



COGNITIVE NEUROSCIENCE

The neural code for “face cells” is not face-specific

Kasper Vinken^{1*}, Jacob S. Prince², Talia Konkle², Margaret S. Livingstone¹

Face cells are neurons that respond more to faces than to non-face objects. They are found in clusters in the inferotemporal cortex, thought to process faces specifically, and, hence, studied using faces almost exclusively. Analyzing neural responses in and around macaque face patches to hundreds of objects, we found graded response profiles for non-face objects that predicted the degree of face selectivity and provided information on face-cell tuning beyond that from actual faces. This relationship between non-face and face responses was not predicted by color and simple shape properties but by information encoded in deep neural networks trained on general objects rather than face classification. These findings contradict the long-standing assumption that face versus non-face selectivity emerges from face-specific features and challenge the practice of focusing on only the most effective stimulus. They provide evidence instead that category-selective neurons are best understood by their tuning directions in a domain-general object space.

INTRODUCTION

High-level visual areas in the ventral stream contain neurons that are category-selective, in that they respond more to images of one category than to images of others. The most compelling examples are “face cells,” which are defined by a higher response to faces than to non-faces (“face selectivity”) and form a system of clusters throughout the inferotemporal cortex (IT) (1, 2). These clusters are large enough to reliably manifest as face-selective patches in functional imaging studies, where they are surrounded by non-face-selective regions (3, 4). A critical question is whether the face selectivity of face cells arises from face-specific circuits that process information about only faces and that are computationally distinct from circuits that process other objects (5–7). At the neural level, mechanisms that process only faces would involve encoding face-specific features—a holistic face context or actual face parts—that are computed nonlinearly from lower-level input features that apply to all kinds of objects, such as curvature, color, or shape. Here, we examine whether face cells respond selectively to faces because of features that are specific to faces.

Previous studies probing the nature of face-cell tuning have predominantly used face images, from the earliest studies presenting faces and individual face parts (8) or whole versus scrambled faces (9, 10), to more recent studies characterizing neural responses as a function of the position and arrangement of face parts (11–14). This exclusive focus on features that apply to only faces or vary among only faces has led to an understanding of face cells in terms of a constellation of face parts in a canonical face-like configuration. Responses of face cells to some non-faces (1) are often dismissed in this framework as epiphenomenal (e.g., they just look like a face) and of little interest. The same face bias is found in computational models of face cells, which are typically built to capture only face-to-face variability, and then fit and evaluated neural responses to only faces (12, 15, 16). Sometimes these models are not even applicable to non-faces (12, 17). This bias toward face-specific processing mechanisms in face-cell research raises an important

question: Are stimuli from the “preferred” category sufficient to characterize the tuning of category-selective neurons?

If, however, face cells are part of an integrated object space (18–20), in which face domains are embedded in a gradient of object selectivity, then it may be insufficient to use only face stimuli to characterize face cells. In this, domain-general, account, the tuning of face cells would better be understood in terms of discriminative visuo-statistical properties and would show systematic, meaningful, graded responses for all kinds of images (21). Face cells are defined on the basis of their separability between faces and non-faces, and thus, in a larger-scale population code, they could reflect tuning axes along which the image statistics of faces are particularly distinctive from the broader set of image statistics present in non-faces. Thus, critically, a domain-general account would predict systematic tuning information in the non-face responses of face cells, tuning that is not reducible to just face image statistics, and would be missed by analyzing only face responses.

Conversely, in a domain-specific view, if IT face cells encode high-level, face-specific information (e.g., the presence of actual face parts), then their response profile should be highly nonlinear with respect to nonspecific image characteristics like texture or shape, resulting in a tight tuning for face features (22) or shapes embedded in the holistic context of a face (11, 23). In this domain-specific account, any non-face responses should be either sparse and restricted to objects that look like a face or face part, or less sparse if the neuron multiplexes face-specific with other, independent information (24, 25). In both cases, there would be little information about face selectivity in responses to non-face stimuli. That is, if face selectivity relies categorically on the presence of actual face parts, in the holistic context of a face-like configuration, then face selectivity will not be linearly predictable from the neural tuning for attributes of objects without face parts or face-like configurations.

To explore the domain specificity of face-cell tuning, we analyzed neural responses in and around IT face patches to hundreds of different objects. We investigated the following two questions that distinguish domain-general from domain-specific coding: (i) whether the degree of face selectivity can be inferred from non-face responses and (ii) whether non-face responses contain information about face-cell tuning that cannot be inferred from

¹Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA.

²Department of Psychology, Harvard University, Cambridge, MA 02478, USA.

*Corresponding author. Email: kasper_vinken@hms.harvard.edu

responses to only face stimuli. We found that tuning for non-faces explained the degree of face selectivity (domain-general prediction 1) and that face-cell tuning was poorly estimated from only faces (domain-general prediction 2). Neural responses were not fully explained by a handful of intuitive object properties nor by a deep neural network (DNN) space tuned to face-to-face variability, but instead were accounted for best by complex image characteristics encoded by a DNN that represents an integrated object space. Thus, face selectivity does not reflect a face-specific code, but a preference for image characteristics that may not be intuitively interpretable but nevertheless correlate with faces. Even for highly face-selective sites, the tuning for those image characteristics could not be fully characterized by looking only at face-to-face variability, which has important implications for how face cells are currently studied and modeled. Overall, this conclusion is consistent with the hypothesis that face cells are not categorically different from other neurons, but that they together form a spectrum of tuning profiles in a shared space of discriminative features learned for all kinds of objects, and these tuning profiles are approximated by representations in later layers of artificial neural networks trained on general object classification.

RESULTS

For most of the results reported here, we analyzed recordings at 449 single (84) and multiunit (365) sites, in central IT [in and around the middle lateral (ML) and middle fundus (MF) face regions] of six macaque monkeys, in response to 1379 images: 447 faces and 932 non-face objects with no semantic or perceptual association with faces (Fig. 1, bottom examples). These face and non-face images were separated in terms of high-level faceness as estimated by a computational object recognition model (fig. S1). The majority of the central IT sites showed, on average, higher responses to faces than to non-face objects (333 of 449, ~74%). We quantified the face selectivity of each neural site by calculating a face d' selectivity index, which expresses the response difference between a face and a non-face object in SD units (see Materials and Methods; values >0 indicate a higher average response to faces than to objects). The larger the d' , the more consistent the response difference between faces and non-face objects. The average face d' was 0.84 (SD = 1.16) and ranged between -1.45 and 4.42 . Overall, response reliability was comparable for faces [mean $\rho_F = 0.69$, 95% confidence interval (CI) [0.67, 0.70]] and non-faces (mean $\rho_{NF} = 0.72$, 95% CI [0.70, 0.73]; Fig. 1B). The dynamic range of responses (i.e., normalized difference between minimum and maximum response; see Materials and Methods) was also comparable for faces (mean $DR_F = 0.48$, 95% CI [0.45, 0.51]) and non-faces (mean $DR_{NF} = 0.50$, 95% CI [0.47, 0.52]), but highly face-selective sites tended to have a higher dynamic range for faces (Fig. 1C).

To ensure that our conclusions also apply to classically-defined face cells, we separately report results for the 50 most face-selective neural sites from chronic arrays in the functional magnetic resonance imaging (fMRI)-localized face patches [face $d' > 1.25$, approximately corresponding to a face selectivity index $>1/3$ in our data (1); the total number of sites with face $d' > 1.25$ was 151]. For brevity, we will refer to these sites by the term "canonical face sites". The average face d' of canonical face sites was 2.40 (SD = 0.77) and ranged between 1.29 and 4.42. Despite the high face selectivity of this subset, response reliability was substantial for non-faces

(mean $\rho_{NF} = 0.64$, 95% CI [0.57, 0.70]) and only slightly higher for faces (mean $\rho_F = 0.72$, 95% CI [0.66, 0.77]; $\Delta\rho = 0.08$, $P = 0.0009$, 95% CI [0.04, 0.13]). The dynamic range was lower for non-faces (mean $DR_{NF} = 0.40$, 95% CI [0.33, 0.47]) compared to faces (mean $DR_F = 0.56$, 95% CI [0.49, 0.62]; $\Delta DR = 0.16$, $P = 0.0004$, 95% CI [0.08, 0.24]).

Thus, even face-selective sites showed reliable responses to non-faces, consistent with previous reports (1). Those previous reports showed relatively small responses for non-faces averaged across face cells, with a clear categorical boundary between faces and non-faces; that is, the smallest face response was larger than the largest non-face response (1). Similarly, fMRI blood oxygen level-dependent (BOLD) responses in the human fusiform face area (FFA) exhibit a distinct discontinuity, a category step, between the lowest responses to face images and highest responses to non-face images (21). Is this lack of overlap between face and non-face responses also true for individual neural sites? For each individual site, we used event-related responses to rank the face and non-face images from highest to lowest response magnitude. We then used only odd-trial responses for further analyses. Out of all 449 sites, 342 (30 canonical face sites) had a significantly higher response to the five best non-faces than to the five worst faces, whereas only 2 (both canonical face sites) had a significantly higher response to the five worst faces than to the five best non-faces ($\alpha = 0.05$; Fig. 2A). Thus, although 2 out of 50 canonical face sites did show a significant category boundary for the stimuli used here, the vast majority of the sites did not (see also fig. S2 for more analyses).

The lack of a category boundary in individual face sites may seem at odds with reports of a clear category boundary for monkey ML (1) and a category step in human FFA (21). A critical difference is that these studies evaluated a population average response. When we performed image ranking on the average population responses of the 50 canonical face sites, we did find a clear category step (Fig. 2B). This suggests that the categorical fMRI BOLD activations in face regions do not reflect a categorical code carried by single neurons but a property that emerges in a distributed population code where different neurons selectively respond to different images (26). The fact that only the population average shows a category step suggests that different face cells respond differently to different non-face objects, rather than that they all respond to the same objects based on how similar those objects are to faces.

Together, these results mean that responses even at the low end of the response profile are not just noise below some category boundary and that even the most face-selective sites carry consistent information about non-face objects. Normally, face-cell studies disregard these non-face responses in favor of using only faces to characterize neural tuning. Here, we went in the other direction and asked what we can infer about neural tuning from only non-faces.

Responses to non-face objects predict face selectivity

We now ask whether response profiles for non-face images are predictive of the degree of face selectivity, the defining property of face cells and face areas (1, 3, 4). Note that, consistent with the literature, we use the term face selectivity to refer to the face versus non-face response difference. If higher responses to faces are driven entirely by discriminative object features, rather than by features that apply to only faces, then the degree of face (versus non-face) selectivity should be linearly predictable from responses to non-faces alone.

