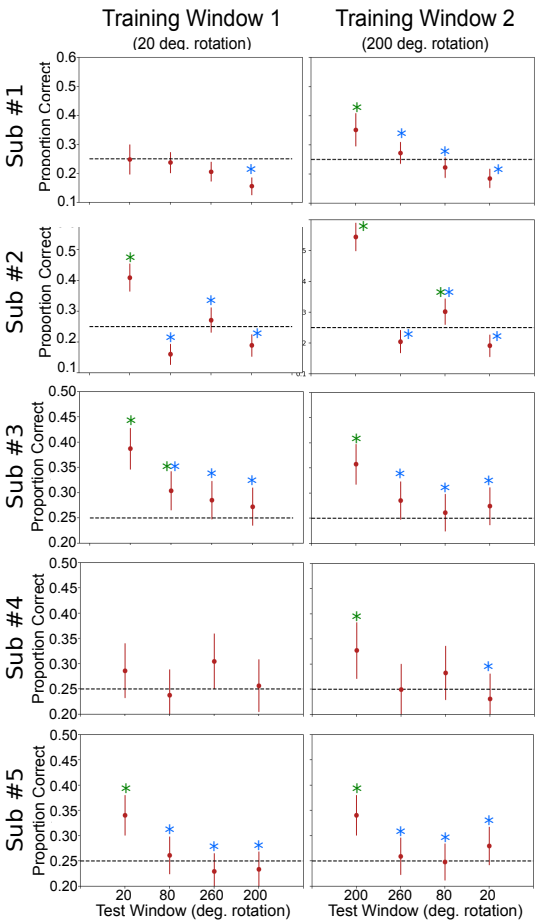

Fig. 16: Classification accuracy results with test data for the decoding of orientation. Each row represents a complete analysis for a single subject. Columns represent different levels of the irrelevant dimension (spatial position) held fixed during training. For example, the leftmost cell for subject 1 displays accuracy results from a decoder that was trained to discriminate grating orientations, while holding the spatial position of the window constant at 20° of rotation. The same decoder was then tested with independent testing data in which the window was held at all spatial positions (20°, 80°, 260°, and 200° of rotation). A green asterisk represents accuracy values significantly above chance (dotted line), leading to the conclusion of invariance from the cross-classification test. As predicted, the cross classification test is susceptible to generating false positives, concluding that orientation is encoded invariantly from spatial position in the V1 of subjects 2 and 3. A blue asterisk represents that accuracy values significantly drop from the value observed at the trained window, leading to the conclusion of no invariance from the classification accuracy invariance test. The test is successful in finding evidence against invariance in the data of each participant.

Theoretically-Driven Tests Can Detect Failures of Invariance In our previous work, we proposed theoretically-driven tests of invariance that were designed to provide more valid conclusions than the cross-classification test (Soto et al., 2018). In short, our work shows that positive evidence for the presence of strict invariance from any decoding-based method is inconclusive. Instead, only failures of invariance can be reliably inferred (see Figure 8). Here, we apply two of these tests on our data set: the classification accuracy invariance test and the decoding invariance test. In contrast to the cross-classification test, both of these theoretically-driven tests try to detect failures of invariance as opposed to confirming the presence of invariance.

The classification accuracy invariance test defines invariance as the case where the probability of correct classification is exactly the same across all levels of the irrelevant dimension. With invariance being the null hypothesis, the test is sensitive to any drop in the classifier's performance across different levels of the irrelevant dimension. We implemented the classification accuracy invariance test by applying an omnibus Chi-Square test on the accuracy estimates from the linear decoder (i.e., testing whether all proportions are the same or some of them are different). Then, we performed pairwise comparisons between accuracy at the training level and each non-training level of the irrelevant dimension. As before, we first applied the classification accuracy test to decoding results from the spatial position classification. Results from the classification accuracy invariance test for spatial position classification are shown in Figure 15. The test did not find support for a lack of invariance in any of the subjects. It is important to note, however, that this null result does not lead to the conclusion that invariance exists. That is, the results shown in Figure 15 are simply inconclusive about invariance.

Next, we applied the classification accuracy invariance test to decoding results from the orientation classification. We expected the classification accuracy invariance test to be more sensitive to failures of invariance in this case, because the decoding accuracy shows a monotonic decline as the testing spatial position moves away from the training position. Indeed, the classification accuracy invariance test shows extreme sensitivity to failures of invariance whenever there is a drop in classifier performance across different levels of the irrelevant dimension (Figure 16). Failures of invariance were detected in every case where the classifier successfully decoded orientations above chance levels in the training window. Interestingly, failures of invariance were also detected in cases where the classifier did not successfully decode

orientation above chance. This is unintuitive, but expected from a theoretical point of view (see Soto et al., 2018), which suggests that a decoder does not have to perform well or be optimal in any way to be able to detect failures of invariance. More specifically, an overall failure of invariance was detected for subject 1 in training windows 1 ($\chi^2(3)$= 14.495, p<.01) and 2 ($\chi^2(3)$= 30.832, p<.001), with the following pairwise comparisons resulting significant: in window 1 20° vs. 200° (z=3.189, p<.01); in window 2: 200° (z=2.334, p<.05), 200° vs 80° (z=3.938, p<.001), 200° vs 20° (z=5.230, p<.001). An overall failure of invariance was detected for subject 2 the training windows 1 ($\chi^2(3)$= 87.893, p<.001) and 2 ($\chi^2(3)$= 168.767, p<.001), with the following pairwise comparisons resulting significant: in window 1, 20° vs 80° (z=8.275, p<.001), 20° vs 260° (z=4.363, p<.001), 20° vs 200° (z=7.209, p<.001); in window 2, 200° vs 260° (z=10.538, p<.001), 200° vs 80° (z=7.354, p<.001), 200° vs 20° (z=10.991, p<.001). An overall failure of invariance was detected for subject 3 the training window 1 ($\chi^2(3)$= 20.126, p<.001) and 2 ($\chi^2(3)$= 14.494, p=.01), with the following pairwise comparisons resulting significant: in window 1, 20° vs 80° (z=2.880, p<.01), 20° vs 260° (z=3.543, p<.001), 20° vs 200° (z=4.013, p<.001); in window 2, 200° vs 260° (z=2.541, p<.05), 200° vs 80° (z=3.424, p<.01), 200° vs 20° (z=2.946, p<.01). An overall failure of invariance was not detected for subject 4 in training windows 1 ($\chi^2(3)$= 3.647, p>.05) and 2 ($\chi^2(3)$= 168.767, p=.062), but one pairwise comparison was found to be significant in window 2, 200° vs 20° (z=2.946, p<.01). An overall failure of invariance was detected for subject 5 in the training window 1 ($\chi^2(3)$= 22.114, p<.001) and 2 ($\chi^2(3)$= 13.654, p<.01), with the following pairwise comparisons resulting significant: in window 1, ; 20° vs 80° (z=2.852, p<.01), 20° vs 260° (z=4.047, p<.001), 20° vs 200° (z=3.901, p<.001), 200° vs 260° (z=2.922, p<.01), 200° vs 80° (z=3.338, p<.01), 200° vs 20° (z=2.171, p<.05).

