nature human behaviour

Article

https://doi.org/10.1038/s41562-025-02218-1

Feature-based encoding of face identity by single neurons in the human amygdala and hippocampus

Received: 2 March 2024

Accepted: 16 April 2025

Published online: 06 June 2025

Check for updates

Runnan Cao[1] ✉, Jinge Wang[2,12], Chujun Lin[3,12], Emanuela De Falco[4,12], Alina Peter[5,12], Hernan G. Rey[6], Peter Brunner[7], Jon T. Willie[7], James J. DiCarlo[5], Alexander Todorov[8], Ueli Rutishauser[9], Xin Li[10], Nicholas J. Brandmeir[11] & Shuo Wang[1,7] ✉

Neurons in the human amygdala and hippocampus are classically thought to encode a person's identity invariant to visual features. However, it remains largely unknown how visual information from higher visual cortical areas is translated into such a semantic representation of an individual person. Here, across four experiments (3,581 neurons from 19 neurosurgical patients over 111 sessions), we demonstrate a region-based feature code for faces, where neurons encode faces on the basis of shared visual features rather than associations of known concepts, contrary to prevailing views. Feature neurons encode groups of faces regardless of their identity, broad semantic categories or familiarity; and the coding regions (that is, receptive fields) predict feature neurons' response to new face stimuli. Together, our results reveal a new class of neurons that bridge perception-driven representation of facial features with mnemonic semantic representations, which may form the basis for declarative memory.

How the human brain encodes and stores different face identities in memory is one of the most fundamental and intriguing questions in neuroscience. Two extreme hypotheses have been proposed. The 'feature-based model' posits that face representations are encoded by a broad and distributed population of neurons[1–4]. In this model, recognizing a particular individual requires access to many neurons, with each neuron responding to many different faces that share specific visual features, such as shape and skin tone[5,6]. A specific type of feature-based coding, axis-based feature coding, has been found in the non-human primate inferotemporal (IT) cortex using single-neuron recordings[7–10] and in humans using functional magnetic resonance imaging[11–13]. In these studies, the activity of neurons/voxels parametrically correlated with facial features along specific axes in face space. At the other extreme, the 'exemplar-based model' posits that explicit facial representations in the brain are formed by highly selective (sparse), yet visually invariant, neurons[14–17]. Identity neurons that selectively respond to many different images of a specific person's face embody exemplar-based coding. Such neurons have been identified in the human amygdala and hippocampus[16,17] and are thought to be part of the building blocks of episodic memory[17]. A hallmark of identity

[1]Department of Radiology, Washington University in St Louis, St Louis, MO, USA. [2]Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV, USA. [3]Department of Psychology, University of California San Diego, La Jolla, CA, USA. [4]Laboratory of Cognitive Neuroscience, École Polytechnique Fédérale de Lausanne, Route Cantonale, Lausanne, Switzerland. [5]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [6]Department of Neurosurgery, Medical College of Wisconsin, Milwaukee, WI, USA. [7]Department of Neurosurgery, Washington University in St Louis, St Louis, MO, USA. [8]Booth School of Business, University of Chicago, Chicago, IL, USA. [9]Departments of Neurosurgery and Neurology, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [10]Department of Computer Science, University at Albany, Albany, NY, USA. [11]Department of Neurosurgery, West Virginia University, Morgantown, WV, USA. [12]These authors contributed equally: Jinge Wang, Chujun Lin, Emanuela De Falco, Alina Peter. ✉e-mail: r.cao@wustl.edu; shuowang@wustl.edu

neurons is that their responses are clustered by high-level conceptual or semantic relatedness (for example, Bill Clinton and Hillary Clinton) rather than by shared visual features[18–20].

Feature-based and exemplar-based models of face processing are not mutually exclusive; they may operate at different stages of processing, with the former thought to give rise to the latter. However, the neural computations bridging these two forms of encoding remain poorly understood. To address this question, we seized a unique opportunity to directly record single-neuron activity in the human brain and utilized computational algorithms based on deep learning to examine face coding in the human amygdala and hippocampus (Fig. 1a). We identify a potential key missing link between the representation of specific facial features (feature-based coding) and the representation of specific people (exemplar-based coding) in the amygdala and hippocampus. We show that a subset of neurons in the human amygdala and hippocampus carry a 'region-based feature code' for face identity, suggesting an intermediate representation linking feature-based coding and exemplar-based coding.