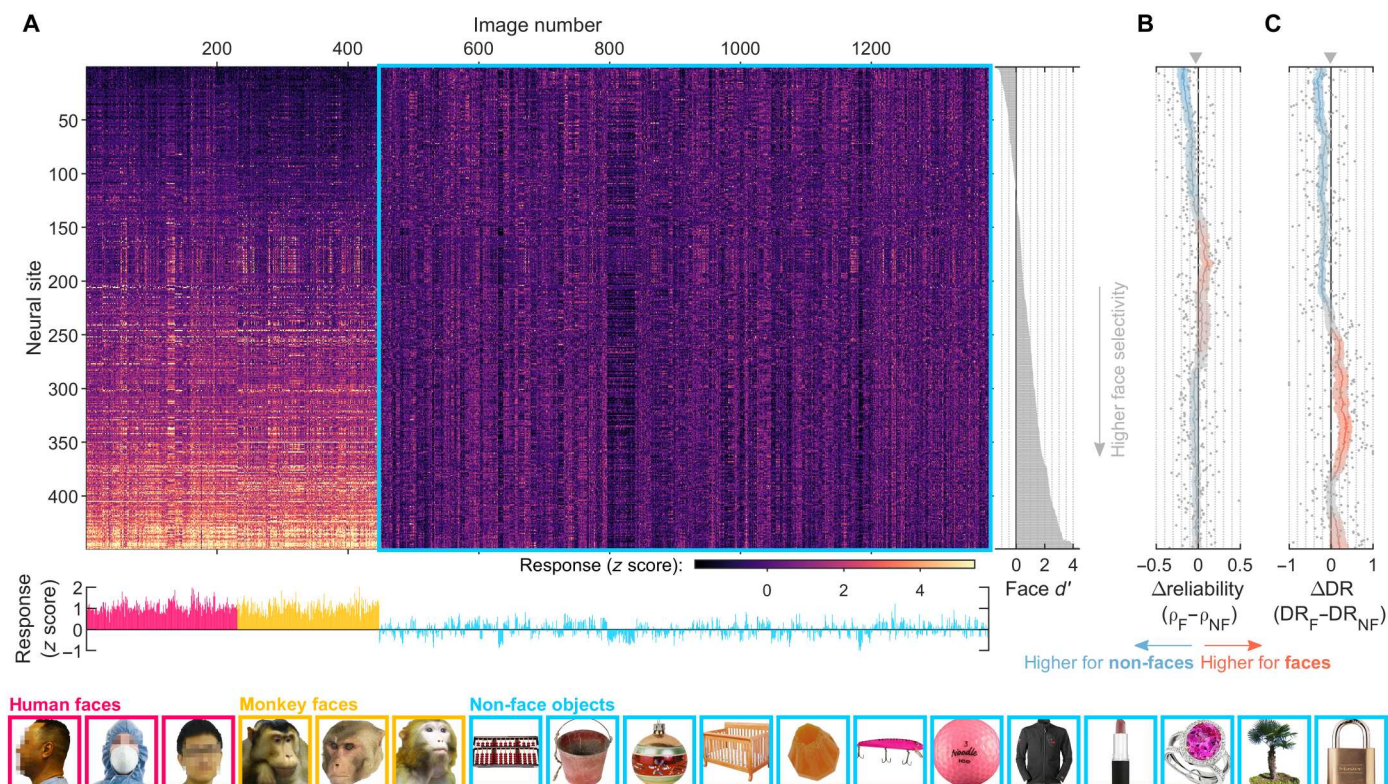


Fig. 1. Neural sites show reliable responses to non-faces regardless of face selectivity. (A) Responses of 449 central IT sites (top) and population averages (bottom marginal) to 230 human faces (pink), 217 monkey faces (yellow), and 932 inanimate non-face objects (blue; examples at the bottom; human faces were anonymized using pixelization). Responses were normalized (z-scored) per site using the means and SDs calculated from non-face object images only (blue outline). Sites were sorted by face selectivity (face d' , right marginal). (B) Difference in response reliability for faces (ρ_F) versus non-faces (ρ_{NF}). Each marker represents a single neural site, the same sorting as in (A). Shaded line and error bounds indicate a moving average (window = 51 sites) and 95% bootstrap CIs (calculated by resampling sites). Orange: Higher response reliability across face images. Light blue: Higher response reliability across non-face images. (C) Difference in the dynamic range of trial-averaged responses for faces (DR_F) versus non-faces (DR_{NF}). Same conventions as (B).

At a later point in this manuscript, we will look at predicting selectivity profiles across individual faces.

We took for each neural site the vector of non-face responses and standardized (z-scored) the average responses to each image to remove the effects of mean firing rate and scale (SD of firing rate). Next, we fit a linear regression model to predict the measured face d' values, using the standardized responses to non-face objects as predictor variables (see Materials and Methods). We used leave-one-session/array-out cross-validation to ensure that the model cannot exploit spurious correlations of simultaneously recorded neural activity. The results in Fig. 3A show that, using all 932 non-face object images, the model explained 65% of the out-of-fold variance in neural face d' ($R^2 = 0.65$, $P < 0.0001$, 95% CI [0.60, 0.69], Pearson's $r = 0.82$). This means that the response profiles for exclusively inanimate, non-face objects can explain most of the variability in face selectivity between neural recording sites. The explained variance increased monotonically as a function of the number of non-face images used to predict face selectivity, starting from ~10% for a modest set of 25 images Fig. 3B. Thus, image-level responses for non-face objects must be determined by features related to the neural site's category-level face selectivity.

The response profile across face images was less predictive of face selectivity ($R^2 = 0.36$, $P < 0.0001$, 95% CI [0.28, 0.43], Pearson's $r = 0.60$), even when the number of non-face images was subsampled to

match the number of face images (Fig. 3B; face selectivity predicted from 1000 subsamples of 447 non-face images: mean $R^2 = 0.59$, min $R^2 = 0.43$, max $R^2 = 0.68$). Across all 1000 subsamples, predictions from non-faces consistently had a lower mean squared error (MSE), and thus higher prediction accuracy, compared to predictions from faces (mean $\Delta\text{MSE} = -0.31$, min $\Delta\text{MSE} = -0.43$, max $\Delta\text{MSE} = -0.09$). This difference was even more pronounced for the subset of canonical face sites (mean $\Delta\text{MSE} = -1.09$, min $\Delta\text{MSE} = -1.50$, max $\Delta\text{MSE} = -0.57$). Notably, the reduced predictivity from faces cannot be explained by a lack of the dynamic range of responses to faces because the dynamic range for canonical face sites was higher for faces than for non-faces (mean $\Delta\text{DR} = 0.16$, $P = 0.0004$, 95% CI [0.08, 0.24]; see also Fig. 1C).

To demonstrate that more accurate predictions of face selectivity from only non-faces are not trivial results, we ran our analyses on the artificial unit responses of DNNs. We found that, when the network was pretrained with only faces, the face selectivity of artificial units was predicted less well (and, in some cases, not at all) from only non-faces than from only faces (fig. S3). Thus, units in an artificial neural network that was pretrained on face identification showed the opposite of what we observed in actual face cells.

These results indicate that face versus non-face selectivity is driven by general properties that are more variably represented in non-faces than in faces. This was despite the observation that how

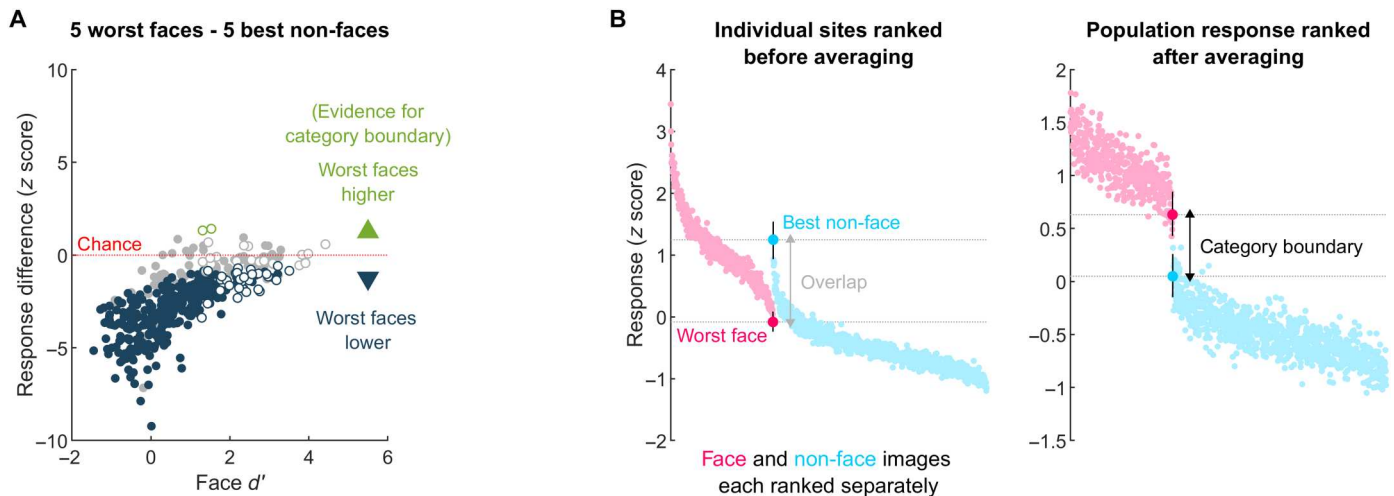


Fig. 2. Individual face sites do not show a category boundary. (A) Scatter plot of the response difference between the five worst faces and the five best non-faces (y axis), and face selectivity (x axis). The best and worst stimuli were identified using even-trial responses and then the response difference was calculated from odd-trial responses. Open markers indicate canonical face sites. Green markers: Sites for which the five worst faces had a higher response (two-sample t test; $\alpha = 0.05$). Blue markers: Sites for which the five best non-faces had a significantly higher response. A category boundary would require a higher response for the worst faces. (B) Odd-trial responses of the 50 canonical face sites to all stimuli ranked independently for faces and non-faces using even-trial responses. The two large, saturated markers indicate the responses to the worst face and the best non-face, with 95% bootstrap CI error bars calculated by resampling sites. Left: Stimuli were ranked for each site individually, before averaging responses across sites. Right: Stimuli were ranked after averaging across sites. Only the population response averaged across canonical face sites shows a category boundary.

face-like an image is (high-level faceness values assessed by DNN; fig. S1) varied more among faces than among non-faces. In the next section, we explore what these object properties could be.