The decoding invariance test is a second theoretically-driven test that was designed to identify failures of invariance (Soto et al., 2018). Unlike the previous two tests, the decoding invariance test does not make use of classification accuracy estimates. Instead, it directly relies on certain properties of the decoding probability distributions for individual stimuli. That is, linear classifiers like the one used here perform classification of a new data point by computing a decision variable z, representing the distance of the data point from the classifier's hyperplane separating two classes. When the decision variable is larger than some criterion value (usually zero), the output is one class, whereas when the decision variable is smaller than the criterion the output is the other class. Instead of comparing simple accuracy estimates, the decoding invariance test compares the full distributions of such decision variables, or *decoding distributions*.

The analysis followed the same steps and rationale as the classification accuracy invariance test presented earlier, but instead of computing accuracies and testing their differences, we obtained decision variables from the trained classifier, and used those to estimate decoding distributions using kernel density estimation. For each pair of stimuli differing in the irrelevant dimension (e.g., 0° vs 45° grating orientation, when the decoded variable was spatial position) we computed the distance between decoding distributions using a discretized $L1$ metric, which corresponds to the area highlighted in yellow in Figure 9. Then, we summed a number of such $L1$ metrics across values of the decoded dimension (e.g., the two spatial windows, when the decoded variable was spatial position), which produced an $L1_j^G$ statistic (see Equation 2). Simply put, while a single $L1$ metric is analogous to the accuracy of the classifier for a single decoded label, the $L1_j^G$ statistic is analogous to the overall decoding accuracy across all labels. The only difference is that $L1_j^G$ measures distances between decoding distributions, rather than accuracies. A positive result on the decoding invariance test suggests a failure of strict neural invariance for the given comparison. Also, we must note that, in theory, the decoding invariance test should provide more information about such violations than the decoding accuracy invariance test (see Figure 8).

We began by applying the decoding invariance test to decoding results from the spatial position classification. Figure 17 shows the results of the decoding invariance test on spatial position classification for all five subjects, repeated for four different training orientations. We expected the test to be most sensitive to failures of invariance in comparison to other operational tests. In line with our theoretical prediction, the test found failures of invariance in many cases where accuracy-based tests either found false positives (i.e., cross-classification) or failed to detect failures of invariance (i.e., classification accuracy invariance; see Figure 15). Overall, we found that the decoding invariance test detected failures of invariance in the data of all five participants (17 out of 20 analyses). More specifically, the decoding invariance test detected failures of invariance for subject 1 in the following cases: in training orientation 1, 0° vs 45° ($L1_j^G$=43.248, p<.05), 0° vs 135° ($L1_j^G$=43.413, p<.05), 0° vs 90°

