



Face neurons encode nonsemantic features

Alexandra Bardon^{a,1}, Will Xiao^{b,c,1}, Carlos R. Ponce^b, Margaret S. Livingstone^{b,2}, and Gabriel Kreiman^{d,e,2}

Contributed by Margaret S. Livingstone; received October 13, 2021; accepted February 17, 2022; reviewed by Marlene Behrmann and Sabine Kastner

The primate inferior temporal cortex contains neurons that respond more strongly to faces than to other objects. Termed “face neurons,” these neurons are thought to be selective for faces as a semantic category. However, face neurons also partly respond to clocks, fruits, and single eyes, raising the question of whether face neurons are better described as selective for visual features related to faces but dissociable from them. We used a recently described algorithm, XDream, to evolve stimuli that strongly activated face neurons. XDream leverages a generative neural network that is not limited to realistic objects. Human participants assessed images evolved for face neurons and for nonface neurons and natural images depicting faces, cars, fruits, etc. Evolved images were consistently judged to be distinct from real faces. Images evolved for face neurons were rated as slightly more similar to faces than images evolved for nonface neurons. There was a correlation among natural images between face neuron activity and subjective “faceness” ratings, but this relationship did not hold for face neuron–evolved images, which triggered high activity but were rated low in faceness. Our results suggest that so-called face neurons are better described as tuned to visual features rather than semantic categories.

face neurons | semantic tuning | neural coding | visual cortex

Ventral stream neurons fire selectively to visual features, such as optimally oriented bars, particular colors, specific curvatures, or certain object categories. A famous type of category-selective neurons is the “face neuron.” Recordings in monkey inferior temporal cortex (IT) by Gross and colleagues (1–3) revealed neurons that responded more strongly to images of faces than to other objects, such as hands and eyeless faces. Face-selective neurons, whose tuning properties are stable for at least months (4), tend to cluster within millimeter-wide patches on the cortex (5). Face-selective neural signals have also been found in humans by intracranial field potential recordings (6); subsequently, by noninvasive measurements (e.g., ref. 7); and recently, with unit recordings (8).

A paragon of category selectivity in the visual cortex, face neurons are as extensively studied as they attract controversy in the interpretation of their tuning properties. The debate centers on whether category-selective neural signals truly represent a semantic category (9–11) or whether they represent visual features that correlate with, but are dissociable from, any object category. The semantic view is often referred to as “word models.” Word models, being essentially ambiguous, are difficult to formally define and to falsify (12). A semantic category–selective neuron should exclusively respond to the namesake category. However, IT neurons ostensibly selective for nonface categories in fact respond to typical features even when they are separated from the preferred category (13–16). Face-selective neurons are known to respond to round objects and other nonface objects in a graded manner (5), and face-selective voxels in functional magnetic resonance imaging (fMRI) show significant responses to face pareidolia images (17). Nevertheless, there is scant evidence that nonface stimuli can drive responses as strong as faces in face-selective neurons.

A diagnostic finding would be to identify nonface stimuli that strongly activate face neurons (i.e., counterexamples to the word model). Finding such stimuli would require efficient exploration of the vast space of images within limited neuron recording time. Here, we recorded spiking activity from both face-selective and nonface-selective neurons and used the XDream method to find strongly activating images (18, 19). XDream uses a broad image prior, does not depend on a predictive model of neuronal responses, and rather uses adaptive search with closed-loop neuron recording. In face-selective neurons, XDream led to images that elicited comparable responses with faces. To evaluate semantic tuning to faces, we quantified human perception to define how face-like images were using a series of six experiments ranging from open-ended (subjects entered one-word descriptions) to more structured (subjects answered whether or not an image looked like a face). Our results show that subjects did not perceive evolved images as faces. Yet, evolved images tailored to face-selective neurons were perceived as more face-like than nonface object images and evolved images tailored to nonface-selective neurons. Moreover, among natural images, there was a significant correlation between subjects’ ratings for “faceness”

Significance

Face neurons, which fire more strongly in response to images of faces than to other objects, are a paradigmatic example of object selectivity in the visual cortex. We asked whether such neurons represent the semantic concept of faces or, rather, visual features that are present in faces but do not necessarily count as a face. We created synthetic stimuli that strongly activated face neurons and showed that these stimuli were perceived as clearly distinct from real faces. At the same time, these synthetic stimuli were slightly more often associated with faces than other objects were. These results suggest that so-called face neurons do not represent a semantic category but, rather, represent visual features that correlate with faces.

Author affiliations: ^aDivision of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125 ^bDepartment of Neurobiology, Harvard Medical School, Boston, MA 02115; ^cDepartment of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02134; ^dBoston Children’s Hospital, Harvard Medical School, Boston, MA 02115; and ^eCenter for Brains, Minds and Machines, Cambridge, MA 02115

Author contributions: A.B., W.X., C.R.P., M.S.L., and G.K. designed research; A.B., W.X., C.R.P., and M.S.L. performed research; A.B., W.X., and G.K. analyzed data; and A.B., W.X., C.R.P., M.S.L., and G.K. wrote the paper.

Reviewers: M.B., Carnegie Mellon University; and S.K., Princeton University.

The authors declare no competing interest.

Copyright © 2022 the Author(s). Published by PNAS. This article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/).

¹A.B. and W.X. contributed equally to this work.

²To whom correspondence may be addressed. Email: margaret.livingstone@hms.harvard.edu or gabriel.kreiman@childrens.harvard.edu.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2118705119/-DCSupplemental>.

Published April 4, 2022.

and face neuron responses. This correlation was present even among nonface images. However, evolved images received relatively low faceness ratings yet evoked some of the highest firing rates from face neurons. These results show that face neurons are not tuned to the semantic concept of faces but, instead, respond to visual attributes associated with faces.

Results

We synthesized preferred stimuli for neurons in monkey ventral visual cortex using the XDream algorithm (18, 19). Evolved images strongly activated both face neurons and nonface neurons (Fig. 1 *D* and *E*). For face neurons, evolved images were as effective stimuli as face images (Fig. 1*D*). We conducted a series of six behavioral experiments to evaluate how human subjects perceived these evolved images (Fig. 1 *A–C*). For comparison, we included natural images depicting faces and nonface objects (book, butterfly, car, chair, cloud, dog, fruit, tree, and wheel) and abstract drawings. Images included in the main analyses are shown in *SI Appendix, Fig. S1*. Numbers in the text refer to mean \pm SD unless otherwise noted.

For each neuron, we used its mean background-subtracted firing rate to faces (r_f) and to nonface objects (r_{nf}) to define a face selectivity index (FSI) = $(r_f - r_{nf}) / (r_f + r_{nf})$ that summarizes the neuron's selectivity for faces over nonface objects. We operationally defined face-selective neurons (face neurons for brevity) to be neurons with FSI greater than 0.5 recorded in central inferior temporal cortex (CIT) to exclude recordings from the posterior lateral (PL) face patch in posterior IT (PIT) (20). PL neurons often have high FSI but are known to respond to single eyes instead of whole faces (21). We defined nonface neurons to be neurons with an FSI less than zero recorded from any area, although all but one included neuron were recorded in CIT. Based on these criteria, we included in the main analyses 39 images evolved from face neurons and 47 images evolved from nonface neurons. Similar results were observed using more lenient criteria admitting more neurons (*SI Appendix, Figs. S6 and S7*).

Experiment 1: One-Word Description. In the first experiment, we sought to use an open-ended approach. We asked subjects to use one word to describe each image [i.e., basic-level description (22)]. Fig. 2 *A–C* shows example images and their descriptions. When presented with a photo of a child, subjects gave consistent descriptions, such as “child,” “girl,” and “kid”; when presented

with an illustration of clouds, subjects described it with words such as “sky” and “cloud.” For five chair photos, subjects consistently answered “chair” (86%) (Fig. 2*D*). In contrast, for 10 face photos, subjects used a variety of words like “woman” (21%), “man” (15%), or “girl” (12%) (Fig. 2 *D* and *E*), thereby using more specific words than “face” as the basic-level description. Notwithstanding the sometimes varied answers, all natural object images (book, butterfly, car, chair, cloud, dog, face, fruit, tree, or wheel) were adequately described (*SI Appendix, Fig. S2 D and E*), indicating that subjects followed the task directions. To assess how subjects described images that did not afford a single obvious label, we included abstract drawings. Abstract drawings received a variety of descriptions, including “painting” (13%), “art” (12%), and “flower” (7%), among others (Fig. 2 *D* and *E*, third column; also, example descriptions of a single image are in *SI Appendix, Fig. S2A*).