Tuning to color and simple shape does not explain face selectivity

The prediction of individual sites' face selectivity from their non-face response profiles indicates that face-selective sites respond more to some non-face objects than to others. To quantify this, for each non-face image, we took the vector of z-scored responses across all neural sites (columns outlined in blue in Fig. 1A) and correlated it with the vector of face d' values. A positive correlation means that the image tended to elicit a higher response in more face-selective neural sites, and vice versa for a negative correlation. Figure 3C shows the 40 most positively and 40 most negatively correlated non-face images. By inspection, a potential interpretation is that face cells simply encode how "face-like" an object is. For example, a cookie or a clock does, in some ways, resemble a face more than a chair or a microscope does. However, even in the most face-selective sites, we found that individual sites consistently violated this interpretation by preferring some non-face objects over some faces, suggesting a more primitive explanation than perceived faceness (see Fig. 2 and figs. S1 and S2). In addition, the ability to predict face d' from non-face responses was not explained by only the most face-like non-face images (fig. S4).

By inspection, objects with responses positively correlated with face selectivity tended to be tan or red and round, whereas objects negatively correlated with face selectivity tended to be elongated or spiky. These observations raise the question of whether these simple object properties can explain the gradient of face selectivity and the gradient of non-face responses. Previous work has suggested that the majority of face cells are tuned to elongation/aspect ratio (11),

and the featural distinction between spiky versus stubby-shaped objects has recently been offered as an intuitive description of one of the two major axes in IT topography, including face patches (18). Similarly, properties like roundness, elongation, and spikiness were shown to account for object representations outside face-selective regions in anterior IT (27).

For each object, we computed the following seven properties: elongation, spikiness, circularity, Lu/v' color coordinates (L refers to luminance, u' and v' are chromaticity coordinates; see Materials and Methods), and a low-level face-correlation index [defined as the averaged pixel-wise correlation between a non-face and all face images, serving as a quantification of low-level, contrast-based face configuration (28)]. Each neural site's selectivity for these properties was quantified by computing the Spearman's rank correlation between the object property and the neural response (rows outlined in blue in Fig. 1A). Face d' correlated negatively with selectivity values for elongation and spikiness and positively with selectivity values for face-correlation, circularity, redness (u'), and yellowness (v' ; Fig. 3D). That is, these properties were correlated with the information encoded by face cells, but how much of the variance in face selectivity can they explain together? We fit a model (same methods as for Fig. 3A) to predict face d' as a linear combination of the selectivity values for these properties. Figure 3E shows that the combined model explained only ~13% of the out-of-fold variance in observed face d' ($R^2 = 0.13$, $P < 0.0001$, 95% CI [0.06, 0.19], Pearson's $r = 0.39$), falling short of the 65% explained by the non-face responses themselves. Thus, only a fraction of the link between non-face responses and face selectivity can be explained by color and simple shape properties.

However, these intuitively interpretable features represent a trade-off between simplicity and the ability to capture the rich complexity of object properties in our visual environment. We will

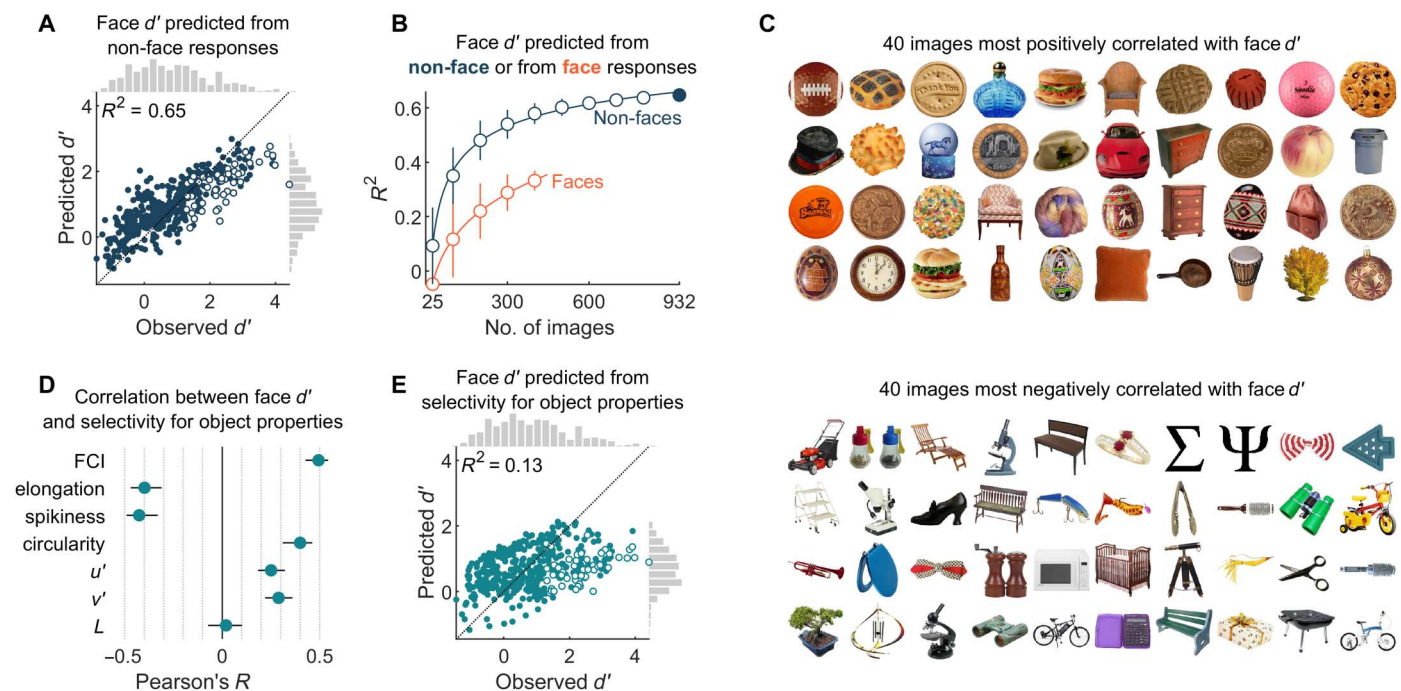


Fig. 3. Face selectivity is predicted by the response profile for non-face objects, but not by tuning to color and simple shape. (A) Observed face d' values and the values predicted from the pattern of responses to all 932 non-face objects (outlined in blue in Fig. 1A). Each marker represents a single neural site. Open markers indicate canonical face sites. The dotted line indicates $y = x$. (B) Out-of-fold explained variance as a function of the number of non-face (dark blue) or face (orange) images used to predict face d' (means \pm SD for randomly subsampling images 1000 times). The filled marker indicates the case shown in (A). (C) The 40 non-faces with the highest positive (top) or lowest negative (bottom) correlation with face d' (sorted left to right, top to bottom). (D) Correlations between face d' and selectivity for object properties (FCI, face-correlation index; error bars: 95% bootstrap CIs, calculated by resampling sites). (E) Face d' values predicted from selectivity for the seven object properties of (C). Same conventions as (A).

address this next by leveraging the representational capacity of DNNs trained on natural images.

Face selectivity and non-face responses share a common encoding axis

We asked whether the link between category-level face selectivity and non-face responses could be better explained by statistical regularities encoded in convolutional DNNs. We used two DNN architectures [Inception (29) and AlexNet (30)], pretrained on three different image datasets to do general object categorization [ImageNet (31)], scene categorization [Places365 (32)], or face identity categorization [VGGFace2 (33)]. Note that ImageNet also contains some images with faces (albeit no separate face category), so what sets it apart from VGGFace2 is not the absence of faces, but the fact that it represents an integrated object space, including faces. The image-statistical regularities, or DNN features, encoded by a pre-trained network are not necessarily intuitively interpretable, like face parts or spikiness, but they have been shown to explain a substantial amount of variance in IT responses (34, 35). After pretraining, DNN activations to images can be linearly mapped to neural responses to obtain a DNN encoding model, which provides an estimate of the direction ("encoding axis") in the DNN representational space associated with the response gradient. Thus, generating a DNN encoding model involves a pretraining phase, where the model learns a basis set of DNN features optimal for object/scene/face classification, followed by a linear mapping phase where these DNN features are fit to neural responses using

a separate training set of images and corresponding responses. If common image characteristics account for both face selectivity and non-face responses, then the encoding model fit on responses to only non-face images should also predict face versus non-face selectivity and possibly also image-level face response profiles (see the next section). We refer to encoding model fit on only non-face responses as "non-face encoding models" (regardless of the pretraining history of the base DNN) and to encoding model fit on only face responses as "face encoding models".

We first calculated the explained variance in face d' for the non-face encoding models of each object-pretrained inception layer (Fig. 4A) and found that it increased from 6% for the input pixel layer up to the highest value of 57% ($R^2 = 0.57$, $P < 0.0001$, 95% CI [0.51, 0.62], Pearson's $r = 0.76$; Fig. 4, A and B) for inception-4c (yellow marker in Fig. 4A). This implies that in the inception-4c representational space, a single non-face encoding axis largely captures both image-level responses for non-face objects and category-level selectivity between faces and non-faces. Because this axis was fit for each site independently, this analysis cannot rely on spurious correlations between non-face responses and face selectivity. Therefore, we rule out the possibility that non-face responses result from tuning that does not contribute to, but is spatially correlated with, face selectivity in IT (see also fig. S5). The fact that the explained variance in face d' is low for early DNN layers and increases to its maximum in inception-4c implies that mid- to high-level image characteristics best explain the link between non-face responses and face selectivity [see fig. S6 for a meta-model combining the