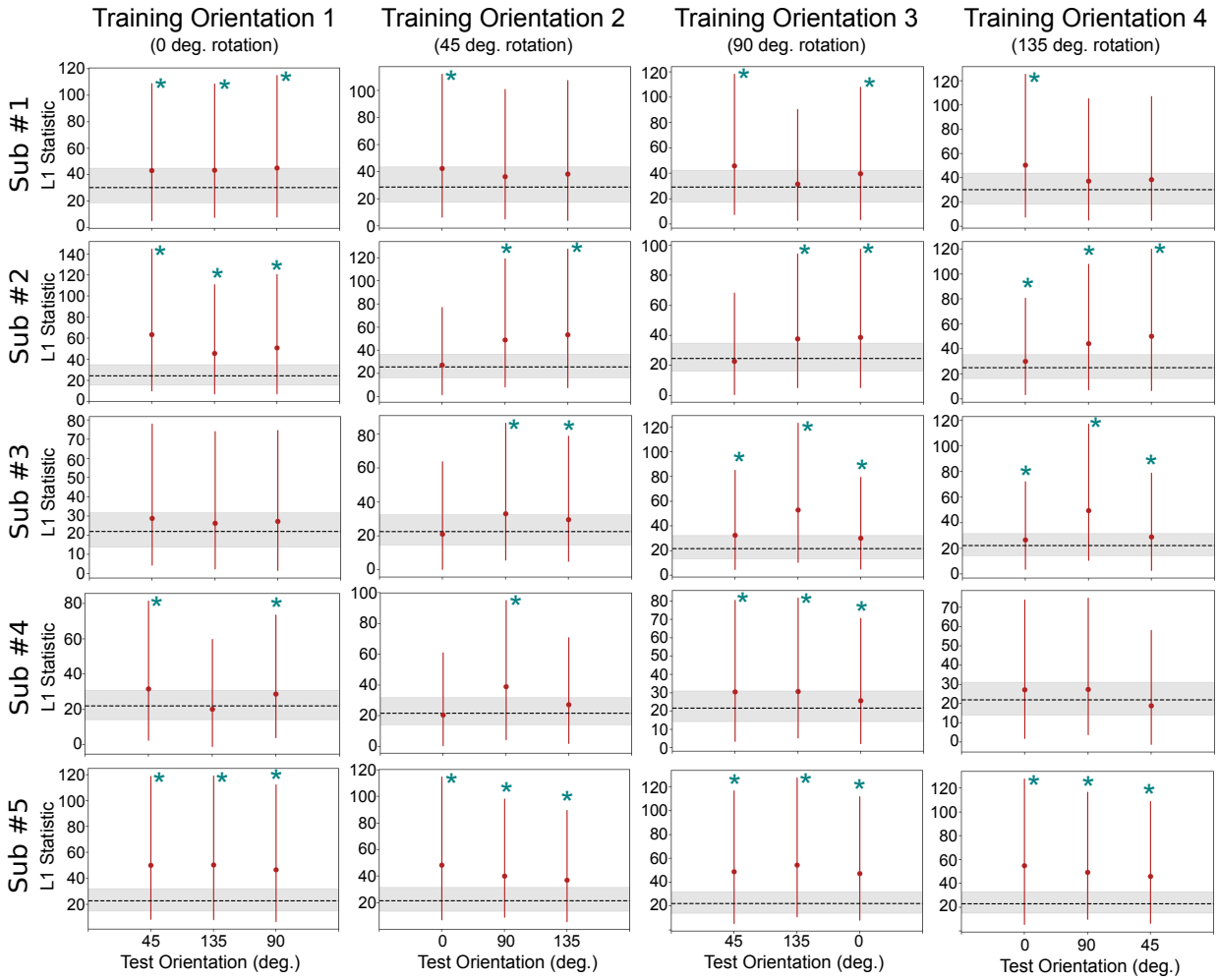

Fig. 17: Decoding invariance test results with spatial position as the target dimension. Each row represents a complete analysis for a single subject. Columns represent different levels of the irrelevant dimension (orientation) during training. The y-axis shows the $L1_j^G$ statistic, which quantifies the magnitude of violations of decoding invariance. Bars represent 90% bootstrap confidence intervals on the $L1_j^G$ statistic. The dotted line and surrounded gray area represent the expected value and 90% bootstrap confidence interval for the $L1_j^G$ statistic when no differences exist between two distributions. Teal asterisks represent significant failures of invariance according to a permutation test. The test successfully found evidence for the absence of invariance in 18 out of 20 analyses.

($L1_j^G$=43.413, p<.05); in training orientation 2, 45° vs 0° ($L1_j^G$=42.402, p<.05); in training orientation 3, 90° vs 45° ($L1_j^G$=45.957, p<.05), 90° vs 0° ($L1_j^G$=39.771, p<.05), in training orientation 3, 135° vs 0° ($L1_j^G$=50.459, p<.01). The test detected failures of invariance for subject 2 in the following cases: in training orientation 1, 0° vs 45° ($L1_j^G$= 63.331, p<.001), 0° vs 135° ($L1_j^G$=45.551, p<.001), 0° vs 90° ($L1_j^G$= 50.828, p<.001); in training orientation 2, 45° vs 90° ($L1_j^G$=49.146, p<.001), 45° vs 135° ($L1_j^G$= 53.428, p<.001); in training orientation 3, 90° vs 135° ($L1_j^G$=37.829, p<.01), 90° vs 0° ($L1_j^G$= 38.863, p<.01); in training orientation 4, 135° vs 0° ($L1_j^G$= 29.941, p<.05), 135° vs 90° ($L1_j^G$= 44.375, p<.001), 135° vs 45° ($L1_j^G$=50.289, p<.001). The test detected failures of invariance for subject 3 in the following cases: in training orientation 2, 45° vs 90° ($L1_j^G$= 33.199, p<.05), 45° vs 135° ($L1_j^G$=29.709, p<.05); in training orientation 3, 90° vs 45° ($L1_j^G$=32.414, p<.05), 90° vs 135° ($L1_j^G$=52.844, p<.001), 90° vs 0° ($L1_j^G$=30.163, p<.05); in training orientation 4, 135° vs 0° ($L1_j^G$= 26.472, p<.05), 135° vs 90° ($L1_j^G$=49.387, p<.001), 135° vs 45° ($L1_j^G$= 29.003, p<0.05). The test detected failures of invariance for subject 4 in the following cases: in training orientation 1, 0° vs 45° ($L1_j^G$= 31.555, p<.05), 0° vs 90° ($L1_j^G$=28.671, p<.05); in training orientation 2, 45° vs 90° ($L1_j^G$=39.067, p<.001); in training orientation 3, 90° vs 45° ($L1_j^G$= 30.418, p<.05), 90° vs 135° ($L1_j^G$=30.632, p<.05), 90° vs 0° ($L1_j^G$= 25.684, p<.05). Finally, the test detected failures of invariance for subject 5 in the following cases: in training orientation 1, 0° vs 45° ($L1_j^G$=50.238, p<.001), 0° vs 135° ($L1_j^G$=50.439, p<.001), 0° vs 90° ($L1_j^G$= 46.737, p<.001); in training orientation 2, 45° vs 0° ($L1_j^G$=48.398, p<.001), 45° vs 90° ($L1_j^G$=40.084, p<.001), 45° vs 135° ($L1_j^G$=37.133, p<.001); in training orientation 3, 90° vs 45° ($L1_j^G$=48.960, p<.001), 90° vs 135° ($L1_j^G$=54.352, p<.001), 90° vs 0° ($L1_j^G$=47.402, p<.001); in training orientation 4, 135° vs 0° ($L1_j^G$=54.942, p<.001), 135° vs 90° ($L1_j^G$=49.427, p<.001), 135° vs 45° ($L1_j^G$=45.854, p<.001).

From these results, it is apparent that the decoding invariance test is sensitive to failures of invariance known to exist in the underlying neural code, even when the decoder's overall accuracy during generalization testing reaches a ceiling (see Figure 15). Even when the classifier reached perfect accuracy across all levels of the irrelevant dimension, as was the case in subject 2's decoding results, the decoding invariance test successfully detected failures of invariance. This highlights the value of developing theory-based tests in neuroimaging analyses.