How did subjects describe images evolved by visual neurons? An example image and its responses are shown in Fig. 2*C*. The neuron that gave rise to this image was recorded in the middle lateral (20) face patch in CIT and showed high selectivity to faces (FSI = 1.18; background-subtracted firing rate response to 10 human faces: 26 ± 10 spikes/s; 20 monkey faces: 18 ± 14 spikes/s; 22 nonface objects: -4 ± 5 spikes/s). This evolved image elicited a stronger neuronal response than face images did (20 last-generation evolved images: 64 ± 19 spikes/s). The subjects described the evolved image using words such as “face” (19%), “monkey” (16%), “painting” (5%), and “art” (5%). For all images evolved from face neurons, top descriptions included “monkey” (14%), “dog” (5%), “art” (5%), “animal” (4%), and “face” (4%) (Fig. 2 *D* and *E*, fourth column). For all images evolved from nonface neurons, similar words were used, including “bird” (7%) and “dog” (5%), but words like “monkey” and “face” were less frequently used (3 vs. 14% and 2 vs. 4%, respectively) (Fig. 2 *D* and *E*, fifth column). The words used to describe evolved images were more heterogeneous than the words used to describe control object images but were comparable with the diversity of words used to describe abstract drawings. This is indicated by the frequency of the top word. The top word accounted for about 14 and 7% of descriptions of images evolved from face and nonface neurons, respectively, compared with 13% for abstract drawings, 21% for face photos, and 86% for chair photos.

To better quantify and summarize the descriptions, we calculated the similarity from descriptions to category labels using

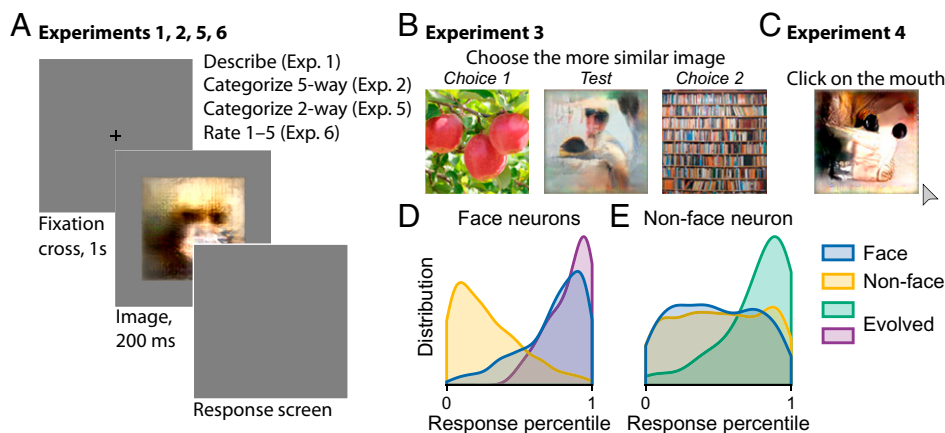


Fig. 1. Overview of the study. (*A–C*) Schematics of experiments. (*A*) In experiments 1, 2, 5, and 6, each trial began with a center cross shown for 1 s followed by an image shown for 200 ms and then, the response screen. (*B*) In experiment 3, three images were presented together in each trial. The subject was instructed to select the side that was more similar to the center image. (*C*) In experiment 4, images were presented individually, and the subject was instructed to “click on the mouth.” (*D* and *E*) Distributions of normalized neuronal responses to face, nonface, and evolved images for face-selective (*D*) and nonface-selective neurons (*E*). The distribution is over images and neurons.

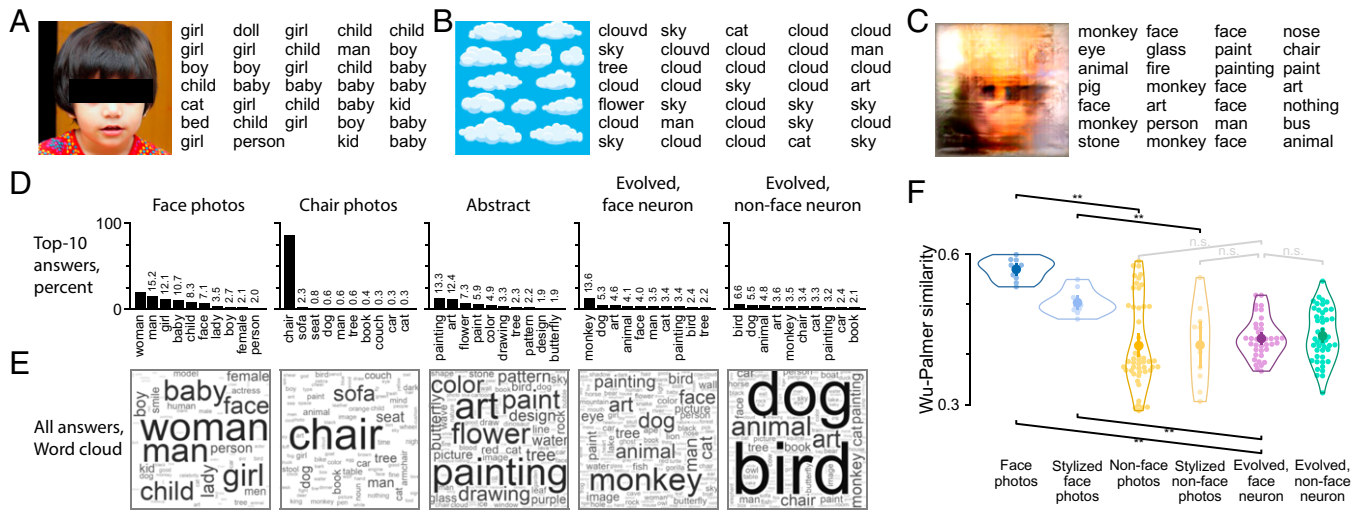


Fig. 2. Experiment 1: one-word (basic-level) description. (A–C) Three example images and example responses are shown. (D) The top 10 descriptions along with frequency are shown for each image group. Frequencies lower than 20% are indicated by numbers above bars. (E) Description frequency for each image group is visualized using a word cloud. (F) WP semantic similarity was calculated between subject-provided descriptions. The swarm plot shows WP similarity between the descriptions of any image and descriptions of face photos (each face photo was compared with the other 9 face photos; other images were compared with all 10 face photos). Each small point represents an image. The horizontal spread within each group is for visualization only. Open contour indicates the kernel density estimate for the points. The inner thick bar and point indicate data mean and bootstrap 95% CI of the mean. Only indicated pairs were tested. n.s., not significant. $**P < 0.01$, one-tailed permutation test with test direction indicated by the slope of the square bracket, false discovery rate (FDR) corrected across seven tests.

the Wu–Palmer (WP) semantic similarity measure (23), a metric ranging from zero to one based on how close two words are in the WordNet hierarchy (24). For example, the first row in *SI Appendix, Fig. S2F* is the WP similarity of each image to the word “book,” averaged over all descriptions of each image. As expected, descriptions of book images had high similarity to the word book. Descriptions of other object images also mostly matched their respective category names, resulting in a strong diagonal in *SI Appendix, Fig. S2F*. One exception is the category of face photos. Due to the structure of the WordNet hierarchy, WP similarity is lower (0.31) between face and woman (most common description of face images) than between woman and for example, “butterfly” (0.57), “tree” (0.63), and “dog” (0.67). *Materials and Methods* describes in more detail how WP similarity is defined and how it was calculated for these example word pairs.