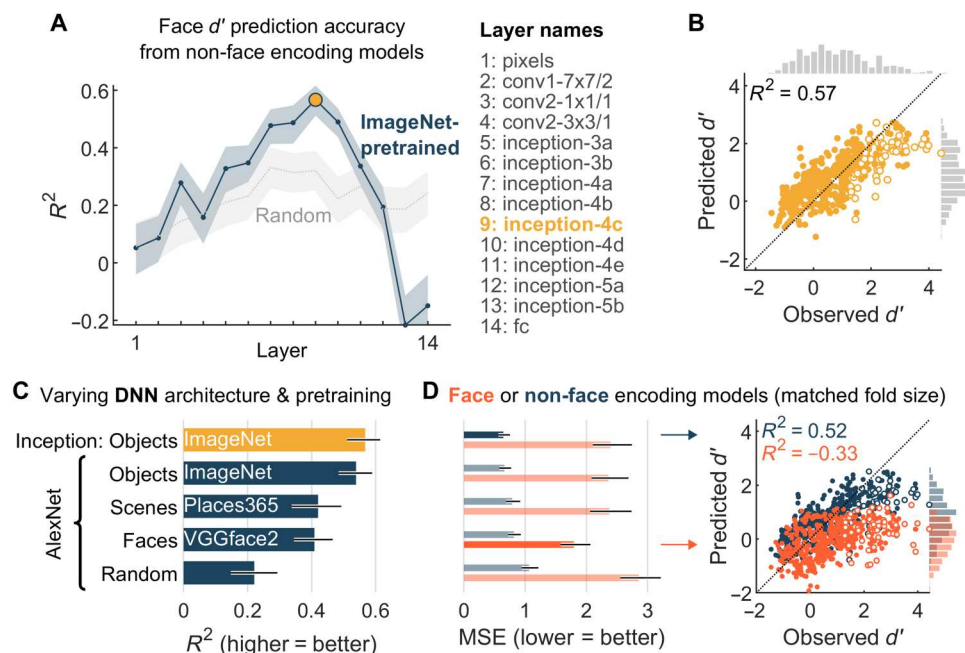


Fig. 4. DNN encoding model fit exclusively on non-face responses predict face selectivity. (A) The amount of variance in neural face d' , explained by non-face encoding models [successive layers of Inception (29); see Materials and Methods; with 95% bootstrapped CI, calculated by resampling sites]. For each neural site, we fit 14 separate encoding models using only responses to non-face objects: one for each DNN layer, starting from the input space in pixels. Light gray: Models based on a randomly initialized DNN that was not pretrained. (B) Observed face d' values and the values predicted by the inception-4c layer model [yellow marker in (A)]. Each marker represents a single neural site (open markers = canonical face sites). The dotted line indicates $y = x$. (C) Explained variance in face d' (with 95% bootstrapped CI, calculated by resampling sites) predicted by non-face encoding models based on various DNNs (best layer): Inception pretrained on object classification [same as in (A) and (B)]; AlexNet (30) randomly initialized or pretrained on object, scene, or face classification. (D) Left: MSE (with 95% bootstrapped CI, calculated by resampling sites) for face d' predictions from face or non-face encoding models, using the same base models as (C). For this comparison, non-face encoding models were fit by subsampling stimuli to match the training-fold size of face encoding models. Nonfaded bars indicate the best face/non-face encoding model (lowest MSE). Right: Observed versus predicted face d' values from the best face (orange) and best non-face (blue) encoding model.

information from non-face response profiles (Fig. 3A) and the non-face DNN encoding model (Fig. 4B)].

The prediction of face selectivity did not rely on DNN units that themselves are maximally activated by a face but on units whose responses distinguish moderately (absolute average d' values of ~ 0.2 or smaller; compare to the face-cell criterion of $d' > 1.25$) between faces and non-faces (fig. S7). This suggests that face selectivity is not driven by features that are maximally present in a face but by a combination of many features that apply to all kinds of objects yet are differentially prevalent in faces versus non-faces.

Next, we asked how the ability to predict face selectivity depended on the architecture or the pretraining set of the DNN base model. We found that non-face encoding models based on object-pretrained AlexNet performed nearly as well as object-pretrained Inception, but scene-pretrained and even face-pretrained versions of AlexNet performed significantly worse (Fig. 4C, nonoverlapping CIs). Even for the canonical face sites with the strongest face selectivity, predictions from the object-pretrained non-face encoding model (AlexNet) had a lower MSE compared to predictions from the face-pretrained non-face encoding model ($\Delta\text{MSE} = -0.42$, $P = 0.0222$, 95% CI $[-0.80, -0.09]$). Thus, the link between responses to non-faces and face selectivity is best captured by a base set of image characteristics that represent an integrated object space, rather than by a base set optimized specifically for faces.

Complementing this result, the face encoding models (i.e., for which we used only neural responses to face images to derive an encoding model) were bad at predicting the degree of face selectivity: Every single face encoding model had a negative R^2 , meaning that the model predictions were less accurate than simply using the average observed face d' as a prediction for each site. Non-face encoding models (with training-fold size matched to face encoding models) performed better regardless of the DNN architecture or pretraining set (Fig. 4D, left). The face encoding model performed best when it was based on a face-pretrained AlexNet (nonoverlapping CIs), but the prediction accuracy was still poor ($R^2 = -0.33$, 95% CI $[-0.47, -0.20]$, Pearson's $r = 0.49$; Fig. 4D, right) with a substantially higher MSE compared to predictions from the best non-face encoding model ($\Delta\text{MSE} = -1.15$, $P < 0.0001$, 95% CI $[-1.36, -0.95]$). This difference was more pronounced for the subset of canonical face sites ($\Delta\text{MSE} = -3.21$, $P < 0.0001$, 95% CI $[-4.26, -2.31]$).

Up to this point, we have focused on neural sites located in central IT. Figure S8 and the accompanying Supplementary Text show that, like neurons in central IT, face selectivity in anterior lateral (AL) face patch was also linked to tuning for non-face objects. Thus, the encoding axis in an integrated object space (and not a face-specific space), estimated from responses to only non-face objects, captured face versus non-face selectivity of face cells in both central and anterior IT.

Image-level predictions of face and non-face encoding models

Up to this point, we have focused on face versus non-face selectivity, the defining property of face cells. In the previous section, we found that encoding axes estimated from non-face responses best predicted a neural site's degree of face selectivity. Beyond such category selectivity d' measures, do these models also explain selectivity for individual images?

We first asked how well the non-face versus face encoding models (both based on object-pretrained inception-4c) captured the neural population representation of all stimuli, using representational similarity analysis (36). For each stimulus pair, we computed the population response dissimilarity (cosine distance between rows in the stimulus \times site matrix), resulting in three representational dissimilarity matrices: one for the neural data, one for the non-face encoding model, and one for the face encoding model (Fig. 5A). We compared these dissimilarity matrices by computing Spearman's rank correlation (r_s) using off-diagonal elements. Overall, the non-face encoding model (with training-fold size

matched to the face encoding model) captured the neural representational geometry quite well (mean $r_s = 0.71$, 95% CI [0.69, 0.72]; canonical face sites: mean $r_s = 0.69$, 95% CI [0.66, 0.72]), significantly better than did the face encoding model (mean $r_s = 0.52$, 95% CI [0.49, 0.55]; canonical face sites: mean $r_s = 0.32$, 95% CI [0.25, 0.39]), by correctly capturing the dissimilarity of individual non-faces to faces and by separating monkey from human faces. The face encoding model also separated monkey from human faces but failed to capture the dissimilarities of individual non-faces to faces.

The representational similarity analysis provides an overall picture of the population representation, but how well do these models predict the selectivity for individual face images per neural site? For each neural site, we computed prediction accuracy (\bar{r}) as Pearson's correlation coefficient between observed and predicted responses (concatenated from all test folds) normalized by the response reliability. Because this and the next analyses evaluated accuracy separately for faces and non-faces, we excluded 58 sites (mean face $d' = 0.45$; SD = 1.10) that had response reliability

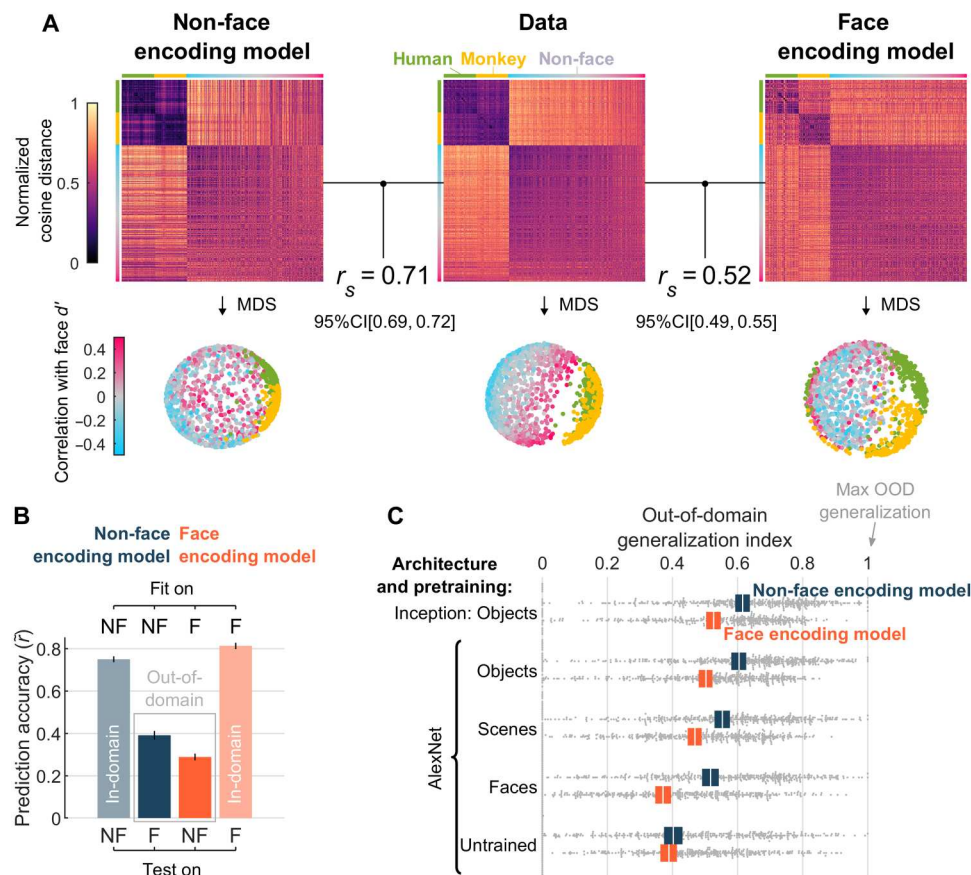


Fig. 5. Image-level predictions of the face and non-face encoding models. (A) Representational geometry reproduced by the face and non-face encoding model (object-pretrained inception-4c). Top: Dissimilarity matrices for all stimulus pairs, non-faces are sorted by the correlation with face d' (see Fig. 3C). Spearman's rank correlation (r_s) was used to compare neural and model representations using off-diagonal elements. Bottom: Visualization of dissimilarity matrices using metric multi-dimensional scaling (criterion: stress). (B) Image-level accuracies of the face and non-face encoding model, separately for the face (F) and non-face (NF) images (object-pretrained inception-4c). Accuracy (\bar{r}) was computed as the Pearson's r between observed and out-of-fold predicted responses, normalized by the response reliability. For the non-face encoding model, face images are out of domain (OOD), whereas for the face encoding model, non-face images are OOD. Error bars: 95% bootstrapped CI, calculated by resampling sites. (C) OOD generalization for encoding models based on various DNNs, showing results for the layer with the best OOD accuracy, thus allowing for the face and non-face encoding models to be each based on a different layer. Colored boxes: 95% bootstrapped CI, calculated by resampling sites; white line: mean; gray dots: individual sites.

below a threshold of 0.4 (see Materials and Methods) for either faces or non-faces, leaving 391 remaining sites (mean face $d' = 0.90$; SD = 1.16; $N = 38$ canonical face sites). This exclusion did not qualitatively affect the results. For faces, the non-face encoding model (mean $\tilde{r} = 0.39$) had a lower prediction accuracy than the face encoding model (mean $\tilde{r} = 0.81$; mean $\Delta\tilde{r} = -0.42$, $P < 0.0001$, 95% CI $[-0.44, -0.40]$; using object-pretrained inception-4c). Conversely, for non-faces, the non-face encoding model (mean $\tilde{r} = 0.75$) had a higher prediction accuracy than the face encoding model (mean $\tilde{r} = 0.29$; mean $\Delta\tilde{r} = 0.46$, $P < 0.0001$, 95% CI $[0.45, 0.48]$). Thus, each encoding model generalized best within the same domain as the stimuli used for fitting the model to neural responses (Fig. 5B).