Results

Identity neurons

We recorded from 2,082 neurons in the amygdala and hippocampus (collectively referred to as the medial temporal lobe (MTL)) of 12 neurosurgical patients (5 males; 38 sessions in total; Supplementary Table 1 and Supplementary Fig. 1) while they performed a one-back task. In this task, patients were instructed to respond whenever an identical image was repeated (Fig. 1b; accuracy = 75.2% ± 20.0%, mean ± s.d. across sessions). Participants viewed 500 natural face images of 50 celebrities (10 different images per celebrity/identity). A total of 1,577 neurons had an overall average firing rate greater than 0.15 Hz and we restricted our analysis to this subset of neurons, which included 753 neurons from the amygdala, 505 neurons from the anterior hippocampus, and 319 neurons from the posterior hippocampus (Supplementary Table 2).

To select 'identity neurons', we first used a one-way (1 × 50) analysis of variance (ANOVA) to identify neurons with a significantly unequal response to different identities ($P < 0.05$) in a window of 250–1,250 ms following stimulus onset. We imposed a second criterion to identify which identities a neuron was selectively responding to (selected identities): the neural response to such an identity was required to be at least 2s.d. above the mean of the neural responses to all identities. A total of 155 identity neurons satisfied both criteria (9.83%, binomial $P = 1.67 \times 10^{-15}$; Fig. 1d,i, and Supplementary Fig. 2a and Supplementary Table 2; see Supplementary Results for analysis of reliability of identity selectivity), consistent with previous work[16,18,21,22]. Of the 155 identity neurons, 53 responded to a single identity only (here referred to as 'single-identity (S-ID) neurons'[16,21]) and the remaining 102 neurons each responded to multiple identities (here referred to as 'multiple-identity (M-ID) neurons'[18,19]; Supplementary Fig. 2b–g; see also Supplementary Information). On average, M-ID neurons encoded 2.48 ± 0.61 identities (Supplementary Fig. 2h).