To avoid this problem in comparing against category labels, we compared subject-provided descriptions with each other, averaging over all word pairs between each image pair (*SI Appendix, Fig. S2H*). Consistent with the analyses so far, natural images in the same category received similar descriptions, visible as diagonal groups indicating high similarity in *SI Appendix, Fig. S2H, Upper Left*. For instance, descriptions of book images had an average WP similarity of 0.78 ± 0.07 to other book image descriptions compared with a similarity of 0.38 ± 0.05 to descriptions of all other images.

We focused on whether descriptions of images evolved by face neurons were similar to descriptions of face photos. Fig. 2F shows the similarity of descriptions per image to face photo descriptions averaged over face photos (each face photo was compared with the other 9 face photos; other images were compared with all 10 face photos). By this measure, face photo descriptions were most similar to other face photo descriptions (WP similarity 0.57 ± 0.02). As expected, nonface object descriptions were significantly less similar to face photo descriptions (0.42 ± 0.09 ; $P < 10^{-5}$, one-tailed permutation test). Relative to face photo descriptions, descriptions of images evolved from face neurons were significantly less similar (0.43 ± 0.04) than descriptions of other face photos ($P < 10^{-5}$) while comparable with descriptions

of nonface object images (0.42 ± 0.09 ; $P = 0.11$) and descriptions of images evolved from nonface neurons (0.44 ± 0.05 ; $P = 0.32$).

To assess whether the quantification of semantic similarity depended on WordNet, we considered an alternative word similarity measure based on word embeddings, LexVec (25). Conclusions were similar using this alternative metric, except that images evolved from face neurons were described as slightly more similar to faces than nonface objects or images evolved from nonface neurons (0.31 ± 0.06 vs. 0.20 ± 0.08 and 0.27 ± 0.06 ; $P = 0.001$ and $P < 10^{-5}$, respectively) (*SI Appendix, Fig. S2L*; also, *SI Appendix, Fig. S2 F–L* compares the WP and LexVec metrics).

Evolved images are limited by the underlying image generator, which can approximately depict objects but does not produce photorealistic images (18). To account for this “stylistic” constraint, we generated “stylized” images from face and object images as additional controls. A stylized image is a natural image converted to a synthetic image that the generator could produce and that was optimized to resemble the original image. All stylized images are shown in *SI Appendix, Fig. S1*. Stylized faces were most commonly described as face (30%), man (20%), and woman (13%) (*SI Appendix, Fig. S2 D and E*). Stylized nonface images were also mostly described as the correct object, as indicated by the near diagonal in *SI Appendix, Fig. S2 F and G* (stylized nonface images include one abstract and one for each category on the y axis in that order). Thus, the image generator style, although nonrealistic, retained enough information to allow for reliable object identification. Compared with face photo descriptions, stylized face image descriptions had WP similarity of 0.50 ± 0.02 and LexVec similarity of 0.51 ± 0.04 , both significantly higher than for descriptions of face neuron evolved images (both $P < 10^{-5}$). Thus, generator style alone did not account for the low proportion of descriptions indicating faces in the evolved images.

In summary, evolved images were described by a variety of words in open-ended categorization. The most common descriptions of evolved images related to animals or art, but there was much less agreement in the description of evolved images among observers than for any natural image category. In terms of

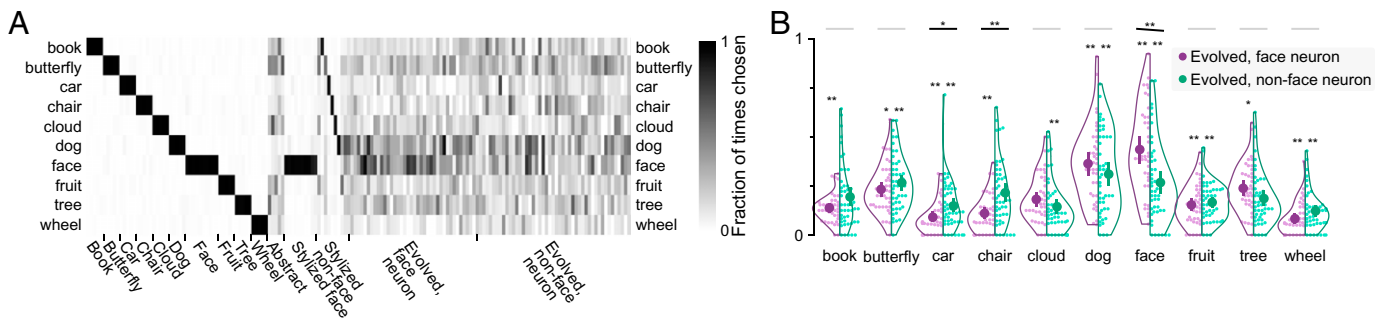


Fig. 3. Experiment 2: five-way categorization. (A) Subjects were presented with an image and were asked to choose among five category labels (Fig. 1A). The heat map shows the fraction of trials a label was chosen when it was an available option. Thus, the fraction ranges from zero to one in all cases. Each column corresponds to an image. Each row corresponds to a categorization option. (B) The swarm plot shows the fraction of trials each label was chosen (if available) separately for images evolved by face neurons (purple) and nonface neurons (green). Each point represents one evolved image. Open contour indicates the kernel density estimate. The inner thick bar and point indicate data mean and bootstrap 95%-CI of the mean. *On violin, $P < 0.05$, two-sided binomial test for difference from chance = 0.2, FDR corrected across 20 tests; **on violin, $P < 0.01$, two-sided binomial test for difference from chance = 0.2, FDR corrected across 20 tests; *on the black line, $P < 0.05$, **on the black line, $P < 0.01$, permutation test, FDR corrected across 10 tests, the test was one tailed for the face option (face neurons evolved greater than nonface neurons evolved) and two tailed otherwise.

similarity to face photo descriptions, descriptions of images evolved by face neurons were comparable with descriptions of nonface object images and images evolved from nonface neurons.

Experiment 2: Five-Way Categorization. Experiment 1 allowed for a wide range of subjective word choices. To more directly test the hypothesis that face neurons have a specific categorical preference for faces, we conducted a forced choice categorization task. We assessed 10 categories in total (the y axis in Fig. 3A), including face. In each trial, after the image was shown, five category options were presented, and subjects were asked to “choose the most appropriate category.” For nonface object images, the options always included the correct category, while the four other options were randomly composed from the remaining nine categories on a trial by trial basis. For abstract drawings and evolved images, all five options were randomly chosen from the 10 categories. We report in Fig. 3 the fraction of trials in which a category was chosen for an image normalized by the number of trials in which that category was available as an option for that image.

Natural images were almost always correctly classified (0.97 to 0.99 accuracy), evident as strong diagonal elements in Fig. 3A. For abstract drawings, choices were dispersed, with the most common choices being butterfly (0.45 ± 0.16) and cloud (0.33 ± 0.19). Stylized face images were almost always categorized as “face.” Stylized nonface images were also usually correctly categorized (in Fig. 3A, values near the diagonal that correspond to the stylized image categories were usually high).

Evolved images, like abstract drawings, received a wide range of categorization choices. Nevertheless, the choices were not random. For both groups of evolved images, the fraction of times a label was chosen was significantly different across label options (face-neuron-evolved images: $P = 1 \times 10^{-27}$; nonface-neuron-evolved images: $P = 7 \times 10^{-9}$, Kruskal–Wallis test), and several options were chosen significantly differently than expected by chance (asterisks for individual violins in Fig. 3B). Images evolved from face neurons were most commonly labeled as face (0.44 ± 0.22) or dog (0.36 ± 0.20). The same two categories were the top choices for nonface neuron evolved images, with lower frequency (dog: 0.31 ± 0.21 , face: 0.27 ± 0.21). Thus, both groups of evolved images were categorized as face less than half of the time when face was an available option and significantly less than face photos (0.97 ± 0.01 ; both $P < 10^{-5}$, one-tailed permutation test). The differences between face and nonface neuron evolved images were significant for the categories “face” ($P = 0.003$), “chair” ($P = 0.003$), and “car” ($P = 0.043$; one-sided test for face, two-sided test for other categories, FDR corrected across 10 tests). The category

face was chosen significantly more often for images evolved from face neurons than those from nonface neurons; the reverse was true for chair and car.