Does this mean that face-to-face response variability is partially determined by face-specific features? This could be the case. However, an alternative explanation is that the DNN encoding models overfitted on the stimulus domain used for mapping the model to neural data. That is, if only non-face images are used to fit the encoding model of a face cell in this object-trained DNN, then the encoding model will overfit to non-face-image statistics. Similarly, if only face images are used to fit the encoding model of an object-trained DNN, then the encoding model will overfit to face-image statistics. Using model simulations, we empirically tested and confirmed this overfitting hypothesis, showing that lower out-of-domain prediction accuracy is expected even when face-to-face response variability is not determined by face-specific features (fig. S9). Thus, a lower out-of-domain prediction accuracy by itself cannot be interpreted as evidence for domain-specific features.

To further test this possibility of face-overfitting on the neural data, we fit a face encoding model on either human faces or monkey faces only and evaluated the model using the faces of the held-out species. By doing this, we decreased the dependence between training and test folds in terms of low-level features, which can inflate generalization accuracy, while staying within the domain of faces and thus retaining dependence in terms of facial features (eyes, mouths, noses, etc.). The prediction accuracy of the face encoding model (mean $\tilde{r} = 0.34$, averaged across held-out species) was not significantly better than that of a non-face encoding model with matched training-fold size (mean $\tilde{r} = 0.32$; mean $\Delta\tilde{r} = -0.02$, $P = 0.0800$, 95% CI $[-0.04, 0.00]$). For canonical sites, there was a small advantage for the face encoding model (mean $\Delta\tilde{r} = -0.08$, $P = 0.0294$, 95% CI $[-0.16, -0.02]$), but this advantage disappeared when canonical sites were matched in response reliability for faces and non-faces (mean $\Delta\tilde{r} = -0.03$, $P = 0.3234$, 95% CI $[-0.11, 0.03]$, after regressing out the reliability difference). Thus, despite the shared facial features between monkey and human faces, a face encoding model fit on one species was not markedly better at predicting response variability among the faces of the held-out species than the non-face encoding model.

This raises the question of which model was better at predicting out-of-domain responses. After all, if a model truly captures the attributes that a neuron is tuned to, then it should generalize beyond images similar to the training set. We computed a generalization index (GI; see Materials and Methods) that quantifies how close an encoding model's out-of-domain prediction accuracy (e.g., on faces for the non-face encoding model) is to the within-domain prediction accuracy (e.g., on faces for the face encoding model). The higher the GI, the better the out-of-domain generalization. The

encoding model fit using non-faces achieved significantly better out-of-domain generalization than did the face encoding model (mean $\Delta\text{GI} = 0.09$, $P < 0.0001$, 95% CI $[0.07, 0.11]$; canonical face sites: mean $\Delta\text{GI} = 0.09$, $P = 0.0187$, 95% CI $[0.02, 0.17]$; Fig. 5C, object-pretrained inception). This is an important result, as it goes directly against intuitions that responses to non-faces just reflect the degree to which they look like a face. Here, we see that variation among non-face images is better able to extrapolate and predict response variation among faces than the other way around. Furthermore, the gap between non-face and face encoding models was largest for a face-pretrained AlexNet, suggesting that an encoding model primed to capture face-to-face variability is more likely to overfit on neural responses to face images and thus less likely to capture the actual tuning axis of face cells.

Last, encoding models based on object-pretrained AlexNet generalized better from faces to non-faces (and vice versa) than those based on face-pretrained AlexNet (nonoverlapping CIs for Alexnet—objects and Alexnet—faces in Fig. 5C). This suggests that, like face versus non-face selectivity, face-cell responses to individual non-faces are best captured by a base set of image characteristics that represent an integrated object space, rather than by a base set optimized for discriminating between faces.

In sum, non-face encoding models best captured the overall representational geometry and performed better on out-of-domain generalization than face encoding models, even for highly face-selective neurons. This suggests that models that capture only face-to-face variability, or that are fitted on only face-to-face response variability, are more prone to overfitting and thus do not capture underlying attributes.

DISCUSSION

In this study, we investigated the tuning for non-face objects in neural sites in and around face-selective regions ML/MF and in AL of macaque IT. The neural sites spanned a graded spectrum of face versus non-face selectivity, ranging from not face-selective to strongly face-selective (Fig. 1). We found that face selectivity was linearly related to responses to non-face objects: The response profile for non-faces could predict the degree of face selectivity across neural sites, while the prediction from the face-response profile was significantly worse (Fig. 3). Interpretable object properties such as roundness, spikiness, or color explained only a fraction of the relationship between face selectivity and the response profile to non-faces. Instead, image attributes represented in higher layers of an object classification-trained DNN could best explain this link: The DNN encoding axis estimated from responses to non-face objects could predict the degree of face selectivity (Fig. 4) and predict variation among individual face images (Fig. 5). In contrast, encoding models fit on only face responses performed less well overall on predictions of face versus non-face selectivity (Fig. 4) as well as out-of-domain image-level responses (Fig. 5). Last, face-pretrained DNNs which directly learn features that capture face-to-face variation were significantly worse at predicting the overall responses of face cells, whether fit with faces or non-face images. Thus, tuning in macaque IT face patches is not face-specific.

Broadly, our results imply that face-cell responses to non-faces are determined by discriminative object features that also explain face versus non-face selectivity. Therefore, at its core, face selectivity in the ventral stream should not be considered a semantic code

dissociable from visual attributes. Nor does it require features that can strictly be considered face parts, either based on their visual characteristics or based on a face-like configuration (8–14). We also showed that neurons did not respond based on the high-level resemblance of a stimulus to a face (figs. S1 and S2). Instead, our results indicate that the degree of face selectivity in the macaque IT cortex is a correlate of the underlying tuning for multiway object discrimination, without specialized nonlinear tuning for face-specific features. In other words, face selectivity in both central and anterior IT did not depend on the presence of a face-specific feature (e.g., an actual eye), but instead can be linked to features that also discriminate between non-faces. Such a neural code, that is not restricted to sparse responses for only the preferred stimulus domain, could be more flexibly used for across-domain generalization and may form the basis for generalization from very few exemplars of unfamiliar object classes.

Whether the encoding axes of face cells also include additional attributes that apply to and vary among only faces remains an open question. We did find that the non-face encoding models explained responses to faces less well than did face encoding models (and vice versa)—which could indicate different axes for faces and objects, but it could also be a consequence of model overfitting (fig. S9; see discussion below). We do not exclude the possibility that neural responses become absolutely face-specific in regions downstream of AL, perhaps as early as the anterior medial face patch, which we did not investigate.

Is it even possible for a neuron to compute visual features that apply to only faces? Face-specific responses could be achieved in a biologically plausible way by applying a response threshold to a graded face-selective input (see fig. S5). Such categorical tuning is also computationally plausible: In AlexNet, we found a substantial number of units with strictly category-specific output for the stimulus set of the current experiment. Most of these units were found in the penultimate layer (fc7; i.e., right before the classification output) and depended on the DNN training set: In face-trained AlexNet, 19% of fc7 units were activated by only faces and <1% by only non-faces; in object-trained AlexNet, <1% of fc7 units were activated by only faces and 21% by only non-faces. However, these proportions were negligible in the earlier model layers which best explained the neural data (relu4-pool5, with each <1% face-specific units). These observations suggest that strictly category-specific processing, independent of other objects, may not occur at the level of the ventral stream, which is usually best approximated by convolutional/pooling layers preceding fc7 (37–39).

The rapid presentation of randomly interleaved stimuli might raise concerns about face-to-face adaptation given the high proportion of face stimuli (~18% of the total; see Materials and Methods). However, adaptation of IT responses is proportional to the feature similarity between successive stimuli (40–42) and thus attribute-specific (not category-specific), presumably based on the shared input that a neuron receives for attributes shared between stimuli (43). We found only negligible category-selective suppression: The net response to a face preceded by a face was slightly lower relative to a face preceded by a non-face [Median = 97.3%, interquartile range (IQR) = 100.2 to 94.3%], and this reduction was comparable to that of a non-face preceded by a non-face relative to a non-face preceded by a face (Median = 98.1%, IQR = 101.2 to 93.3%). This is consistent with the hypothesis that face selectivity is explained by shared attributes of faces and non-faces because the

response components that encode those attributes were similarly adapted for faces and non-faces.

Our claim that the neural code for face cells is not face-specific is not an argument against operational face selectivity (i.e., larger responses to faces than to objects) or against a role for face cells in the perception of faces. Face cells do respond more to faces and, correspondingly, are causally involved in the processing of faces (44–46). However, consistent with our claim that face selectivity depends on domain-general features, in macaque studies, microstimulation of face or body patches affects the perception of stimuli from other categories, though to a lesser extent than the perception of images from the patch category (44, 45, 47). The fact that face-patch (or body-patch) stimulation affects the perception of non-face (or non-body) objects could reflect the extent to which the encoded features apply to each object, rather than whether the object belongs to a particular category. Alternatively, as proposed by Schalk *et al.* (46), face domain–stimulation effects on non-face object perception could arise if the stimulation produced a hallucination of a face. However, Azadi *et al.* (48) recently reported that optogenetic stimulation of macaque IT does not result in a detectable hallucination, but rather a distortion of whatever object the animal is viewing, and the detectability of the effect depends on what the animal is looking at.