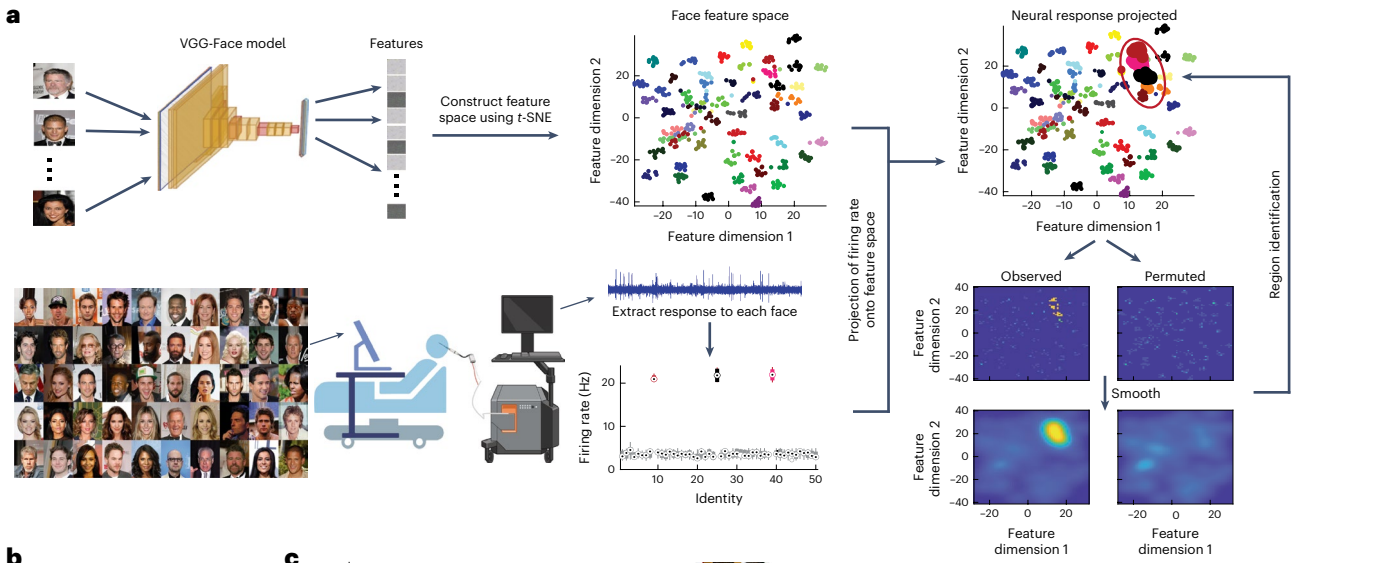
Feature-based coding of face identities

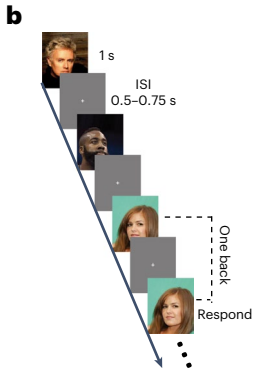
We next asked whether the multiple identities that M-ID neurons responded to were related. Some M-ID neurons are known to encode conceptually related identities (for example, Bill Clinton and Hillary Clinton)[18–20], showing that the response of these M-ID neurons is independent of visual features[16,17,21]. However, it is unknown whether M-ID neurons can also encode visually (rather than conceptually) similar identities (that is, identities sharing a similar visual appearance, for example, similar face and eye shape, skin or hair colour; see refs. 19,23 for analysis of low-level physical image similarities). To answer this question, we applied the following steps: (1) we extracted visual features from each stimulus and used these features to construct a feature space that indicated the relationship between the stimuli based on their appearance (Fig. 1a top left, and Supplementary Figs. 3 and 4); (2) we recorded single-neuron activity from the amygdala and hippocampus and extracted the neural response to each stimulus for each individual neuron (Fig. 1a bottom left); (3) for each neuron, we projected its response to each stimulus onto the feature space by multiplying its response magnitude to the corresponding stimulus location in the feature space, resulting in a response-weighted two-dimensional (2D) spike density map (Fig. 1a top right and Supplementary Fig. 5); and (4) to perform statistical analysis, select significant neurons and identify coding regions, we conducted a permutation test by shuffling the response to each stimulus and comparing the observed vs permuted spike density maps (Fig. 1a bottom right and Supplementary Fig. 5). We based our subsequent analyses on the selected neurons with coding regions.

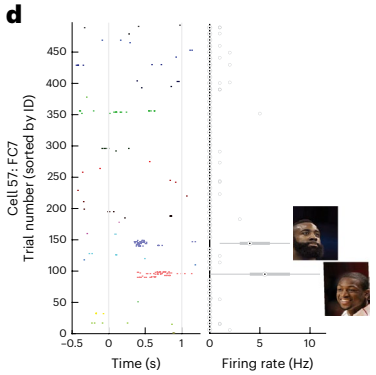
First, we extracted visual features from the pictures shown to the patients using a pre-trained deep neural network (DNN) Visual Geometry Group (VGG)-16 trained to recognize faces (Fig. 1a; see Supplementary Fig. 3a,b for DNN architecture and visualization of DNN features). Facial features were represented by the activation of thousands of units in the network (that is, the DNN features); we reduced the dimensionality of the DNN features to construct a 2D face feature space using *t*-distributed stochastic neighbour embedding (*t*-SNE) for each DNN layer (Fig. 1a,c,e,j and Supplementary Fig. 4a; note that quantifications are in this *t*-SNE space but we replicated our results in the full dimensional space of the DNN (Supplementary Fig. 2k); also note that the pairwise distance between faces in the full dimensional space is preserved in the *t*-SNE space as shown in Supplementary Fig. 3d). The dimensions of the face feature space represented the major variations in faces that led to successful recognition of the identities by the DNN. Note that this face feature space was derived solely from the images shown to participants without using any neural responses (Fig. 1a). We first compared the pattern of neural responses of M-ID neurons to features of the later (higher-level) DNN layers. In these later layers, faces of the same identity are clustered (see below). Later, we continued to compare the response of neurons to features in the earlier DNN layers, in which face identities are not clustered.