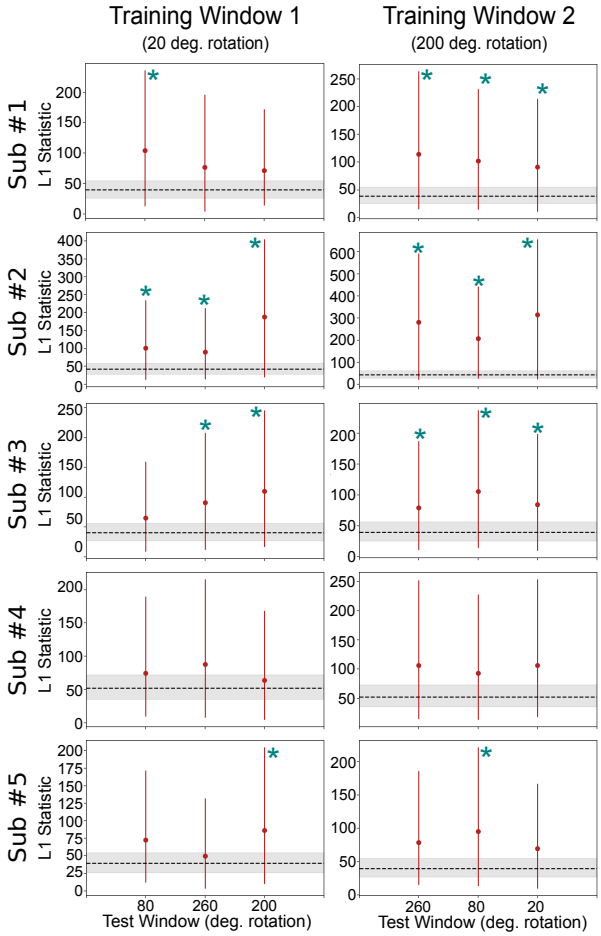

Fig. 18: Decoding invariance test results with orientation as the target dimension. Each row represents a complete analysis for a single subject. Columns represent different levels of the irrelevant dimension (spatial position) during training. The y-axis shows the $L1_j^G$ statistic, which quantifies the magnitude of violations of decoding invariance. Bars represent 90% bootstrap confidence intervals on the $L1_j^G$ statistic. The dotted line and surrounded gray area represent the expected value and 90% bootstrap confidence interval for the $L1_j^G$ statistic when no differences exist between two distributions. Teal asterisks represent significant failures of invariance according to a permutation test. The test successfully found evidence for the absence of invariance in 8 out of 10 analyses.

We also applied the decoding invariance test to decoding results from the classification of orientation. Figure (18) shows the results of the decoding invari-

ance test on orientation classification for all five subjects, repeated for two different training windows. As in the spatial position analysis, we expected the test to be more sensitive to failures of invariance than methods based on classification accuracy. Contrary to our expectations, in this case the decoding invariance test detected failures of invariance less frequently than the classification accuracy invariance test. The decoding invariance test detected failures of invariance in the data of four out of five participants (eight out of ten analyses) when orientation was treated as the target dimension. More specifically, the decoding invariance test detected failures of invariance for subject 1 in the following cases: in training window 1, 20° vs 80° ($L1_j^G$= 104.508, p<.001); in training window 2, 200° vs 260° ($L1_j^G$= 114.291, p<.01), 200° vs 80° ($L1_j^G$= 102.444, p<.01), 200° vs 20° ($L1_j^G$= 91.252, p<.01). The test detected failures of invariance for subject 2 in the following cases: in training window 1, 20° vs 80° ($L1_j^G$= 101.049, p<.01), 20° vs 260° ($L1_j^G$=90.324, p<.05), 20° vs 200° ($L1_j^G$= 187.638, p<.001); in training window 2, 200° vs 260° ($L1_j^G$= 281.977, p<.001), 200° vs 80° ($L1_j^G$= 208.153, p<.001), 200° vs 20° ($L1_j^G$= 314.683, p<.001). . The test detected failures of invariance for subject 3 in the following cases: in training window 1, 20° vs 260° ($L1_j^G$= 90.956, p<0.01), 20° vs 200° ($L1_j^G$=109.779, p<.001); in training window 2, 200° vs 260° ($L1_j^G$= 79.337, p<.05), 200° vs 80° ($L1_j^G$= 105.765, p<.001), 200° vs 20° ($L1_j^G$= 84.353, p<.05). The test did not detect any failures of invariance for subject 4 (all p>.05). Finally, the test detected failures of invariance for subject 5 in the following cases: in training window 1, 20° vs 200° ($L1_j^G$=86.353, p<.05); in training window 2, 200° vs 80° ($L1_j^G$=95.574, p<.01).

In comparison to the classification accuracy invariance test, the decoding invariance test appears to be more sensitive to detecting failures of invariance in cases where the decoder's performance reaches ceiling levels (Figure 17). However, when classification accuracy is well below ceiling levels, as in decoding of orientation, the test is less sensitive than classification accuracy invariance, probably due to a lower statistical power of the permutation test involved.

Simulation Results

Simulation 1: False Positive Invariance Can Result From Features Of The Measurement Model

Our empirical results in the preceding section clearly show that the cross-classification test consistently generates false positives. Using procedures common in the previous literature, the test provided evidence of encoding of spatial position that is invariant to grating orientation, and encoding of orientation that is invariant to spatial position, in V1. The cross-classification test generated false positives both in cases where we expected absence of strict invariance and in cases where partial invariance could potentially hold. Yet, some researchers might argue that they are interested in studying a form of partial invariance that was not tested in our experiment, and that this form of invariance can still be detected by the commonly-used cross-classification test. Here we propose that it is impossible to determine whether positive evidence of invariance is due to real invariance of neural representations versus "false positive" invariance, produced by the transformation and mixing of neural responses from different populations that occurs at each voxel, for any sensible definition of invariance at the neural level. To provide evidence for such a general claim, we resort to simulation work (for details on the models and procedures used in these simulations, see *Simulations* in the *Materials and Methods* section).