Experiment 3: Image Similarity. In experiment 3, we aimed to assess visual similarity without using words as in experiments 1 and 2. We asked subjects to “choose the more similar image” from one of two choice images (Fig. 1B). Choice images were drawn from the same natural images used in experiments 1 and 2. In each trial, choice images were randomly drawn from 2 of the same 10 categories as in experiment 2. We tested each evolved image at least once with each of 45 possible category pairings. Fig. 4A shows the fraction of evolved images for which subjects chose the category indicated on the y axis over the one on the x axis (thus, entries i, j and j, i sum to one) separately for evolved images from face (Fig. 4A, Left) and nonface (Fig. 4A, Right) neurons. For example, the entry in row 2, column 1 shows that for the majority (76%) of face neuron evolved images, subjects chose photos of butterflies as being more similar than books. If subjects did not prefer either category, selecting by chance should correspond to a value of 0.5 on average, indicated by white color. For images evolved from face neurons, subjects preferred the category dog significantly above chance for five of the nine possible comparisons. Subjects preferred face to a lesser degree, doing so significantly above chance for three of nine comparisons. For images evolved from nonface neurons, subjects preferred the categories dog and butterfly significantly above chance in two of nine comparisons.

To summarize the pairwise comparisons and to directly compare the two groups of evolved images, we calculated the preference for each category averaged over alternatives (Fig. 4B). For images evolved from face neurons, the top choices were dog ($70 \pm 16\%$), face ($68 \pm 18\%$), and butterfly ($54 \pm 17\%$). For images evolved from nonface neurons, the top choices were butterfly ($66 \pm 19\%$), dog ($65 \pm 16\%$), and face ($56 \pm 22\%$). Subjects chose the option face more often for images evolved from face neurons than nonface neurons ($P = 0.018$, one-tailed permutation test, FDR corrected across 10 tests). Butterfly also showed a statistically significant difference ($P = 0.018$, two-tailed permutation test) and was chosen less often for face neuron than nonface neuron evolved images.

These results show that face neuron evolved images were more visually similar to face photos than nonface neuron evolved images were. Nevertheless, both groups of evolved images were more similar to dog images than faces.

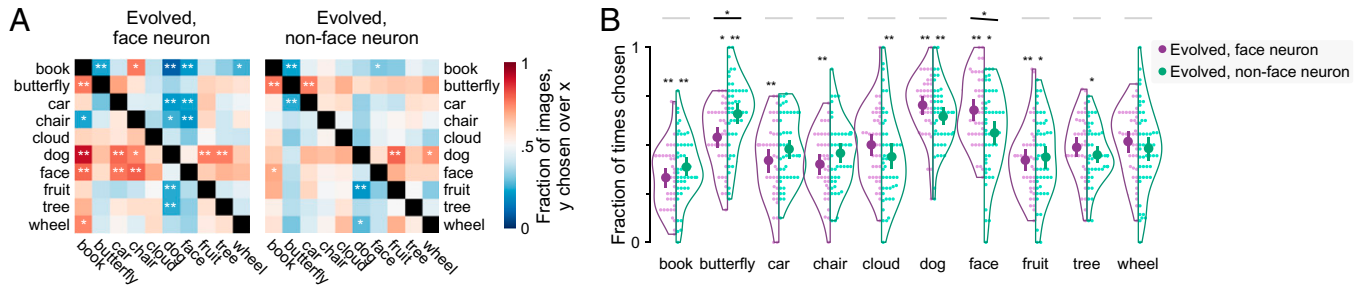


Fig. 4. Experiment 3: image similarity. (A) Subjects were presented with three images and asked to select whether the left or right image was more similar to the center one (Fig. 1B). Each evolved image was tested at least once for each option pair (10 choose 2 = 45 pairs). The heat map shows the fraction of images for which the y category was chosen over the x category for images evolved from face neurons (Left) or nonface neurons (Right). * $P < 0.05$, two-sided binomial test for difference from chance = 0.5, FDR corrected across 90 pairwise tests; ** $P < 0.01$, two-sided binomial test for difference from chance = 0.5, FDR corrected across 90 pairwise tests. (B) The swarm plot shows the fraction of trials a category was favored when it was an option (thus, possible values range from zero to one). Plot conventions follow those in Fig. 3B. *On violin, $P < 0.05$, **on violin, $P < 0.01$, two-sided binomial test for difference from chance = 0.5, FDR corrected across 20 tests; *on the black line, $P < 0.05$, permutation test, FDR corrected across 10 tests, the test was one tailed for the face option (face neurons evolved greater than nonface neurons evolved) and two tailed otherwise.

Experiments 4 and 5: Metrics Tailored to Measure Face Semblance. In two further experiments, we specifically measured whether and how much the evolved images resembled faces. First, we reasoned that if subjects perceived a face in any image, they should be able to locate specific features of the face. An intuitive assay is to ask the subjects to locate an “eye.” However, any small dark spot in an image could be interpreted as an eye. Additionally, the two eyes, if present, would add ambiguity to eye localization. Therefore, in experiment 4, we asked subjects to “click on the mouth (if unsure, make your best guess)” (Fig. 1C). The task instructions are ill defined for nonface objects, such as chairs. We reasoned that if subjects did not see a face in an image, they would click on inconsistent locations. Thus, we quantified the consistency of click locations across subjects using the entropy of the distribution of click locations. An entropy of zero indicates

that subjects always clicked on the same location (after binning), and higher entropy indicates more varied click locations.

Subjects located the mouth in face photos consistently and correctly (an example is shown in Fig. 5A, Upper Left). In comparison, clicks were dispersed, as predicted, in nonface object images (an example is shown in Fig. 5A, Lower Left). The entropy of clicks was significantly different across image groups (Fig. 5B) ($P = 7 \times 10^{-11}$, Kruskal–Wallis test). Ad hoc pairwise comparisons confirmed that clicks were more consistent for face photos (entropy = 0.9 ± 0.3) than for nonface object images (2.6 ± 0.9 ; $P < 10^{-5}$, one-tailed permutation test, FDR corrected across seven tests). Clicks were also more consistent for face photos and stylized face images (1.4 ± 0.3) compared with face neuron evolved images (2.2 ± 0.5 ; both $P < 10^{-5}$). Clicks were slightly more consistent for images evolved from face neurons than for