Which attributes underlie category selectivity in face cells

What attributes underlie face selectivity, if not strictly face-specific features? The better performance of object-pretrained DNNs suggests that face selectivity is best explained by discriminative object features and is in line with previous observations (17, 37, 38). These features are not entirely low-level or spatially localized because non-face encoding models based on pixels or earlier DNN layers did not predict face selectivity and later DNN layers did. These later layers encode image statistics that correlate with the presence of high-level visual concepts, such as object parts, body parts, or animal faces (49). However, these image statistics are not discrete or categorical in nature, and they correlate with mid-level feature distinctions, such as curvy versus boxy textural statistics (50) and spiky versus stubby shapes (18). These descriptors are useful for providing general intuitions about the nature of the visuo-statistical features underlying object representation, but we suspect they should not be taken as a claim about a simple underlying basis set for object representation. For example, in the present data, selectivity for spikiness, color, aspect ratio, roundness, and pixel-level face configuration correlated with face selectivity, yet, in a cross-validated regression, these properties explained only a small part of the variance in face selectivity (Fig. 3). Thus, the tuning of face cells may not be reducible to intuitively interpretable human labels like faceness or roundness but may comprise a complex mixture of attributes that emerge in a distributed coding framework (51, 52).

Implications for a face bias in face-cell and face domain research

The fact that non-face responses allowed us to infer information about face-cell responsiveness that could not be characterized using only faces implies that we need to explore responses to non-face objects to fully understand the tuning of face cells. This idea is a substantial departure from most previous approaches (including from our own laboratory), which first use face and non-face images to identify face cells, but then use only faces to further

characterize face-cell tuning (11–13). Similarly, computational models of face cells are often not evaluated on non-faces (15–17) or the model may represent only face-to-face variation that does not apply to non-face objects (12). While these previous studies represent important milestones in the effort to understand face-to-face selectivity, our current work suggests that models built to discriminate only faces and optimized for only face responses may not reproduce important properties of face cells.

The worse characterization of neuronal face selectivity by their responses to faces cannot be explained by a narrower range of responses for faces nor by a lower response reliability. For the canonical face sites, both the response reliability and the response range were higher for faces compared to non-faces, yet face selectivity and out-of-domain image selectivity were still predicted worse from face responses than they were from non-face responses. This highlights one of the most important implications of this study, namely, that tuning that is inferred from only faces may not capture the actual stimulus attributes that explain face-cell responses in a generalizable way. In other words, a tuning model fit on only faces may overfit (or underfit) on faces, and, because of the pixel-level similarity and thus dependency between individual face images, this problem will go unnoticed when the model is evaluated on only face images (fig. S9) (53). Similarly, the semantic-categorical or parts-based views of face selectivity may reflect a human bias in interpretations that have overfit on the (limited) categorical or parts-based stimulus sets used in past experiments. Thus, we believe that an experimental bias toward the preferred category leads to a category-specific bias in understanding, and we suggest that future studies on the tuning of IT neurons should not be limited to a single preferred stimulus domain that occupies a narrow part of the stimulus space and/or has high physical homogeneity.

Although this study addresses the response properties of individual neurons, it has implications for understanding functional specificity at a larger scale in the ventral visual pathway. Our microscale results on the face selectivity of individual neurons have meso- and macroscale corollaries, as well as developmental and teleological implications. At the macroscale is the related question of whether face domains, comprising clusters of face-selective neurons, process only faces (5), and the degree to which other category-selective domains are specific for their preferred category. A mesoscale question is whether the circuitry of face, or other category, domains is computationally distinct from the neuronal circuits that process other objects. The answers to these questions are not necessarily either domain-specific or domain-general at all three levels (54), though these levels can be logically linked.

The microscale domain-specific view that face cells encode information specific to faces, which could rely on the nonlinear detection and gated processing of facial features (7), is consistent with macro-level specificity—that face domains (clusters of face neurons) process only faces. Both micro- and macro-domain specificity are often linked with the mesoscale idea that face-processing circuitry is computationally distinct from mechanisms that process other objects (5–7). Furthermore, at the teleological and developmental level, if face (or body, or scene) domains are specialized to process specifically only their preferred category, possibly using circuitry optimized to process that category, then it is logical to invoke evolutionary pressure for developing such innate domains, to serve the recognition of these biologically important categories. As McKone and Kanwisher (55) put it: domain-specific

mechanisms are “highly specialized processors that operate on specific kinds of information, that develop early, and that are likely to be evolutionarily conserved.”

Alternatively, our microscale results that face neurons show domain-general response properties are consistent with a macroscale domain-general hypothesis spanning these same levels: that category-selective regions are not functionally discrete but are part of an integrated object space supported by non-specific visuo-statistical characteristics (18, 27). In this domain-general view, at the microscale, the tuning of face-selective neurons is explained by a linear combination of visual attributes that apply to all kinds of objects, and at the mesoscale, the circuitry in different domains is the same, just with different inputs. Furthermore, domain-general tuning would be consistent, both developmentally and teleologically, with domain-general genetic programs that interact with prenatal spontaneous activity and postnatal experience to sculpt neuronal selectivity according to the statistics of the environment (56), constrained by a map-based proto-architecture (57).

Thus, at both a macroscale and a developmental view of the IT cortex, our results are consistent with evidence converging on a unified organization based on domain-general tuning for texture, shape, and curvature that underlie and support categorical distinctions (18, 19, 50, 58–61). These features may be scaffolded in retinotopy and, hence, receptive-field scale, which is present at birth, and likely require patterned visual experience to develop (57, 62). Before this study, it was not known whether IT face selectivity is predicted by domain-general visual characteristics, or whether IT neurons additionally rely on category-specific features to generate face selectivity. Our results support the notion that category maps can be accounted for by tuning in an integrated feature space (18, 20).

In conclusion, we show that the neural code of IT face cells is not face-specific, in the sense that (i) face versus non-face selectivity can be predicted from responses to non-faces, which are best modeled by features optimized for untangling all kinds of objects and (ii) non-face responses provide information about face-cell tuning that is not well characterized by face images. This does not mean that face cells are not substantially involved with face processing, but that macaque face patches are not modules strictly specific to faces. Features that apply only to faces or explain only face-to-face variability are not a sufficient explanation of face cells, and understanding tuning in the context of an integrated, domain-general object space is required. This conclusion is consistent with the hypothesis that face cells are not categorically different from other neurons but that they together form a spectrum of tuning profiles in a shared space of discriminative features learned for all kinds of objects. More generally, these results challenge the practice of focusing on only the most effective stimulus or category to study neural tuning.

MATERIALS AND METHODS

Animals

Eight adult male macaques (8 to 12 kg) were used in this experiment: six rhesus macaques (*Macaca mulatta*) aged 4 to 13 years old (four provided by the New England Primate Research Center and two from Harvard Medical School) and two pigtailed macaques (*Macaca nemestrina*) aged 10 and 11 years old (provided by Johns Hopkins). Seven were implanted with chronic microelectrode

arrays in the lower bank of the superior temporal sulcus: five monkeys at the location of the middle face region (ML and MF) and two monkeys at the location of the anterior face region (AL). One monkey had a recording cylinder for acute recordings implanted over the middle face region. All procedures were approved by the Harvard Medical School Institutional Animal Care and Use Committee (protocol #ISO00001049) and conformed to National Institutes of Health guidelines provided in the Guide for the Care and Use of Laboratory Animals.

Behavior

The monkeys were trained to perform a fixation task. They were rewarded with drops of juice to maintain fixation on a spot in the middle of the screen (a 53-cm LCD monitor in front of the monkey). Gaze position was monitored using an ISCAN system (ISCAN, Woburn, MA). MonkeyLogic (<https://monkeylogic.nimh.nih.gov/>) was used as the experimental control software. As long as the monkey maintained fixation, images were presented at a size of 4 to 6 visual degrees and at a rate of 100 ms on and 100 to 200 ms off. Images were presented foveally for acute recordings and at the center of the mapped receptive field for chronic recordings. An average of 9.7 (SD = 3.4) trials were presented per stimulus.

Recording arrays

Five monkeys were implanted with 32-channel floating microelectrode arrays (Microprobes for Life Sciences, Gaithersburg, MD) in the middle face region, identified by a fMRI localizer (see below). One monkey had an acute recording chamber positioned over the middle face region (identified by fMRI), and neuronal activity was recorded using a 32-channel NeuroNexus Vector array (Ann Arbor, MI) that was inserted each recording day. The two remaining monkeys were implanted with 64-channel NiCr microwire bundle arrays (Microprobes for Life Sciences, Gaithersburg, MD) (63) in the AL face region, identified by fMRI localizer in one monkey and based on anatomical landmarks in the other (64).

fMRI-guided array targeting

In all but one monkey, the target location of face patches was identified using fMRI. Monkeys were scanned in a 3T TIM Trio scanner with an AC88 gradient insert using four-channel surface coils (custom-made by A. Maryam at the Martinos Imaging Center), using a repetition time of 2 s, echo time of 13 ms, flip angle (α) of 72°, iPAT = 2, 1-mm isotropic voxels, matrix size of 96 × 96 mm, and 67 contiguous sagittal slices. Before each scanning session, monocrySTALLINE iron oxide nanoparticles (12 mg/kg; Feraheme, AMAG Pharmaceuticals, Cambridge, MA, USA) were injected into the saphenous vein to enhance contrast and measure blood volume directly. To localize face-selective regions, 20-s blocks of images of either faces or inanimate objects were presented in randomly shuffled order, separated by 20 s of a neutral gray screen. Additional details are described in (62).

Stimuli

During the experiments, the monkeys were presented with a total of 2550 images with objects on a white background that were also presented in (65). Most of those images were from (66), but some of the human face images and the monkey face images were from our laboratory. For the purpose of this study, we selected a priori a subset of 932 images of inanimate objects that are not face-like (e.g., no jack-

o'-lanterns, masks, and toys with a head) and 447 close-up images of human and macaque faces (~18% of the total images), which varied in identity and viewpoint, with or without headgear or personal protective equipment worn by humans in the laboratory. The goal of this stimulus selection was to have a set of faces and a set of non-faces that were distinct from each other in terms of high-level face-ness, which we confirmed using a computational object recognition model as a proxy for perceptual ratings (fig. S1).