The first simulation was designed to show that the cross-classification test can consistently generate false positives even when no sensible definition of invariance (strict or partial) holds at the neural level. We began by defining two sets of encoding models, corresponding to two levels of the irrelevant dimension, where we ensured that no invariance holds of any kind by encoding the target dimension through neural channels with homogeneous features (i.e., evenly spaced position, same maximum activity, same width) at level 1 of the irrelevant dimension (see Figure 11, top), and through neural channels with completely randomized features at level 2 of the irrelevant dimension (see Figure 11, bottom). Then, we produced false positive invariance by optimizing the weights of the measurement model such that the voxel-wise activity values were similar across the two levels of the irrelevant dimension (Figure 11). Finally, we sampled data from both models and used them as input to a linear SVM classifier. As in the preceding empirical analyses, the decoder was trained on data from the first level model and tested on independent data from both the first and second level models (Figure 12). This entire procedure was repeated 200 times per simulation run, and we present the average results across simulations. We performed twenty simulation runs, where we gradually increased the measurement noise in each voxel (from standard deviation 1 to 20, in steps of 1). Figure 19a shows the decoding accuracy results from this simulation. The most important values are represented by the blue curves, which represent performance of the classifier

in the non-trained level of the irrelevant dimension. Whenever accuracy is above chance, represented by the dotted line, the cross-classification test leads to a conclusion of invariance in a situation where no invariance exists (i.e., false positive invariance). The cross-classification accuracy score was much higher than chance across all levels of noise, even as measurement noise was drastically increased. In fact, cross-classification accuracy seems to reach an asymptotic value, similar to that of accuracy with the original training level of the irrelevant dimension (i.e., orange line in the figure), around 30% correct with high values of measurement noise. Figure 19b shows the proportion of false positives for the cross-classification test, which consistently remained above the nominal $\alpha = .05$, represented by the dotted line, across all levels of noise.

These results suggest that a suitable selection of measurement model is sufficient for inducing false positives in the cross-classification test, even when the underlying encoding distributions themselves show no form of invariance. At the very least, the simulation shows that the false positive rate of the cross-classification test is much higher than the widely accepted value of 5%. The magnitude of this inflation of the false positive rate is unknown, and probably depends on properties of encoding of the stimulus dimensions under study in the particular regions of interest under study. As those properties are usually unknown in studies using cross-classification, the danger of false-positive invariance is pervasive in the literature using the test.

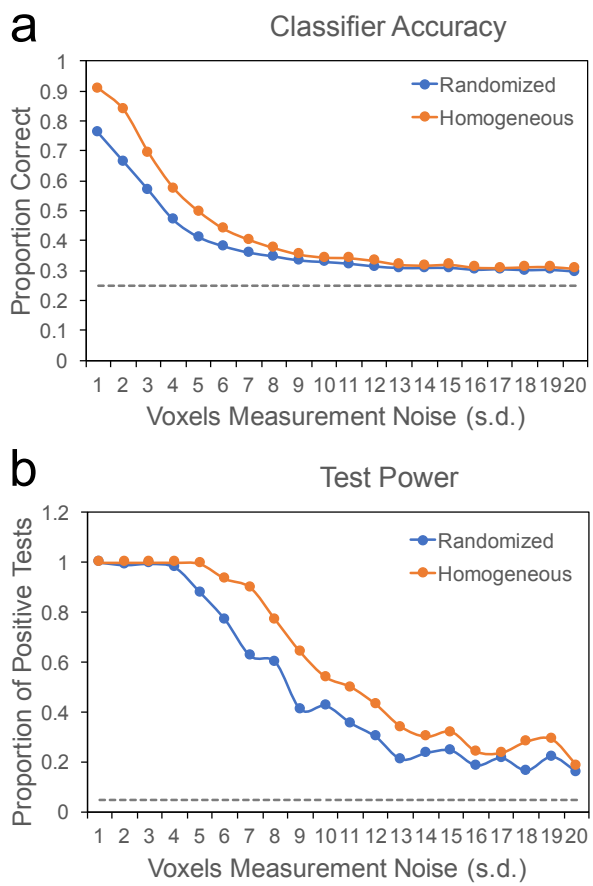

Fig. 19: Decoding results from simulation 1. (a) Classifier accuracy scores for model-generated data from both levels of the irrelevant dimension. The *y*-axis represents accuracy scores, while the *x*-axis represents level of measurement noise (in units of standard deviation). Note that classifier accuracy never dips below chance levels (dotted line) for either model. (b) Proportion of cross-classification tests that positively concluded in favor of invariance. The *y*-axis represents proportion of false positives, and *x*-axis represents measurement noise. At all levels of measurement noise, the proportion of false positives remains higher that the accepted 5% threshold (dotted line).

Simulation 2: False Positive Invariance Results From Encoding Schemes Known to Be Implemented in Cortex

It can be argued that the conditions that lead to false positive invariance in the first simulation, namely the explicit selection of the measurement weights that produce similar voxel-wise activity patterns across levels of the irrelevant dimension, are unlikely to occur in real fMRI experiments. As indicated above, the simulation was performed to provide evidence of the danger of false positive invariance, but we do not know how pervasive this issue is in real data.

However, a stronger case can be made if we show that an encoding scheme known to be used in the brain is likely to lead to false positive invariance using the cross-classification test. Such a case is illustrated by the experimental results shown in Figure 15, involving decoding of spatial position from V1 voxels, across variation in grating orientation. The

reason why the cross-classification test provides evidence for invariance in this case is that the code for spatial position in V1 is spatially smooth, whereas the code for grating orientation is fine-grained (Guest and Love, 2017). Figure 13 illustrates in a more general way why a combination of spatially smooth coding of the target dimension and fine-grained coding of the irrelevant dimension would lead to findings of false-positive invariance. Each column in the figure represents a different voxel, which contains a number of neurons represented by small circles. Each neuron is selective for one of two levels of the target dimension, represented by the colors red and yellow. Note first that there is no invariance at the neural level in this example. The population of neurons encoding the target dimension at the first level of the irrelevant dimension (top) is completely different from the population encoding the target dimension at the second level of the irrelevant dimension (bottom). However, the relative proportion of neuron types (i.e., red vs. yellow) remains the same across levels of the irrelevant dimension. This produces completely invariant voxelwise patterns of activity.