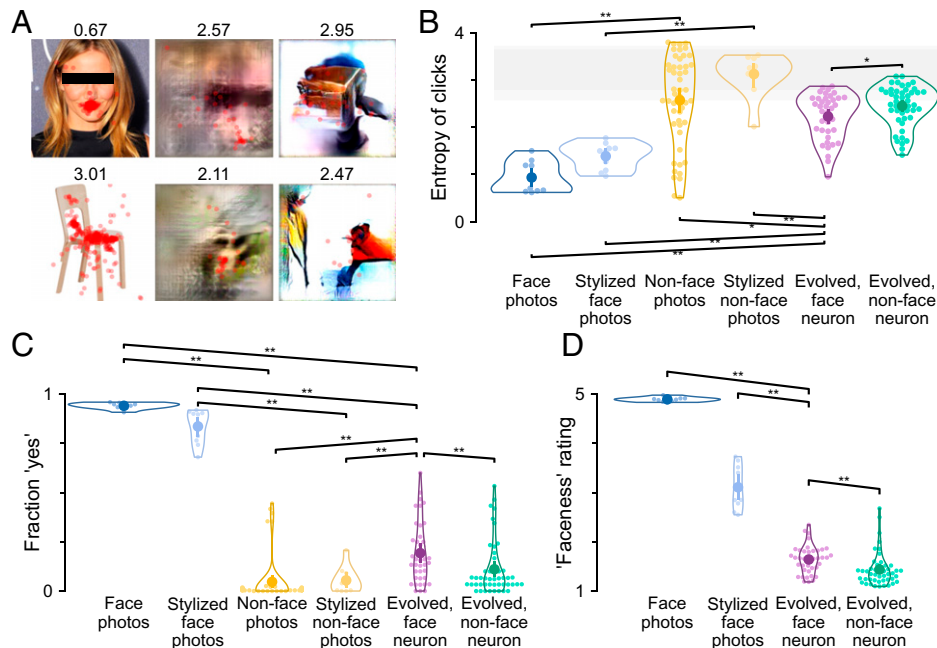


Fig. 5. Experiments 4 to 6: locating the mouth, binary classification, and rating of faceness. (A) In experiment 4, subjects were asked to click on the mouth (Fig. 1C). Click locations are shown for six example images. (B) The swarm plot shows the entropy of click locations for each image across subjects. Shading indicates permutation 95% CIs at the image (lighter gray) or group level (darker gray) when click locations were permuted across images to represent the null hypothesis that there was no difference across images. (C) In experiment 5, subjects indicated whether or not an image contained a face (Fig. 1A). The swarm plot shows the fraction of yes answers for each image across subjects. Plot conventions follow those in Fig. 2F. (D) In experiment 6, subjects provided a faceness rating between one (not a face) and five (most face like) to each image (Fig. 1A). The swarm plot shows the average faceness ratings for each image across subjects. FDR correction was across all three tests performed. In B–D, plot conventions follow those in Fig. 2F. * $P < 0.05$, ** $P < 0.01$, one-tailed permutation test with test direction indicated by the slope of the square bracket, FDR corrected across all seven tests performed in B, C, and all three tests performed in D.

images evolved from nonface neurons (2.5 ± 0.4 ; $P = 0.017$). As an alternative way to quantify click consistency, the spread of a Gaussian fitted to click locations per image showed the same between-group differences (SI Appendix, Fig. S3A). Click consistency was not explained by center bias, which did not differ across groups (SI Appendix, Fig. S3B).

In experiment 5, we asked directly, “Is there a face in this image?” (Fig. 1A). No definitions or further explanations were provided, so participants had to interpret the question based on their own concept of the word face. Face photos were unequivocally reported as faces (fraction of yes responses = 0.94 ± 0.02) (Fig. 5C), even when distorted in the stylized versions (fraction of yes responses = 0.84 ± 0.09). Nonface object images were clearly reported as not face (0.05 ± 0.12 ; 0.01 ± 0.01 when excluding five images of dogs, which were grouped with nonface object images but received a fraction of yes responses of 0.39 ± 0.04). Images evolved from face neurons received a fraction of 0.19 ± 0.16 yes responses, less than face photos ($P < 10^{-5}$, one-tailed permutation test, FDR corrected across seven tests) but more than nonface object images ($P < 10^{-5}$) and slightly more than images evolved from nonface neurons (0.11 ± 0.14 ; $P = 0.007$).

Experiment 6: Faceness Rating. The results of experiment 5 showed that evolved images were not consistently classified as faces in a binary choice task. In an additional experiment, we measured a graded value of faceness rather than just the binary distinction of face or nonface. For each image, subjects were asked to “rate the faceness of this image on a scale of 1 to 5.” As in experiment 5, we did not elaborate on the definition of faceness, nor did we tell the subjects what kind of faces to expect. Subjects used the whole scale for rating. Face photos received ratings of 4.90 ± 0.04 (Fig. 5D), and nonface object images received ratings of 1.2 ± 0.1 (Fig. 6A). All evolved images received low ratings (Fig. 5D). There was a small, although statistically significant, difference between the ratings assigned to face neuron evolved images (1.6 ± 0.3) and nonface neuron evolved images (1.4 ± 0.3 ; $P = 0.002$, FDR corrected across three tests). The low rating was not fully explained by generator style, as stylized face images received ratings of 3.1 ± 0.4 .

These ratings provide a unique opportunity to compare face neuron responses with a graded measure of face semantic content based on human perception. To this end, we collected faceness ratings for 131 additional images (SI Appendix, Fig. S4A) that were included as reference images in some of the evolution experiments. Faceness ratings for this set of images are presented in Fig. 6A. Subjects gave the highest rating to face photos, as expected. Human faces with various modifications, such as faces wearing personal protective equipment or faces including bodies, received slightly lower ratings. Monkey faces received lower ratings still. Nevertheless, all of the images mentioned so far received ratings of at least three. Other animate images (e.g., elephant) and inanimate objects with face-like features (e.g., jack-o'-lantern) received a wide range of ratings centered around 2.5 (other animate: 2.7 ± 0.5 ; inanimate face like: 2.5 ± 0.9). As expected, the lowest ratings (1.2 ± 0.1) were given to nonface objects. Nonface objects had the most similar ratings to both groups of evolved images (1.6 and 1.4 , respectively).

Did the faceness ratings, a graded measure of semantic face content, correspond well with neuronal responses that would be considered face selective using a conventional metric based on binary categories? We described the responses of each neuron by two numbers: 1) the FSI and 2) the correlation between neuronal firing rates and image faceness ratings. This analysis is illustrated for four example face neurons in Fig. 6B, where the FSI is indicated

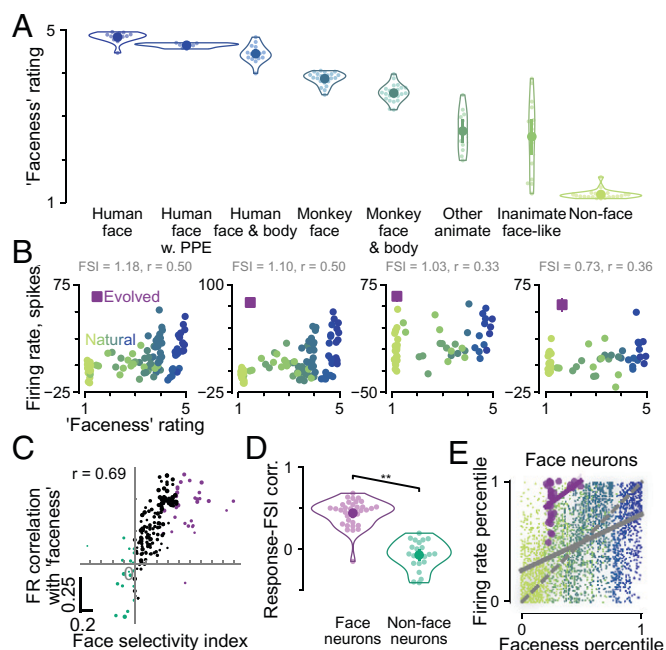


Fig. 6. Face neuron responses were correlated with graded ratings of faceness. (A) The swarm plot shows faceness ratings for 131 natural images of objects grouped by category. The images are shown in SI Appendix, Fig. S4A. (B) Responses of example face neurons are compared with image faceness ratings, including the natural images in A (green to blue dots) and the evolved image for each neuron (purple square). The text indicates the FSI of the neuron and the correlation coefficient between natural image faceness and neuronal responses. (C) The face selectivity of 220 visual neurons as quantified by FSI is compared with the correlation between faceness rating and neuron spiking responses. Each point represents a neuron. Face and nonface neurons are colored purple and green, respectively. The size of each point is scaled relative to the neuron’s trial to trial self-consistency. (D) The swarm plot compares firing rate with image faceness correlation values (y values in C) for face and nonface neurons. In A and D, plot conventions follow those in Fig. 2F. $**P < 0.01$, one-tailed permutation test. (E) Image faceness and firing rates were normalized as percentiles within each neuron and then pooled over face neurons separately for natural and evolved images. Each point represents one image response in one face neuron. Green to blue scatter represents natural images, and shading represents their distribution (kernel density estimate). Purple scatter represents evolved images. Solid lines represent linear regression fits. The dashed gray line is the identity line.

alongside the full range of responses in relation to image faceness ratings. The neuronal responses were correlated with faceness with coefficients ranging from 0.33 to 0.5 for the example neurons. This correlation was positive for all face neurons but one (CIT, FSI > 0.5) and indeed, for most neurons with an FSI > 0 (Fig. 6C and D). Across 220 neurons for which we both evolved images and showed the rated natural images, there was a clear positive relationship between the firing rate–faceness correlation and the FSI (Fig. 6C) (Pearson’s $r = 0.69$; $P = 4 \times 10^{-32}$). Face neurons had higher response correlations to image faceness ($r = 0.44 \pm 0.16$) than nonface neurons (-0.07 ± 0.17 ; $P < 10^{-5}$, one-tailed permutation test) (Fig. 6D). Although face photos received high faceness ratings and although neurons with high FSI responded strongly to faces by definition, it is not a forgone conclusion that face neuron responses should be positively correlated with faceness ratings at the image level. Unlike FSI, faceness ratings are a graded rather than binary measure of face vs. nonface. Indeed, when the firing rate–faceness analysis was restricted to only nonface object images, the responses of more face-selective neurons were still more correlated with faceness ratings (SI Appendix, Fig. S4C) (Pearson’s $r = 0.41$; $P = 2 \times 10^{-10}$).