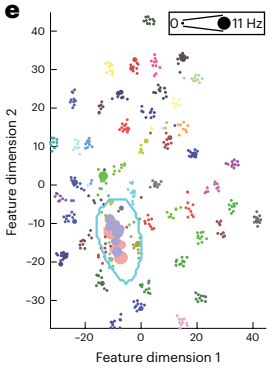
Fig. 1 | Feature-based neuronal coding of face identities. a, Overview of procedure and analysis pipeline. Faces from 50 celebrity identities were used for neural recordings. We used the VGG-Face model to extract visual features from each image and constructed a face feature space using *t*-SNE. We then projected the firing rate onto the face feature space to obtain a spike density map in the feature space. Lastly, by comparing the observed vs permuted spike density maps, we identified coding regions within the feature space. **b**, One-back task. **c**, The face feature space constructed by *t*-SNE for the DNN layer FC7. **d**–**m**, Two example neurons that encoded visually similar identities. **d**–**h**, Cell 57. **i**–**m**, Cell 128. **d**,**i**, Neuronal responses to 500 faces (50 identities). Trials are aligned to face stimulus onset and are grouped by individual identity. In each boxplot, the central mark represents the median, box edges indicate the 25th and 75th percentiles, whiskers extend to non-outlier extremes, and circles denote outliers. **e**,**j**, Projection of the firing rate onto the FC7 feature space. Each colour represents a different identity. The size of the dot indicates the firing rate. **f**,**k**, Estimate of the spike density in the feature space. By comparing the observed (top) vs permuted (bottom) responses, we could identify a region where the observed neuronal response was significantly higher in the feature space. This region was defined as the tuning region of a neuron. This region was defined as the tuning region of a neuron (delineated by the red and cyan outlines in **c**,**e**,**j**). **g**,**l**, A range of new faces (shown in dashed rectangles) synthesized based on the latent code of the anchor faces (that is, faces from the encoded identities at the extremes of each feature dimension) using StyleGAN2. Coordinates of the synthesized faces were linearly interpolated in the feature space. The original faces are plotted using the same coordinates as in **e** and **j**. **h**,**m**, A range of new faces synthesized around the average face using StyleGAN2. The standard deviation of the visual features across all faces from the encoded identities in the region determined the axis for extrapolating to a series of new faces.

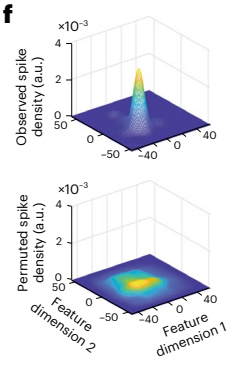
https://doi.org/10.1038/s41562-025-02218-1