Of course, in a more realistic scenario, the proportion of neuron types would not be identical for every voxel across levels of the irrelevant dimension. The goal of our second simulation was to investigate to what extent false positive invariance is a problem when different but homogeneous populations of neurons encode a target dimension at different values of the irrelevant dimension, and to what extent the problem persists regardless of differences in the contribution of each population to voxelwise activity. With this goal in mind, we created a model in which a target dimension is encoded in a completely non-invariant manner, with one subpopulation of neurons responding whenever the irrelevant dimension is at level 1, and a different subpopulation of neurons responding whenever the irrelevant dimension is at level 2. Both subpopulations were modeled using a standard homogeneous encoding model (see above), but note that this similarity in tuning functions is not equivalent to invariance, as each channel responded *only* at one of the levels of the irrelevant dimension. In other words, our simulation assumes that populations encoding the target dimension are completely separated across levels of the irrelevant dimension, but they encode the target dimension in a similar way (just as neurons in Figure 13 have two selectivity types across levels of the irrelevant dimension). As before, we report the averaged results from 200 simulations in each run. Measurement noise was set to a fixed level across simulations (s.d.=5, which in our previous simulation produced accuracies around

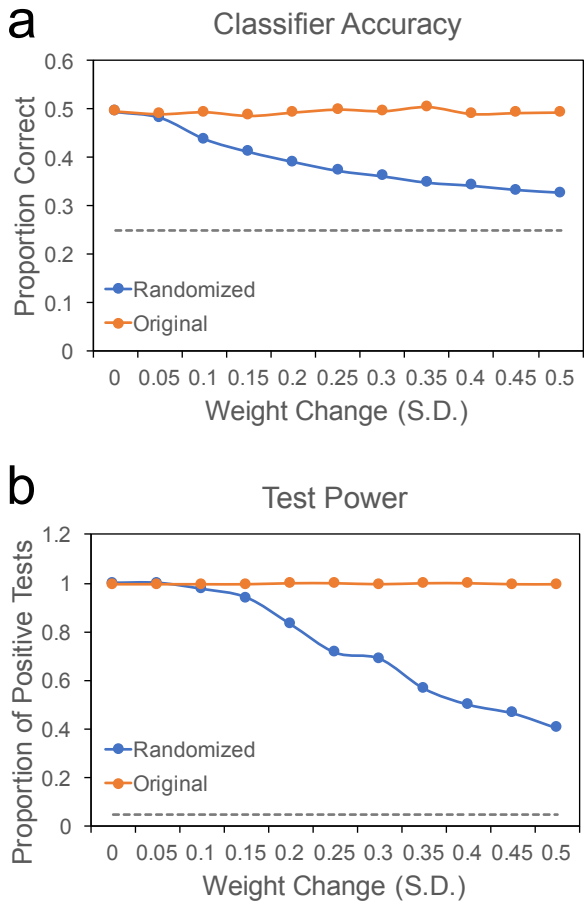

Fig. 20: Decoding results from simulation 2. (a) Classifier accuracy scores for model-generated data from both levels of the irrelevant dimension. The *y*-axis represents accuracy scores, while the *x*-axis represents magnitude of noise added to measurement weights for the second level model. Note that classifier accuracy never dips below chance levels (dotted line) for either model. (b) Proportion of cross-classification tests that positively concluded in favor of invariance. The *y*-axis represents proportion of false positives, and the *x*-axis represents magnitude of noise added to measurement weights for the second level model. The proportion of false positives remains well above the accepted 5% threshold (dotted line) at noise levels with a standard deviation five times the average value of the original weights.

40%-50%, see Figure 19). In each simulation run, we increased the difference in the measurement models for the two levels of the irrelevant dimension, by adding random noise to weights of the measurement model as illustrated in Figure 14. We started with a simulation in which the measurement model was identical across levels of the irrelevant dimension, which corresponds to the case illustrated in Figure 13. The standard deviation of the weight noise was then gradually increased in 0.05 steps, to a maximum of 0.5, which represents five times the average weight value. That is, in the final models the contribution of each neuron type (e.g., neurons selective to a value of 0 in the target dimension) was widely different across levels of the irrelevant dimension.

The results from this simulation are shown in Figure 20. As in the first simulation, the cross-classification accuracy scores consistently remain above chance levels as the perturbation of measurement weights increased (Figure 20a, blue curve). Yet again, the cross-classification test consistently showed false positives at a rate much higher than the nominal 5% (Figure 14b). High levels of false-positive invariance were present even when the weight noise standard deviation was five times as large as the average weight values. These results suggest that, when two *completely separate* neural populations use similar codes to represent a target dimension across levels of an irrelevant dimension, false positive invariance is likely to be found regardless of high random variation in the measurement model. The fact that the two populations are completely separate ensures that no sensible definition of neural invariance holds in the current simulation.

Discussion

Here, we have provided empirical and computational evidence supporting the theoretical insight (from Soto et al., 2018) that the cross-classification test, and neuroimaging decoding tests in general, cannot support a valid conclusion of invariance of neural representations, for any sensible definition of invariance at the neural level. We used decoding of orientation and spatial position in V1 to show that the cross-classification test leads to conclusions of invariance when such invariance is known to not hold. We performed two separate decoding analyses: one testing whether encoding of grating orientation is invariant to changes in spatial position, and another one testing whether encoding of spatial position is invariant to changes in grating orientation. We applied the popular cross-classification test (Allefeld and Haynes, 2014; Anzellotti and Caramazza, 2014; Kaplan et al., 2015) to the results of our decoding analyses. When applied to decoding of spatial position, the cross-classification test consistently found false positive evidence of invariance across orientation. In the analysis of orientation, the cross-classification test found false positive evidence of invariance across spatial position in two out of five participants.

In comparison, the results of two theory-driven tests of invariance (the classification accuracy invariance and decoding invariance tests; proposed by Soto et al., 2018) correctly found evidence against invariance in our analyses. In the decoding of spatial position, the classification accuracy invariance test was unable to find such evidence, probably because all accuracies were at ceiling levels, but the decoding invariance test did find evidence against invariance despite such ceiling effects. On the other hand, in the decoding of orientation, both tests were sensitive to violations of invariance, but in this case the classification accuracy invariance test proved to be more sensitive, probably due to its superior statistical power.

It is always possible for researchers to point out that they are interested in studying a form of neural invariance that was not tested in our experiment, and that this form of invariance can still be detected by the commonly-used cross-classification test. For this reason, we performed two simulations showing that, more generally, the cross-classification test produces evidence of invariance when no sensible definition of invariance holds at the neural level. The first simulation showed that one can always find a measurement model (i.e., a model linking neural activity to the indirect measures obtained in neuroimaging) that induces false positives in the cross-classification test, even when the underlying neural encoding shows no form of invariance. At the very least, this simulation shows that the false positive rate of the cross-classification test is much higher than the widely accepted value of 5%. It is unknown by how much the false positive rate of the cross-classification test is inflated, and the precise value probably depends on properties of stimulus encoding in the particular brain regions under study. As those properties are usually unknown in studies that use cross-classification, the danger of false-positive invariance is pervasive in the literature that has used the test up to this point. The second simulation showed that an encoding scheme known to be used in the brain (spatially smooth coding of the target dimension and fine-grained coding of the irrelevant dimension; see Guest and Love, 2017) is likely to lead to false positive invariance using the cross-classification test, regardless of random variation in the measurement model.