Was high faceness rating for an image a necessary condition for strong responses to that image from face-selective neurons? For

the four example face neurons in Fig. 6B, evolved images (purple squares) elicited higher responses than all natural images, including those with comparable faceness ratings. To summarize across face-selective neurons, we overlaid one plot per neuron similar to Fig. 6B by normalizing, per neuron, all image responses (evolved or natural) as a percentile relative to natural image responses. We did the same for image faceness ratings. The summary plot in Fig. 6E shows that for face neurons overall, neuronal responses positively correlated with faceness, but evolved images consistently evoked higher firing rates than would be expected by their faceness rating in comparison with the regression fit for natural images.

Discussion

Face-selective neurons in primate visual cortex have been extensively studied (2, 3, 5, 26, 27). These so-called face neurons tend to cluster on the cortex into “face patches” that are stereotyped across individuals (5). Yet, do face neurons truly represent the semantic concept of “face,” or are they better described as responding to visual features typical of but dissociable from faces? If face neurons are semantic, they should respond exclusively to face images. Conversely, if face neurons respond to some nonface images as strongly as to realistic faces, these neurons are better described as visual feature selective. Face neurons have been observed to respond moderately to inanimate objects, such as clocks or round fruits (5, 28). However, because responses to faces are yet stronger, these observations do not seriously challenge the semantic hypothesis. There is so far no convincing evidence that nonface images can activate face neurons as strongly as face images do. A recently developed algorithm, XDream, can generate images that strongly activate visual neurons, including face neurons (18, 19), for which evolved images elicited as high of responses as face photos (Fig. 1D). Thus, these synthetic stimuli provide an opportunity to test the alignment between face neuron selectivity and the semantic category of face. If face neurons are semantic, evolved images that strongly activate face neurons should be perceived to be as face like as possible compared with realistic faces, with stylized faces (to account for limits of the image generator), or with monkey faces (to account for possible species-specific preference). Conversely, if face neuron evolved images are less face like than any of the comparisons, the evolved images would constitute counterexamples to the semantic hypothesis.

To measure perceptual similarity, we conducted psychophysics experiments on Mechanical Turk (with pilot experiments in the laboratory) (SI Appendix, Fig. S5) to assess whether subjects perceived faces in images evolved by the XDream algorithm, comparing images generated for face neurons and nonface neurons and comparing with natural object images. We probed perception of evolved images in six psychophysics experiments progressively more focused on face semblance from open-ended description (experiment 1) to binary face detection (experiment 5) and faceness ratings (experiment 6). To capture the ambiguous semantic category of faces impartially, we deliberately gave minimal instructions, no definitions, and no feedback on correctness. Our experiments consistently showed that subjects did not interpret evolved images as faces. Subjects described evolved images using a wide variety of words that were different from the words they used to describe face photos (experiment 1), chose the label face much less often than for face photos in forced choice categorization (experiments 2 and 5), did not consistently choose face images as more visually similar (experiment 3), were less consistent in locating facial features than with face photos (experiment 4), and rated evolved images with much lower faceness than face photos (experiment 6). Meanwhile, subjects identified animal-related

attributes in the evolved images, showing a small but significant tendency to associate evolved images more with categories such as monkeys, dogs, or faces than with categories such as books or chairs. Furthermore, in most experiments, there was a small but significant effect showing that images evolved from face neurons resembled face photos slightly more than nonface object images and images evolved from nonface neurons.

Although we focused on 86 evolved images, our conclusions remained the same when we repeated the analyses using relaxed criteria to include more evolved images, either by including intermediate FSI values (SI Appendix, Fig. S6, 315 evolved images) or by using imputed values for neurons with missing or unreliable FSI estimates (SI Appendix, Fig. S7, 187 evolved images). In both repeat analyses, evolved images for more face-selective neurons were perceived as more face like than evolved images for less face-selective neurons, even though neither group closely resembled real faces.

There are a few caveats when interpreting the present results. One is that we studied the selectivity of monkey neurons but tested human perceptual properties. It is not feasible to conduct some of the behavioral experiments (such as experiment 1) in monkeys. We cannot rule out the possibility that monkeys might perceive the evolved images as more (or less) face like than human subjects have reported here. Conversely, putative human face-selective neurons may not respond strongly to the evolved images here. We share all data, stimuli, and code to facilitate future experiments to test whether the evolved stimuli examined here are good counterexamples to human face-selective neural signals or whether evolved images based on human neurophysiological recordings can look more face like to human subjects.

Second, we predominantly studied multi- and single-unit spiking activity, whereas semantic information may be more relevantly encoded at the population level. Studies have generally found the same category selectivity in face patches using noninvasive imaging, single-unit activity, or multiunit activity (8, 29), but detailed tuning patterns likely differ, as they do even among face neurons. Thus, it is conceivable that single neurons may respond to face-related features at the same time that population activity encodes semantic categories (11). Population activity can refer to several distinct concepts—such as the (weighted) average activity of neurons within an anatomical area, within certain cell types, in a cortical layer, or in a minicolumn—that are more or less directly captured by a range of experimental techniques. Mesoscopic signals, like local field potentials or intracranial field potentials, or macroscopic noninvasive signals, like fMRI, magnetoencephalography (MEG), or electroencephalography (EEG), could lead to different conclusions. Future work is needed to test which, if any, level of neural activity encodes semantic selectivity for faces and to elucidate how such selectivity could emerge from the firing of nonsemantic neurons.

Although we find that semantics is an inadequate model for describing the responses of face neurons, subjects nevertheless described face neuron evolved images as face about 4% of the time in open-ended description, when they could have used any of about 7,000 common English nouns we accepted. While this could indicate a human bias toward seeing faces that was possibly enhanced by the presence of faces in control images, this result also suggests that evolved images do contain features that are reminiscent of faces. Moreover, in most experiments, evolved images from face neurons were perceived as slightly more face like than nonface object images and also, than evolved images from nonface neurons. Furthermore, when considering only natural images, face neuron responses were moderately correlated to faceness ratings. This correlation remained even when considering only

nonface object images. This relationship between face-selective neural responses and graded face semblance (even among nonface stimuli) is reminiscent of prior results in the human fusiform face area (30). However, this correlation does not extend to evolved images, which elicited comparable face neuron responses with real faces but were rated with far lower faceness values.

In conclusion, we found that face neurons are tuned to visual features that are correlated with, but are not exclusive to, the semantic category of faces by generating stimuli that evoked face-level responses but that were not perceived to be face like. This view is compatible with evidence from several studies showing a dissociation between the tuning of (nonface-selective) IT neurons and the category for which they are supposedly selective (14–16). Although face neuron selectivity is not coextensive with the semantic category of face, this selectivity might well serve the function of detecting faces and processing facial features. Such a teleological hypothesis cannot be tested by feature tuning alone but requires causal perturbation. The results highlight the challenges associated with word models that are ambiguous and difficult to falsify and emphasize the need for more rigorous and quantitative theories of neuronal tuning along ventral visual cortex.