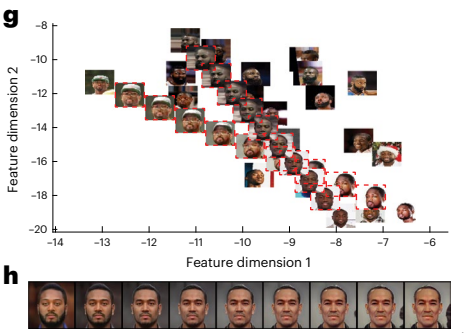


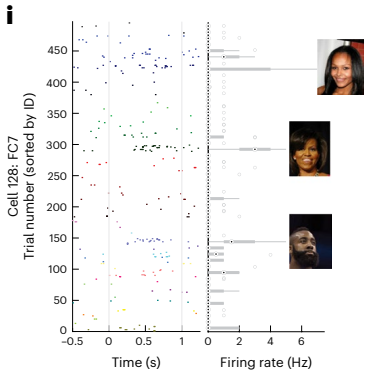


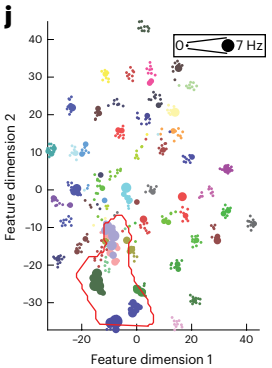


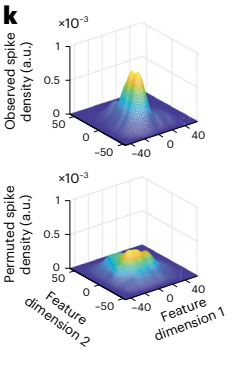


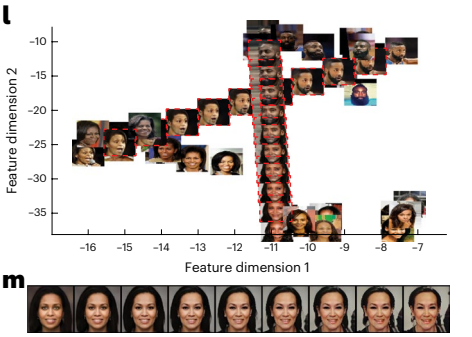












The feature space demonstrated an organized structure. For example, in the second fully connected (FC) layer FC7 (which is towards the top of the DNN hierarchy and demonstrates clustering of identities; Supplementary Figs. 3 and 4a), faces of the same identity were clustered, Feature Dimension 2 represented a gender dichotomy, and darker skinned faces were clustered at the bottom left corner of the feature space (Fig. 1c). Because the DNN had no access to semantic information or conceptual knowledge about the faces (for example, gender, ethnicity, social traits), faces were distributed in the DNN-derived feature space purely based on their visual appearance.

We next asked whether the response of M-ID neurons to seemingly unrelated identities could be understood in this unbiased visual feature space. We projected the neuronal responses of a given neuron to each face onto the feature space (that is, multiplying the firing rate of each face to its corresponding location in the feature space to derive a response-weighted 2D feature map; Fig. 1a,e,j; see Supplementary Fig. 5 for details). This revealed that some M-ID neurons were selective to different identities that were clustered together in the face feature space (Fig. 1e,j; see Supplementary Fig. 6 for more examples), suggesting that these M-ID neurons responded to face identities that were visually similar (for example, similar in face shape or skin tone). To formally quantify the tuning of M-ID neurons (Fig. 1a; see Methods and Supplementary Fig. 5 for a step-by-step illustration of the selection procedure), we estimated a continuous spike density map in the 2D feature space (Fig. 1f,k top) and used a permutation test (1,000 runs; Fig. 1f,k bottom) to identify the region(s) that had a significantly higher spike density above chance (red/cyan outlines in Fig. 1c,e,j; significant pixels were selected with permutation $P < 0.01$ and cluster size thresholds). This region indicates the part of the feature space to which a neuron was tuned. We refer to such neurons as exhibiting 'region-based feature coding' because they coded a certain region in the feature space. Through visualization of visual features, we observed that faces in a tuning region shared a combination of similar features such as skin tone, eye shape, lip thickness and so on (Fig. 1g,l); and synthesized faces around the average face of a tuning region further revealed the variation of visual features in the tuning region (Fig. 1h,m). At the population level, we found that for 42/102 M-ID neurons (41.2%), all the identities to which the neurons responded were clustered in feature space in the same region. We refer to these as 'feature M-ID neurons' (Supplementary Fig. 2a). The remaining M-ID neurons encoded identities distributed in separate locations in the feature space that were not part of the same region, and we referred to these neurons as 'non-feature M-ID neurons'. Therefore, feature M-ID neurons encoded identities that shared similar facial features, that is, they were visually similar.

In the DNN, the level of feature abstraction, and thus the clustering of identities, increases from earlier layers to later layers (Supplementary Figs. 3 and 4a). We therefore expect that feature M-ID neurons best reflect the facial features represented in later DNN layers. Indeed, we observed feature M-ID neurons only in the later DNN layers Pool5 (25 neurons), FC6 (27 neurons), FC7 (30 neurons) and FC8 (30 neurons; some neurons appeared in multiple layers; Supplementary Fig. 4a; see Supplementary Fig. 7a,d for a breakdown of amygdala and hippocampal neurons). The tuning region of an individual feature M-ID neuron covered 1.82–9.82% of the 2D feature space, with a similar mean coverage across layers (Fig. 2a; note that we adjusted the kernel size to be proportional to the feature dimensions such that the percentage of space coverage was independent of the size of the feature space). In contrast, the response of an individual S-ID or non-feature M-ID neuron covered a significantly smaller region in the feature space (Fig. 2a; two-tailed two-sample t-test: $P < 0.001$ for all comparisons). This result was as expected because the identities (and thus the tuning regions) encoded by non-feature M-ID neurons were not contiguous with each other and were further apart (Fig. 2b). As a whole, the neuronal population that we sampled covered 46–52% of the 2D feature space (Fig. 2c and Supplementary Fig. 2i; some areas were encoded by multiple neurons), suggesting that these neurons encoded a variety of visual features; and the tuning regions of non-feature M-ID neurons were more distributed compared with feature M-ID neurons. The distribution of pairwise distance between faces within each neuron's tuning region(s) further showed that feature M-ID neurons had a single, large tuning region, whereas non-feature M-ID neurons had smaller, non-contiguous tuning regions that were more widely distributed across the feature space (Fig. 2d and Supplementary Fig. 2j).