Based on our empirical and computational results,

we conclude that the cross-classification test, when applied to neuroimaging data, is not a valid way to provide evidence for invariance of neural representations. The test should be avoided in its current form, and previous research using the test should be re-evaluated in light of our results. The test produces false positives both in cases where errors would be intuitively expected (e.g., in decoding of spatial position, and at low levels of noise in simulation 2) as well as in situations in which they seem unexpected (e.g., in decoding of orientation, and at high levels of noise in simulations 1 and 2). It is critical to note that, in this study, we knew that orientation and spatial position are encoded in a non-invariant manner in the primary visual cortex. In most studies, however, the cross-classification test is applied to cases in which very little prior information is available about the underlying neural code. It is unknown to what extent the problems highlighted by our two simulations are present in other studies.

Alternatively, we found that theoretically-driven tests like the decoding invariance test and classification accuracy invariance test are valid ways to provide evidence for failures of invariance. Although the tests did not show perfect performance (i.e., they did not detect lack of invariance in every test), they are quite sensitive and, more importantly, they do not lead to invalid positive conclusions of invariance. While comparing the two tests, we found that each seems most sensitive to different information about failures of invariance. The classification accuracy invariance test seems statistically more powerful when classifier performance is below ceiling, as shown by the results of the analysis of orientation. However, as expected from theory, the decoding accuracy invariance test is sensitive to violations of invariance that cannot be captured by the classification accuracy invariance test. In particular, when decoding accuracy is at ceiling, only the decoding invariance test can detect violations of invariance by relying on the more fine-grained information available in the full decoding probability distributions, rather than on the coarse information available in accuracy estimates. When researchers do not have access to large data sets, we recommend the classification accuracy invariance test, as it is more statistically powerful. Precise kernel density estimates, which are essential for accurate results from the decoding invariance test, require large longitudinal datasets like the one used in this study (i.e., at least 100-150 data points per density estimate; see Fortmann-Roe et al., 2012). However, if researchers do have access to large data sets, then we recommend using both tests, as they have complementary strengths.

We must note that these tests are only useful if the goal of a study is to obtain evidence of failures of invariance, which corresponds to context-dependent or configural encoding. There are many cases in which such encoding might be of interest in sensory neuroscience, such as in the study of configural encoding of face parts (e.g., Liu et al., 2009) and other object parts (e.g., Ross et al., 2018), the neural basis of Gestalt processes (e.g., Kubilius et al., 2011), and configural pooling of information across the ventral stream (Wilson and Wilkinson, 2015), among other topics.

Some researchers may still be interested in providing evidence for invariance via fMRI studies. We strongly discourage the use of decoding tests for this purpose. Theoretical arguments (Soto et al., 2018), as well as empirical and computational evidence presented in this study all indicate that invariance of neural representations cannot be validly inferred from neuroimaging decoding studies. If a researcher must use decoding tests to study invariance, they should be aware that the rate of false positives is likely to be higher than the nominal 5% standard of statistical tests. In general, any positive evidence of invariance could result from how neural populations are mixed within voxels, and be unrelated to properties of the underlying code. Even if researchers find decoding tests indispensable, they should refrain from using the cross-classification test, at least by itself. Perhaps a good avenue for future research would be to develop new tests that use decoding or encoding modeling in conjunction with adaptation paradigms (Grill-Spector et al., 2006; Krekelberg et al., 2006). Adaptation has the appeal of allowing to reduce the contribution that specific neural populations make to the BOLD signal in a voxel. However, a lesson to be learned from cross-classification is that any new test proposed to study invariance should be validated and linked to meaningful definitions of invariance before being broadly applied.

More generally, the present study shows the dangers of over-reliance on operational definitions that have only face validity in the study of neural representation through indirect measures. Our study joins other recent reports in the literature (Liu et al., 2018; Gardner and Liu, 2019) in showing that the application of highly sophisticated computational tools can lead to the wrong conclusions when problems of identifiability (e.g., between neural and measurement factors) inherent to neuroimaging are not taken into account.

Our study has largely validated arguments made about the measurement of invariance using neuroimaging that were made purely based on theoret-

ical considerations (Soto et al., 2018). Although we show that decoding tests cannot be trusted as a singular source of positive evidence for invariance, our findings open new opportunities to create more solid, theoretically-driven tests that might provide such evidence, and they provide a validated methodology to test for "configural" stimulus encoding in the brain.

References

Akama H, Murphy B, Na L, Shimizu Y, Poesio M (2012) Decoding semantics across fMRI sessions with different stimulus modalities: a practical MVPA study. *Frontiers in Neuroinformatics* 6:24.

Alink A, Abdulrahman H, Henson RN (2018) Forward models demonstrate that repetition suppression is best modelled by local neural scaling. *Nature Communications* 9:3854.

Alink A, Krugliak A, Walther A, Kriegeskorte N (2013) fMRI orientation decoding in V1 does not require global maps or globally coherent orientation stimuli. *Frontiers in Psychology* 4:493.

Allefeld C, Haynes JD (2014) Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. *NeuroImage* 89:345–357.

Anzellotti S, Caramazza A (2014) The neural mechanisms for the recognition of face identity in humans. *Frontiers in Psychology* 5:672.

Anzellotti S, Caramazza A (2017) Multimodal representations of person identity individuated with fMRI. *Cortex* 89:85–97.

Anzellotti S, Fairhall SL, Caramazza A (2014) Decoding representations of face identity that are tolerant to rotation. *Cerebral Cortex* 24:1988–1995.

Archila-Melendez ME, Valente G, Correia JM, Rouhl RPW, Kranen-Mastenbroek V, Jansma BM (2018) Sensorimotor representation of speech perception: Cross-decoding of place of articulation features during selective attention to syllables in 7t fMRI. *eNeuro* 5:e0252–17.2018.