Data analysis

Firing rates

We defined the neural response as the spike rate in the 100-ms time window starting at a latency of 50 to 100 ms after image onset. The exact latency of the response window was determined for each site individually, by calculating the image-level response reliability at each of the 51 latencies between 50 and 100 ms and picking the latency that maximized that reliability. Firing rates were trial-averaged per image, resulting in one response vector per neural site. For the acute recordings, the images were randomly divided into batches of 255 images, which were presented sequentially to the monkey in separate runs. For these sessions, run differences in median responses were equalized to remove slow trends in responsiveness that were unrelated to the stimuli. To include only visually driven, selective neural sites for further analysis, an a priori response reliability criterion of >0.4 was used. This removed 233 sites (15 from acute recordings and 218 from chronic arrays) that were mostly unresponsive neurons, dead channels, or channels in white matter. This yielded 449 sites (84 single and 365 multiunit) from central IT recordings and 57 sites (2 single and 55 multiunit) from AL recordings.

Response reliability

The firing-rate reliability was determined per neural site. First, for each image, the number of repeated presentations (trials) was randomly split in half. Next, the responses were trial averaged to create two response vectors, one per half of the trials. These two split-half response vectors were then correlated, and the procedure was repeated for 100 random splits to compute an average correlation r . The reliability ρ was computed by applying the Spearman-Brown correction as follows

$$\rho = 2r/(1 + r)$$

Face selectivity

Face selectivity was quantified by computing the d' sensitivity index comparing trial-averaged responses to faces and non-faces

$$d' = (\mu_F - \mu_{NF}) / \sqrt{[(\sigma_F^2 + \sigma_{NF}^2)/2]}$$

where μ_F and μ_{NF} are the across-stimulus averages of the trial-averaged responses to faces and non-faces, and σ_F and σ_{NF} are the across-stimulus SDs. This face d' value quantifies how much higher (positive d') or lower (negative d') the response to a face is expected to be compared to a non-face, in SD units.

Dynamic range

The dynamic range for faces was quantified by first identifying the "best" and "worst" face (highest and lowest response, respectively) using even trials, and then computing the normalized difference in response using the held-out odd trials

$$DR_F = (R_{\text{best F}} - R_{\text{worst F}}) / (R_{\text{max}} - R_{\text{min}})$$

where $R_{\text{best F}}$ and $R_{\text{worst F}}$ are the odd-trial-averaged responses to the best and worst face, and R_{max} and R_{min} are the maximum and minimum odd-trial-averaged responses. The dynamic range for non-faces was computed analogously.

Explained variance

To assess how accurately a model can predict face selectivity (face d'), we calculated the coefficient of determination R^2 , which quantifies the proportion of the variation in the observed face d' values that is explained by the predicted face d' values

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

where y_i are the observed values, \bar{y} is the mean of the observed values, and \hat{y}_i are the predicted values. Note that R^2 will be negative when the observed values y_i deviate more from the predicted values \hat{y}_i than from their own mean \bar{y} .

Generalization index

We computed a GI that quantifies how close a model's out-of-domain prediction accuracy (r_{OOD} ; e.g., on faces for the non-face encoding model) is to the within-domain prediction accuracy for the same images (r_{ID} ; e.g., on faces for the face encoding model)

$$\text{GI} = 1 - (r_{\text{ID}} - r_{\text{OOD}}) / (|r_{\text{ID}}| + |r_{\text{OOD}}|)$$

where prediction accuracy r was computed as Pearson's correlation coefficient between observed and predicted image responses.

Statistical analysis

Unless indicated otherwise, P values were calculated using permutation tests, based on 10,000 iterations. For R^2 and correlations, which calculate the correspondence between two variables, permutation testing was performed by randomly shuffling one of the two variables. For the paired difference between two correlations, the condition labels were randomly shuffled for each pair of observations. Ninety-five percent CIs were calculated using the bias-corrected accelerated bootstrap, based on 10,000 iterations.

The main analysis pipeline (Figs. 3 and 4) was first established using an independent pilot dataset with a smaller number of stimuli, before using the data reported here. Furthermore, all main results were cross-validated across independent train and test splits of the data (neural sites or stimuli, depending on the analysis; see below).

Models

Predicting face selectivity from non-face response profiles

A linear support vector regression model was fit to predict face d' values from response profiles to non-face objects (using the MATLAB 2020a function *fitrlinear*, with the SpARSA solver and default regularization). The responses of each neural site were first normalized (z -scored) using the mean and SD of responses to non-face objects only. The prediction accuracy was evaluated on out-of-fold predictions using leave-one-session/array-out cross-validation: The test partitions were defined as either all sites from the same array (chronic recordings) or all sites from the same session (acute recordings). This ensured that no simultaneously recorded data were ever split over the training and test partitions.

Color and shape properties

For each image, the following properties were computed from the non-background pixels: elongation, spikiness, circularity, and $Lu'v'$ color coordinates. Object elongation was defined on the basis of the minimum Feret diameter F_{min} and the maximum Feret diameter

F_{max} , as follows: $1 - F_{\text{min}}/F_{\text{max}}$. Spikiness was defined on the basis of the object area A_{obj} and the area of the convex hull of the object A_{hull} , as follows: $1 - A_{\text{obj}}/A_{\text{hull}}$. Circularity was defined using the object area and the object perimeter P_{obj} , as follows: $(4A_{\text{obj}}\pi)/(P_{\text{obj}}^2)$. $Lu'v'$ color coordinates were computed assuming standard RGB (red, green, blue).

DNN encoding model

The DNN encoding models were based on convolutional neural networks, used for extracting lower to higher-level image attributes, or DNN features, and a linear mapping between these DNN features and neural responses.

We used several DNNs as a base model for fitting encoding models. The first neural network had the architecture named "Inception" (29) and was trained on the ImageNet dataset (31) to classify images into 1000 object categories. We used the pretrained version of Inception that comes with the MATLAB 2020a Deep Learning Toolbox. Fourteen separate encoding models were created from the Inception network, each based on a subsequent processing step (layer) in the hierarchy: the input layer (pixels), the outputs of the first three convolutional layers, the outputs of each of the nine inception modules, and the output of the final fully connected layer. We refer to each of these encoding models by the name of the processing step (layer) that they were based on. The second, third, and fourth neural networks had the AlexNet architecture (30) and were pretrained in our laboratory on ImageNet, 365-way scene classification (32), or 8631-way face identity classification (33), respectively. For each of the AlexNet-based DNNs, we separately created encoding models for each convolutional, max pool, and fully connected layer.

To fit an encoding model based on DNN layer activations, outputs of a layer were normalized per channel using the SD and mean across all 1379 images (and across locations for pixels and convolutional layers). Next, the dimensionality of the outputs was reduced by applying principal component analysis using all images. Last, a linear support vector regression model was fit to predict neural responses from the principal components of the normalized DNN activations (using the MATLAB 2020a function *fitrlinear*, with the SpARSA solver and regularization parameter lambda set to 0.01; before fitting, the predictors were centered on the mean of the training fold and the responses were centered and standardized using the mean and SD of the training fold). Performance was evaluated on out-of-fold predicted responses concatenated from all test folds. For encoding models fit only on non-faces/faces, we used 10-fold cross-validation for the non-face/face images. In this case, the predicted responses for images that were not included in any of the training folds were computed as the average of the out-of-fold predictions. To compute predicted face d' values for the models, we calculated face d' using out-of-fold predicted responses.

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 to S9
References

REFERENCES AND NOTES

1. D. Y. Tsao, W. A. Freiwald, R. B. H. Tootell, M. S. Livingstone, A cortical region consisting entirely of face-selective cells. *Science* **311**, 670–674 (2006).