Visual similarity vs conceptual association

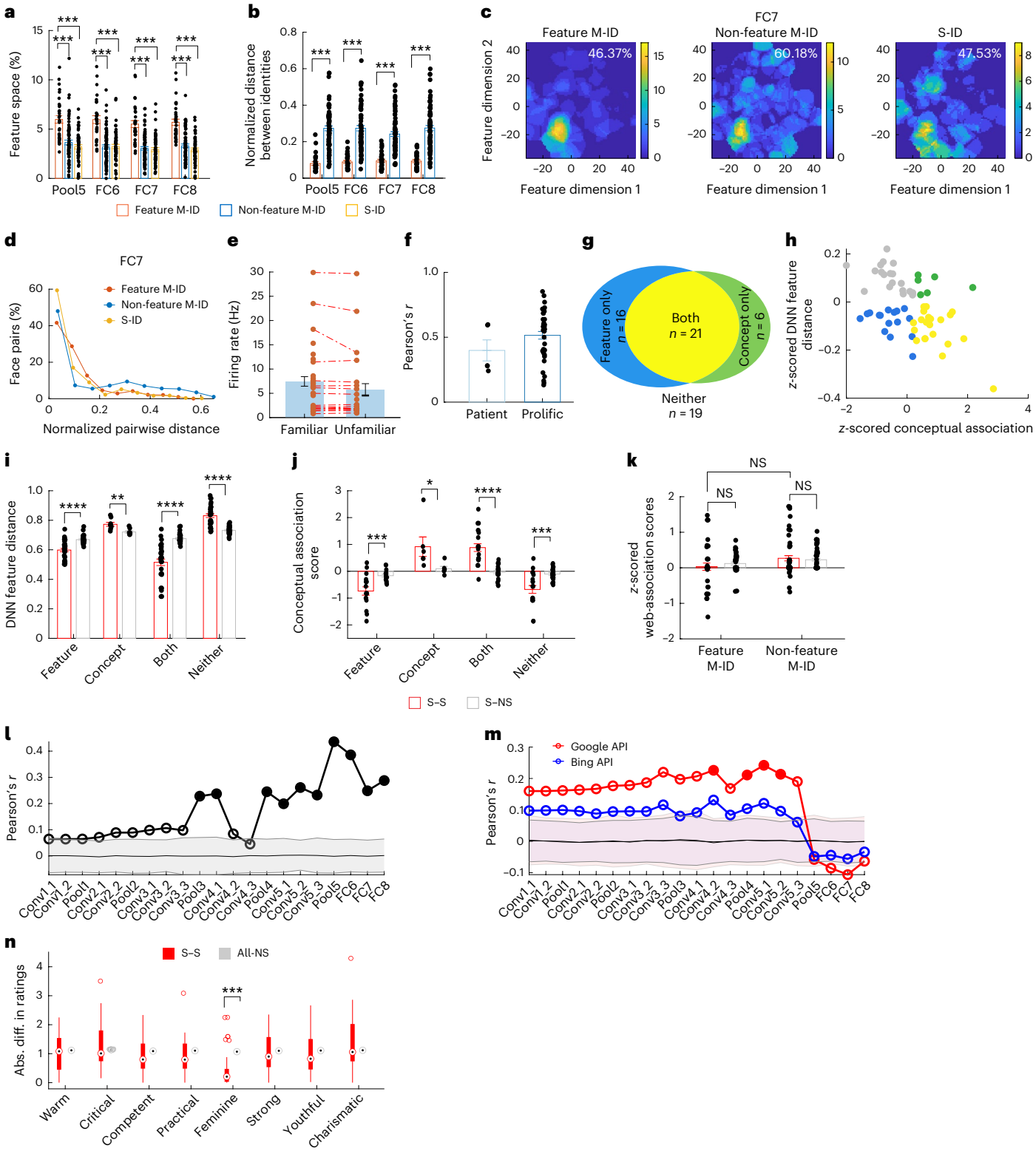
Were the multiple identities to which M-ID neurons responded conceptually related? To address this question, we next assessed whether there was a relationship between the visual similarity of the faces and their conceptual associations (that is, association between identities through concepts).