Benson NC, Butt OH, Datta R, Radoeva PD, Brainard DH, Aguirre GK (2012) The retinotopic organization of striate cortex is well predicted by surface topology. *Current Biology* 22:2081–2085.

Brouwer GJ, Heeger DJ (2009) Decoding and reconstructing color from responses in human visual cortex. *Journal of Neuroscience* 29:13992–14003.

Buchweitz A, Shinkareva SV, Mason RA, Mitchell TM, Just MA (2012) Identifying bilingual semantic neural representations across languages. *Brain and Language* 120:282–289.

Ester EF, Anderson DE, Serences JT, Awh E (2013) A neural measure of precision in visual working memory. *Journal of Cognitive Neuroscience* 25:754–761.

Etzel JA, Gazzola V, Keysers C (2008) Testing simulation theory with cross-modal multivariate classification of fMRI data. *PLOS ONE* 3:e3690.

Fischl B (2012) Freesurfer. *Neuroimage* 62:774–781.

Fortmann-Roe S, Starfield R, Getz WM (2012) Contingent kernel density estimation. *PLoS ONE* 7:e30549.

Garcia JO, Srinivasan R, Serences JT (2013) Near-real-time feature-selective modulations in human cortex. *Current Biology* 23:515–522.

Gardner JL, Liu T (2019) Inverted encoding models reconstruct an arbitrary model response, not the stimulus. *eNeuro* 6:e0363–18.2019.

Gorgolewski K, Burns CD, Madison C, Clark D, Halchenko YO, Waskom ML, Ghosh SS (2011) Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics* 5:13.

Grill-Spector K, Henson R, Martin A (2006) Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences* 10:14–23.

Guest O, Love BC (2017) What the success of brain imaging implies about the neural code. *Elife* 6:e21397.

Henriksson L, Nurminen L, Hyvärinen A, Vanni S (2008) Spatial frequency tuning in human retinotopic visual areas. *Journal of Vision* 8:5.

Hinds OP, Rajendran N, Polimeni JR, Augustinack JC, Wiggins G, Wald LL, Rosas HD, Potthast A, Schwartz EL, Fischl B (2008) Accurate prediction of v1 location from cortical folds in a surface coordinate system. *Neuroimage* 39:1585–1599.

Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology* 195:215–243.

Issa NP, Rosenberg A, Husson TR (2008) Models and measurements of functional maps in V1. *Journal of Neurophysiology* 99:2745–2754.

Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM (2012) Fsl. *Neuroimage* 62:782–790.

Kaiser D, Azzalini DC, Peelen MV (2016) Shape-independent object category responses revealed by MEG and fMRI decoding. *Journal of Neurophysiology* 115:2246–2250.

Kaplan JT, Man K, Greening SG (2015) Multivariate cross-classification: Applying machine learning techniques to characterize abstraction in neural representations. *Frontiers in Human Neuroscience* 9:151.

Knijnenburg TA, Wessels LFA, Reinders MJT, Shmulevich I (2009) Fewer permutations, more accurate p-values. *Bioinformatics* 25:i161–i168.

Krekelberg B, Boynton GM, van Wezel RJA (2006) Adaptation: from single cells to BOLD signals. *Trends in neurosciences* 29:250–256.

Kubilius J, Wagemans J, Op de Beeck HP (2011) Emergence of perceptual gestalts in the human visual cortex: The case of the configural-superiority effect. *Psychological Science* 22:1296–1303.

Liu J, Harris A, Kanwisher N (2009) Perception of face parts and face configurations: an fMRI study. *Journal of Cognitive Neuroscience* 22:203–211.

Liu T, Cable D, Gardner JL (2018) Inverted encoding models of human population response conflate noise and neural tuning width. *Journal of Neuroscience* 38:398–408.

Maloney RT (2014) The basis of orientation decoding in human primary visual cortex: fine- or coarse-scale biases? *Journal of Neurophysiology* 113:1–3.

Man K, Kaplan JT, Damasio A, Meyer K (2012) Sight and sound converge to form modality-invariant representations in temporoparietal cortex. *Journal of Neuroscience* 32:16629–16636.

Ng J, Bharath AA, Zhaoping L (2007) A survey of architecture and function of the primary visual cortex (V1). *EURASIP J. Appl. Signal Process.* 2007:124–124.

Pedregosa F, Eickenberg M, Ciuciu P, Thirion B, Gramfort A (2015) Data-driven hrf estimation for encoding and decoding models. *Neuroimage* 104:209–220.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al. (2011) Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12:2825–2830.

Peirce JW (2007) Psychopy: psychophysics software in python. *Journal of Neuroscience Methods* 162:8–13.

Pouget A, Dayan P, Zemel RS (2003) Inference and computation with population codes. *Annual Review of Neuroscience* 26:381–410.

Pratte MS, Sy JL, Swisher JD, Tong F (2016) Radial bias is not necessary for orientation decoding. *NeuroImage* 127:23–33.

Ramirez FM, Cichy RM, Allefeld C, Haynes JD (2014) The neural code for face orientation in the human fusiform face area. *The Journal of Neuroscience* 34:12155–12167.

Ross DA, Tamber-Rosenau BJ, Palmeri TJ, Zhang J, Xu Y, Gauthier I (2018) High-resolution functional magnetic resonance imaging reveals configural processing of cars in right anterior fusiform face area of car experts. *Journal of Cognitive Neuroscience* 30:973–984.

Sengupta A, Yakupov R, Speck O, Pollmann S, Hanke M (2017) The effect of acquisition resolution on orientation decoding from V1 BOLD fMRI at 7t. *NeuroImage* 148:64–76.

Soto FA, Waldschmidt JG, Helie S, Ashby FG (2013) Brain activity across the development of automatic categorization: A comparison of categorization tasks using multi-voxel pattern analysis. *Neuroimage* 71:284–897.

Soto FA, Vucovich LE, Ashby FG (2018) Linking signal detection theory and encoding models to reveal independent neural representations from neuroimaging data. *PLoS Computational Biology* 14:e1006470.

Van Bergen RS, Ma WJ, Pratte MS, Jehee JFM (2015) Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience* 18:1728–1730.

Wilson HR, Wilkinson F (2015) From orientations to objects: Configural processing in the ventral stream. *Journal of Vision* 15:4.