Materials and Methods

Images. There were three main sources of images: evolved images, control images involving photos and drawings, and stylized images. Of those, 166 were included in main analyses of experiments 1 to 5, and 237 were included in experiment 6. All evolved, control, and stylized images used in main analyses are shown in *SI Appendix, Figs. S1 and S4A*.

Evolved images were generated using the XDream algorithm (18, 19) based on neuronal recordings from macaque IT (18). Briefly, XDream consists of a closed-loop algorithm that starts with noise images and gradually tweaks images to trigger high firing rates by a single neuron or neuronal multiunit. The algorithm has three components: an image generator, a fitness function given by neuronal firing rates, and a genetic search algorithm that selects the best candidate images in each generation and introduces visual variations to these successful images (31). In most of the figures, we consider 39 evolved images from 34 recording sessions in the middle lateral face patch (20) in CIT in four monkeys and 47 images from 39 recording sessions in nonface patches (46 in CIT, 1 in V1) in five monkeys. In Fig. 6C, recording locations were represented by 78 images from CIT, 96 from PIT, 5 from V4, 1 from V2, and 40 from V1.

Control images in experiments 1 to 5 were taken from the internet and included images of books, butterflies, cars, chairs, clouds, dogs, human faces, fruits, trees, wheels, and abstract drawings. We used 10 human face photos and five images for each of 10 nonface object categories. In experiment 6, together with human face photos and stylized faces, we tested 131 additional images that were used to record neuronal responses. These additional images comprised images from a published set (32) and photos taken in the laboratory. Although we included human faces for comparison, we deliberately excluded monkey faces in experiments 1 to 5 to avoid possible priming effects, even as we recognized that some evolved images may be reminiscent of monkey faces. We included monkey faces in experiment 6 because they have been used to record neuronal responses.

Stylized images were based on additional images from each control category (10 faces and 10 nonface objects, 1 from each nonface category). Those images were reconstructed from the image generator underlying XDream by using back propagation to optimize for pixel-level similarity (19).

Neurophysiological Responses. Neuronal responses were recorded using high-impedance intracranial electrodes in floating microelectrode arrays (Micro-Probes) or microwire bundles (4). Background-subtracted firing rates were computed using windows for background and evoked activity that were chosen and fixed before each evolution experimental session. The dataset includes previously published data (18).

Mechanical Turk Experiments. Behavioral experiments were conducted on the Amazon Mechanical Turk platform through psiTurk (33). Initial versions of the experiments were conducted in the laboratory (five subjects) with eye tracking to obtain a baseline level of subject performance that could be compared with the performance on Mechanical Turk, where subjects could not be monitored (*SI Appendix, Fig. S5*). All participants provided informed consent and received monetary compensation for participation in the experiments. All experiments were conducted according to protocols approved by the Institutional Review Board at Boston Children's Hospital. Responses from a Mechanical Turk subject were included if they attained the minimum in-laboratory accuracy, even if the subject did not finish the whole experiment. Each image received at least 25 responses. Details about each experiment are provided in the next sections.

In experiments 1, 2, 4, 5, and 6, generated images were split into eight sets of about 180 images each. Every set included all of the natural images. Each subject completed trials on one image set. In experiment 3, subjects responded to trios of images for 189 to 190 trials. Each subject completed only one experiment. The order of image presentation was randomized in all experiments.

Images were presented in color at a size of 256×256 pixels. No attempt was made to monitor eye movements in the Mechanical Turk experiments, but the images were flashed for 200 ms, thus minimizing the effects of eye movement during image presentation. Subjects provided informed consent and were compensated for participating in these studies. There was no time limit to respond in any of the experiments. Subjects were not allowed to give the same response (e.g., pressing the "1" key) more than five times in a row. No feedback was provided to the subjects.

To avoid introducing bias, we did not inform the subjects of the underlying hypotheses or questions addressed in this study. The experiments were defined by the minimal instructions listed next for each experiment and are shown schematically in Fig. 1 A–C. In particular, no definition of the word "face" was provided, and we left it to participants to interpret the word when it was part of the instructions (experiments 5 and 6).

Experiment 1. A schematic of the experiment is shown in Fig. 1A. A fixation cross in a 500×500 -pixel gray box was shown for 1,000 ms. Then, the test image was flashed for 200 ms. After image presentation, subjects were asked to type one word to describe the image. The answer was accepted during the experiment if it was contained in a set of 6,801 commonly used nouns pulled from the American National Corpus Project (34, 35); otherwise, the subject was prompted to check the spelling and try again. Subjects were excluded from analysis if they responded to under 25% of the questions or gave the same answer over 25% of the time. Example responses are shown in Fig. 2 A–C and *SI Appendix, Fig. S2 A–C*.

Experiment 2. A schematic of the experiment is shown in Fig. 1A. A fixation cross in a 500×500 -pixel gray box was shown for 1,000 ms. Then, the test image was flashed for 200 ms. After image presentation, subjects were asked to "choose the most appropriate category" from five options by typing in the corresponding number. For natural images, the correct category was always an option, while the other four options were chosen randomly from the other nine categories. For abstract drawings and evolved images, all five options were chosen randomly. Subjects were excluded from analysis if their accuracy was less than 84% on natural object images.

Experiment 3. A schematic of the experiment is shown in Fig. 1B. In this experiment, subjects were presented with three images: a test image in the center and two choice images on either side. Subjects were instructed to choose whether the left or right image was more similar to the center image in a two-alternative forced choice manner. The images remained on the screen until subjects made their choice. Subjects pressed the 1 key to select the left image as more similar or 2 for the right image. The test images were always evolved images. The two choice images were randomly chosen from two different categories. Each evolved image was tested at least once for each option pair (10 choose 2 = 45 pairs). Each subject saw 10 control trials where one of the choice images matched the test image. Subjects were excluded from analysis if they matched fewer than 90% of the control trials correctly.

Experiment 4. A schematic of the experiment is shown in Fig. 1C. In this experiment, subjects were asked to "click on the mouth (if unsure, make your

best guess)." No further instructions were provided. In the case of nonface object images (e.g., chairs), subjects still had to click somewhere in the image. The image remained on the screen until the subjects clicked on it. Subjects were excluded from analysis if they failed to click on a region around the mouth in more than one of the normal face photos. (We did not use a percentage accuracy cutoff due to the low number of images tested in the laboratory.)

Experiment 5. A schematic of the experiment is shown in Fig. 1A. A fixation cross in a 500- × 500-pixel gray box was shown for 1,000 ms. Then, the test image was flashed for 200 ms. After image presentation, subjects were asked, "Is there a face in this image?" They responded by pressing the "y" key for "yes" or the "n" key for "no." Subjects were excluded from analysis if they answered yes to fewer than 90% of control face photos.

Experiment 6. A schematic of the experiment is shown in Fig. 1A. A fixation cross in a 500- × 500-pixel gray box was shown for 1,000 ms. Then, the test image was flashed for 200 ms. After image presentation, subjects were asked to "rate the faceness of this image on a scale of 1 to 5." They responded by typing a number from one to five. No further instructions were provided. Subjects were excluded from analysis if they gave an average rating of less than four to face photos.