2. S. Moeller, W. A. Freiwald, D. Y. Tsao, Patches with links: A unified system for processing faces in the macaque temporal lobe. *Science* **320**, 1355–1359 (2008).
3. N. Kanwisher, J. McDermott, M. M. Chun, The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**, 4302–4311 (1997).
4. D. Y. Tsao, W. A. Freiwald, T. A. Knutsen, J. B. Mandeville, R. B. H. Tootell, Faces and objects in macaque cerebral cortex. *Nat. Neurosci.* **6**, 989–995 (2003).
5. N. Kanwisher, Domain specificity in face perception. *Nat. Neurosci.* **3**, 759–763 (2000).
6. E. McKone, N. Kanwisher, B. C. Duchaine, Can generic expertise explain special processing for faces? *Trends Cogn. Sci.* **11**, 8–15 (2007).
7. D. Y. Tsao, M. S. Livingstone, Mechanisms of face perception. *Annu. Rev. Neurosci.* **31**, 411–437 (2008).
8. D. I. Perrett, E. T. Rolls, W. Caan, Visual neurones responsive to faces in the monkey temporal cortex. *Exp. Brain Res.* **47**, 329–342 (1982).
9. C. Bruce, R. Desimone, C. G. Gross, Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J. Neurophysiol.* **46**, 369–384 (1981).
10. R. Desimone, T. D. Albright, C. G. Gross, C. Bruce, Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* **4**, 2051–2062 (1984).
11. W. A. Freiwald, D. Y. Tsao, M. S. Livingstone, A face feature space in the macaque temporal lobe. *Nat. Neurosci.* **12**, 1187–1196 (2009).
12. L. Chang, D. Y. Tsao, The code for facial identity in the primate brain. *Cell* **169**, 1013–1028 (2017).
13. E. B. Issa, J. J. DiCarlo, Precedence of the eye region in neural processing of faces. *J. Neurosci.* **32**, 16666–16682 (2012).
14. D. A. Leopold, I. V. Bondar, M. A. Giese, Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature* **442**, 572–575 (2006).
15. I. Yildirim, M. Belledonne, W. Freiwald, J. Tenenbaum, Efficient inverse graphics in biological face processing. *Sci. Adv.* **6**, ead5979 (2020).
16. I. Higgins, L. Chang, V. Langston, D. Hassabis, C. Summerfield, D. Tsao, M. Botvinick, Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nat. Commun.* **12**, 6456 (2021).
17. L. Chang, B. Egger, T. Vetter, D. Y. Tsao, Explaining face representation in the primate brain using different computational models. *Curr. Biol.* **31**, 2785–2795.e4 (2021).
18. P. Bao, L. She, M. McGill, D. Y. Tsao, A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
19. T. Konkle, A. Caramazza, Tripartite organization of the ventral stream by animacy and object size. *J. Neurosci.* **33**, 10235–10242 (2013).
20. F. R. Doshi, T. Konkle, Cortical topographic motifs emerge in a self-organized map of object space. *Sci. Adv.* **9**, eade8187 (2023).
21. M. Mur, D. A. Ruff, J. Bodurka, P. De Weerd, P. A. Bandettini, N. Kriegeskorte, Categorical, yet graded—single-image activation profiles of human category-selective cortical regions. *J. Neurosci.* **32**, 8649–8662 (2012).
22. S. Yamane, S. Kaji, K. Kawano, What facial features activate face neurons in the inferotemporal cortex of the monkey? *Exp. Brain Res.* **73**, 209–214 (1988).
23. B. Jagadeesh, Recognizing Grandmother. *Nat. Neurosci.* **12**, 1083–1085 (2009).
24. H. Hong, D. L. K. Yamins, N. J. Majaj, J. J. DiCarlo, Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* **19**, 613–622 (2016).
25. S. Salehi, M.-R. A. Dehaqani, H. Esteky, Low dimensional representation of face space by face-selective inferior temporal neurons. *Eur. J. Neurosci.* **45**, 1268–1278 (2017).
26. R. Kiani, H. Esteky, K. Mirpour, K. Tanaka, Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *J. Neurophysiol.* **97**, 4296–4309 (2007).
27. C. Baldassi, A. Alemi-Neissi, M. Pagan, J. J. DiCarlo, R. Zecchina, D. Zoccolan, Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons. *PLOS Comput. Biol.* **9**, e1003167 (2013).
28. S. Baek, M. Song, J. Jang, G. Kim, S.-B. Paik, Face detection in untrained deep neural networks. *Nat. Commun.* **12**, 7328 (2021).
29. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 1–9 (2015).
30. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.*, 1–9 (2012).
31. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
32. B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 1452–1464 (2018).
33. Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, VGGFace2: A dataset for recognising faces across pose and age, in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, Xi'an, China, 15 to 19 May 2018 (IEEE, 2018), pp. 67–74.
34. S. M. Khaligh-Razavi, N. Kriegeskorte, Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLOS Comput. Biol.* **10**, e1003915 (2014).
35. D. L. K. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, J. J. DiCarlo, Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
36. N. Kriegeskorte, M. Mur, P. Bandettini, Representational similarity analysis-connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 1–28 (2008).
37. S. Grossman, G. Gaziv, E. M. Yeagle, M. Harel, P. Mégevand, D. M. Groppe, S. Khuvis, J. L. Herrero, M. Irani, A. D. Mehta, R. Malach, Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nat. Commun.* **10**, 4934 (2019).
38. N. A. R. Murty, P. Bashivan, A. Abate, J. J. DiCarlo, N. Kanwisher, Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nat. Commun.* **12**, 5540 (2021).
39. I. Kalfas, K. Vinken, R. Vogels, Representations of regular and irregular shapes by deep convolutional neural networks, monkey inferotemporal neurons and human judgments. *PLOS Comput. Biol.* **14**, e1006557 (2018).
40. R. Vogels, Sources of adaptation of inferior temporal cortical responses. *Cortex* **80**, 185–195 (2016).
41. B.-E. Verhoef, G. Kayaert, E. Franko, J. Vangeneugden, R. Vogels, Stimulus similarity-contingent neural adaptation can be time and cortical area dependent. *J. Neurosci.* **28**, 10631–10640 (2008).
42. W. De Baene, R. Vogels, Effects of adaptation on the stimulus selectivity of macaque inferior temporal spiking activity and local field potentials. *Cereb. Cortex* **20**, 2145–2165 (2010).
43. K. Vinken, X. Boix, G. Kreiman, Incorporating intrinsic suppression in deep neural networks captures dynamics of adaptation in neurophysiology and perception. *Sci. Adv.* **6**, eabd4205 (2020).
44. S. Sadagopan, W. Zarco, W. A. Freiwald, A causal relationship between face-patch activity and face-detection behavior. *eLife* **6**, e18558 (2017).
45. S. Moeller, T. Crapse, L. Chang, D. Y. Tsao, The effect of face patch microstimulation on perception of faces and objects. *Nat. Neurosci.* **20**, 743–752 (2017).
46. G. Schalk, C. Kapeller, C. Guger, H. Ogawa, S. Hiroshima, R. Lafer-Sousa, Z. M. Saygin, K. Kamada, N. Kanwisher, Facephenes and rainbows: Causal evidence for functional and anatomical specificity of face and color processing in the human brain. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 12285–12290 (2017).
47. S. Kumar, E. Mergan, R. Vogels, It is not just the category: Behavioral effects of fMRI-guided electrical microstimulation result from a complex interplay of factors. *Cereb. Cortex Commun.* **3**, tgac010 (2022).
48. R. Azadi, S. Bohn, E. Lopez, R. Lafer-Sousa, K. Wang, M. A. G. Eldridge, A. Afraz, Image-dependence of the detectability of optogenetic stimulation in macaque inferotemporal cortex. *Curr. Biol.* **33**, 581–588.e4 (2023).
49. B. Zhou, D. Bau, A. Oliva, A. Torralba, Interpreting deep visual representations via network dissection. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 2131–2145 (2019).
50. B. Long, C. Yu, T. Konkle, Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E9015–E9024 (2018).
51. A. J. O'Toole, C. D. Castillo, Face recognition by humans and machines: Three fundamental advances from deep learning. *Annu. Rev. Vis. Sci.* **7**, 543–570 (2021).
52. C. J. Parde, Y. I. Colón, M. Q. Hill, C. D. Castillo, P. Dhar, A. J. O'Toole, Closing the gap between single-unit and neural population codes: Insights from deep learning in face recognition. *J. Vis.* **21**, 1–14 (2021).
53. N. Kriegeskorte, W. K. Simmons, P. S. F. Bellgowan, C. I. Baker, Circular analysis in systems neuroscience: The dangers of double dipping. *Nat. Neurosci.* **12**, 535–540 (2009).
54. N. Kanwisher, Functional specificity in the human brain: A window into the functional architecture of the mind. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 11163–11170 (2010).
55. E. McKone, N. Kanwisher, 17 Does the human brain process objects of expertise like faces? A review of the evidence, in *From Monkey Brain to Human Brain: A Fyssen Foundation Symposium* (MIT Press, 2005), p. 339.
56. T. Konkle, G. A. Alvarez, A self-supervised domain-general learning framework for human ventral stream representation. *Nat. Commun.* **13**, 491 (2022).
57. M. J. Arcaro, M. S. Livingstone, On the relationship between maps and domains in inferotemporal cortex. *Nat. Rev. Neurosci.* **22**, 573–583 (2021).
58. H. P. Op De Beeck, J. A. Deutsch, W. Vanduffel, N. G. Kanwisher, J. J. DiCarlo, A stable topography of selectivity for unfamiliar shape classes in monkey inferior temporal cortex. *Cereb. Cortex* **18**, 1676–1694 (2008).

59. X. Yue, S. Robert, L. G. Ungerleider, Curvature processing in human visual cortical areas. *Neuroimage* **222**, 117295 (2020).
60. A. V. Jagadeesh, J. L. Gardner, Texture-like representation of objects in human visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2115302119 (2022).
61. R. Wang, D. Janini, T. Konkle, Mid-level feature differences support early animacy and object size distinctions: Evidence from electroencephalography decoding. *J. Cogn. Neurosci.* **34**, 1670–1680 (2022).
62. M. J. Arcaro, M. S. Livingstone, A hierarchical, retinotopic proto-organization of the primate visual system at birth. *elife* **6**, e26196 (2017).
63. D. B. T. McMahon, I. V. Bondar, O. A. T. Afuwape, D. C. Ide, D. A. Leopold, One month in the life of a neuron: Longitudinal single-unit electrophysiology in the monkey visual system. *J. Neurophysiol.* **112**, 1748–1762 (2014).
64. M. J. Arcaro, T. Mautz, V. K. Berezovskii, M. S. Livingstone, Anatomical correlates of face patches in macaque inferotemporal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 32667–32678 (2020).
65. C. R. Ponce, W. Xiao, P. F. Schade, T. S. Hartmann, G. Kreiman, M. S. Livingstone, Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell* **177**, 999–1009.e10 (2019).
66. T. Konkle, T. F. Brady, G. A. Alvarez, A. Oliva, Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *J. Exp. Psychol. Gen.* **139**, 558–578 (2010).
67. R. Rajalingham, E. B. Issa, P. Bashivan, K. Kar, K. Schmidt, J. J. DiCarlo, Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* **38**, 7255–7269 (2018).
68. D. Y. Tsao, S. Moeller, W. A. Freiwald, Comparing face patch systems in macaques and humans. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 19514–19519 (2008).
69. R. Lafer-Sousa, B. R. Conway, N. G. Kanwisher, Color-biased regions of the ventral visual pathway lie between face- and place-selective regions in humans, as in macaques. *J. Neurosci.* **36**, 1682–1697 (2016).
70. K. N. Kay, T. Naselaris, R. J. Prenger, J. L. Gallant, Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).

Acknowledgments

Funding: This work was supported by the Research Foundation Flanders, Belgium – postdoctoral fellowship (to K.V.), Alice and Joseph Brooks Fund Postdoctoral Fellow (K.V.), NIH grant R01MH116858-03 (to M.S.L.), and NSF CAREER: BCS-1942438 (T.K.). **Author contributions:** Conceptualization: K.V., T.K., and M.S.L. Data curation: K.V. Formal analysis: K.V. Methodology: K.V. and J.S.P. Investigation: M.S.L. Visualization: K.V. Supervision: T.K. and M.S.L. Writing—original draft: K.V. Writing—review and editing: K.V., J.S.P., T.K., and M.S.L. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are available in our Open Science Framework repository at <https://osf.io/bk67z/> or our Dataverse repository at <https://doi.org/10.7910/DVN/GF5ZK4>.

Submitted 7 December 2022

Accepted 27 July 2023

Published 30 August 2023

10.1126/sciadv.adg1736

The neural code for “face cells” is not face-specific

Kasper Vinken, Jacob S. Prince, Talia Konkle, and Margaret S. Livingstone

Sci. Adv. **9** (35), eadg1736. DOI: 10.1126/sciadv.adg1736

View the article online

<https://www.science.org/doi/10.1126/sciadv.adg1736>

Permissions

<https://www.science.org/help/reprints-and-permissions>

Use of this article is subject to the [Terms of service](#)

Science Advances (ISSN 2375-2548) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Advances* is a registered trademark of AAAS.

Copyright © 2023 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).