Word Similarity Quantified by WP or LexVec Word Embedding–Based Metrics. Given two words w_1, w_2 , the WP similarity (WPS) is defined as

$$\text{WPS} = \frac{\text{depth}(\text{LCS}(w_1, w_2))}{\frac{1}{2}(\text{depth}(w_1) + \text{depth}(w_2))},$$

where depth indicates the number of nodes from the top to arrive at the word in the WordNet hierarchy and LCS refers to the least common subsumer: in other words, the most specific shared category. WP similarity ranges from zero (no relation) to one (identity). In the WordNet hierarchy, a word can have multiple hierarchical definitions; in this case, the highest WPS over all definitions was used. For example, a WPS of 0.31 between "woman" and "face" is calculated as follows: woman = entity → physical entity → causal agent → person → adult → woman; face = entity → physical entity → thing → part → body part → external body part → face; $2/(0.5 \times (7 + 6)) \approx 0.308$. A WPS of 0.57 between "woman" and "butterfly" is calculated by the following: woman = entity → physical entity → object → whole → living thing → organism → person → adult → woman; butterfly = entity → physical entity → object → whole → living thing → organism → animal → invertebrate → arthropod → insect

→ lepidopterous insect → butterfly; $6/(0.5 \times (9 + 12)) \approx 0.571$. The LCS is the same between "woman" and "tree" or "dog" while the latter two have shallower hierarchical definitions (depth = 10 and 9, respectively) than butterfly, resulting in even higher WP similarity to "woman" [$6/(0.5 \times (9 + 10)) \approx 0.632$ and $6/(0.5 \times (9 + 9)) \approx 0.66$, respectively].

LexVec word embedding maps each word to a vector such that the vectorial dot product between two word vectors approximates, conceptually, the log odds (or enhancement in probability) on one word occurring given that the other word occurs nearby (36). We used precalculated word embeddings from Salle and Villavicencio (25), where each word is mapped to a 300-dimensional vector. LexVec similarity between two words was calculated as the dot product between the two corresponding word vectors, and thus, the value can be interpreted as approximate log odds as defined above. The log odds were lower bounded by zero, so the minimum dot product value was approximately zero.

Entropy in Experiment 4. Entropy was calculated by putting x and y coordinates of click locations into 121 bins (11×11 grid on the image). For each bin i in each image, the click probability p_i was calculated as the number of clicks in the bin divided by the total number of clicks on the image. The entropy for each image was calculated as $\sum_i (-p_i \log p_i)$.

Statistical Tests. We conducted pairwise permutation tests by permuting image assignment to categories for 10,000 permutations. P values associated with Pearson's r (experiment 6) were calculated using the exact distribution for the null hypothesis that the two variables were drawn from a bivariate normal distribution with zero covariance, as implemented in the Python library "scipy" (37). One or two tailedness of tests and other types of tests are noted in the text and were implemented using the Python library "scipy.stats." P values for multiple comparisons were corrected to control false discovery rate at the level of 0.05 using the two-stage Benjamini–Krieger–Yekutieli procedure (38) as implemented in the Python library "statsmodels" (39).

Data Availability. Psychophysics and neural recording data have been deposited on the lab website and are publicly accessible (<https://klab.tch.harvard.edu/resources/whatisafaceneuron.html>), including part of data that were collected in previous work (18).

ACKNOWLEDGMENTS. This work was supported by NIH Grants R01EY026025, R01EY16187, and P30EY12196 and NSF Grant CCF-1231216. C.R.P. was supported by a Packard Fellowship.

- 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).
- 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).
- 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).
- D. B. McMahon, I. V. Bondar, O. A. 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).
- D. Y. Tsao, W. A. Freiwald, R. B. Tootell, M. S. Livingstone, A cortical region consisting entirely of face-selective cells. *Science* **311**, 670–674 (2006).
- T. Allison *et al.*, Face recognition in human extrastriate cortex. *J. Neurophysiol.* **71**, 821–825 (1994).
- 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).
- T. Decramer *et al.*, Face neurons in human visual cortex. *bioRxiv* [Preprint] (2020). <https://doi.org/10.1101/2020.10.09.328609> (Accessed 3 February 2022).
- P. Pietrini *et al.*, Beyond sensory images: Object-based representation in the human ventral pathway. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 5658–5663 (2004).
- A. G. Huth, S. Nishimoto, A. T. Vu, J. L. Gallant, A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**, 1210–1224 (2012).
- N. A. Ratan 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).
- G. Buzsáki, *The Brain from Inside Out* (Oxford University Press, 2019).
- D. J. Freedman, M. Riesenhuber, T. Poggio, E. K. Miller, A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J. Neurosci.* **23**, 5235–5246 (2003).
- E. M. Meyers, D. J. Freedman, G. Kreiman, E. K. Miller, T. Poggio, Dynamic population coding of category information in ITC and PFC. *Neurophysiology* **93**, 1342–1357 (2005).
- S. Bracci, J. B. Ritchie, I. Kalfas, H. P. Op de Beeck, The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. *J. Neurosci.* **39**, 6513–6525 (2019).
- P. Bao, L. She, M. McGill, D. Y. Tsao, A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
- S. G. Wardle, J. Taubert, L. Teichmann, C. I. Baker, Rapid and dynamic processing of face pareidolia in the human brain. *Nat. Commun.* **11**, 4518 (2020).
- C. R. Ponce *et al.*, Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell* **177**, 999–1009.e10 (2019).
- W. Xiao, G. Kreiman, XDream: Finding preferred stimuli for visual neurons using generative networks and gradient-free optimization. *PLOS Comput. Biol.* **16**, e1007973 (2020).
- 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).
- E. B. Issa, J. J. DiCarlo, Precedence of the eye region in neural processing of faces. *J. Neurosci.* **32**, 16666–16682 (2012).
- E. Rosch, C. Mervis, W. Gray, D. Johnson, P. Boyes-Braem, Basic objects in natural categories. *Cogn. Psychol.* **8**, 382–439 (1976).
- Z. Wu, M. Palmer, Verb semantics and lexical selection. *arXiv* [Preprint] (1994). <https://arxiv.org/abs/cmp-lg/9406033> (Accessed 3 February 2022).
- C. Fellbaum, *WordNet: An Electronic Lexical Database* (Bradford Books, 1998).
- A. Salle, A. Villavicencio, "Incorporating subword information into matrix factorization word embeddings" in *Proceedings of the Second Workshop on Subword/Character Level Models*, M. Faruqui, H. Schütze, I. Trancoso, Y. Tsvetkov, Y. Yaghoobzadeh, Eds. (Association for Computational Linguistics, New Orleans, LA, 2018), pp. 66–71.
- M. P. Young, S. Yamane, Sparse population coding of faces in the inferotemporal cortex. *Science* **256**, 1327–1331 (1992).
- 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).
- E. M. Meyers, M. Borzello, W. A. Freiwald, D. Y. Tsao, Intelligent information loss: The coding of facial identity, head pose, and non-face information in the macaque face patch system. *J. Neurosci.* **35**, 7069–7081 (2015).
- D. Y. Tsao, M. S. Livingstone, Mechanisms of face perception. *Annu. Rev. Neurosci.* **31**, 411–437 (2008).
- M. Meng, T. Cheria, G. Singal, P. Sinha, Lateralization of face processing in the human brain. *Proc. Biol. Sci.* **279**, 2052–2061 (2012).
- Y. Yamane, E. T. Carlson, K. C. Bowman, Z. Wang, C. E. Connor, A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nat. Neurosci.* **11**, 1352–1360 (2008).

32. 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).
33. T. M. Gureckis *et al.*, psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behav. Res. Methods* **48**, 829–842 (2016).
34. N. Ide, K. Suderman, "The American National Corpus first release" in *LREC* (2004).
35. D. Quintans, *Get common nouns from MASC 3.0.0*. M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva, Eds. (European Language Resources Association, Lisbon, Portugal, 2019). http://www.desiquintans.com/downloads/nounlist/01_Get_common_nouns_from_MASC_3.0.0.pdf. Accessed 3 February 2022.
36. A. Salle, A. Villavicencio, M. Idiart, "Matrix factorization using window sampling and negative sampling for improved word representations" in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, K. Erk, N. A. Smith, Eds. (Volume 2: Short Papers) (Association for Computational Linguistics, Berlin, Germany, 2016), pp. 419–424.
37. P. Virtanen *et al.*, SciPy 1.0 Contributors, SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
38. Y. Benjamini, A. M. Krieger, D. Yekutieli, Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491–507 (2006).
39. S. Seabold, J. Perktold, "Statsmodels: Econometric and statistical modeling with python" in *9th Python in Science Conference*. S. van der Walt, J. Millman, Eds. (SciPy 2010, Austin, TX, 2010).