First, 5 patients we recorded from rated how conceptually related they thought each pair of identities was using an established paradigm

Fig. 2 | Population summary of region-based feature coding in identity neurons and comparison between visual similarity and conceptual association. a, Percentage of feature space covered by the tuning regions of identity neurons (feature M-ID: $n = 25, 27, 30, 30$ for Pool5, FC6, FC7 and FC8, respectively; non-feature M-ID: $n = 77, 75, 72, 72$; S-ID: $n = 53, 53, 53, 53$). **b**, Normalized distance between M-ID neurons' selected identities in the feature space (see **a** for the number of neurons). Error bars denote ±s.e.m. across neurons and dots show individual values. Two-tailed two-sample t-test (Bonferroni correction): $^{***}P < 0.001$. **c**, The aggregated tuning regions of the neuronal population. Colour bars show the number of neurons with tuning regions in a given area. Percentage in the upper right corner shows the percentage of feature space covered by at least one neuron. **d**, Distribution of pairwise distance between faces in each neuron's tuning region(s). S-ID: $n = 53$; feature M-ID: $n = 30$; non-feature M-ID: $n = 72$. **e**, Feature M-ID neurons did not differentiate familiar vs unfamiliar selected identities (two-tailed paired t-test, $P > 0.05$). Each dot represents a neuron. Error bars denote ±s.e.m. across neurons. **f**, Correlation between conceptual association ratings and visual similarity ratings. Error bars denote ±s.e.m. across participants ($n = 5$ for patients and $n = 40$ for general controls) and dots show individual values. **g**, Separate populations of M-ID neurons encoding visual features (that is, visual similarity) and concepts (that is, conceptual association). **h**, Response of all M-ID neurons as a function of visual similarity and conceptual association. Shown are the differences between S–S and S–NS pairs. Each circle represents a neuron. The colour of the circle indicates the classification of the neuron, as shown in **g**. **i**, DNN feature distance for each subpopulation of M-ID neurons (feature only: $n = 16$; concept only: $n = 6$; both: $n = 21$; neither: $n = 19$). **j**, Conceptual association ratings for each subpopulation of M-ID neurons. S–S, pairs of identities that a neuron was selective to. S–NS, pairs of identities where a neuron was selective to one but not selective (NS) to the other. Error bars denote ±s.e.m. across neurons and dots show individual values. Two-tailed two-sample t-test (uncorrected for multiple comparisons): $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ and $^{****}P < 0.0001$. **k**, Web-association score for M-ID neurons. Error bars denote ±s.e.m. across neurons and dots show individual values. Left: feature M-ID neurons ($n = 38$). Right: non-feature M-ID neurons ($n = 54$). We applied a two-tailed one-sample t-test to compare within each neuronal group and a two-tailed two-sample t-test to compare between neuronal groups (uncorrected for multiple comparisons). NS, not significant. **l**, Correlation between patients' visual similarity ratings and DNN feature similarity (that is, the negative of the DNN feature distance) for each DNN layer. **m**, Pearson correlation between DNN feature similarity and web-association scores (1,225 pairs of identities). Filled circles represent significant correlations (permutation test: $P < 0.05$; Bonferroni correction for multiple comparisons across DNN layers) and open circles represent non-significant correlations. Shaded area denotes ±s.d. across 1,000 permutation runs. **n**, The absolute difference (abs. diff.) in social trait judgements between identity pairs ($n = 42$ feature M-ID neurons). All-NS, all other pairs (that is, all excluding S–S pairs). In each boxplot, the central mark represents the median, box edges indicate the 25th and 75th percentiles, whiskers extend to non-outlier extremes, and circles denote outliers. Two-tailed two-sample t-test (uncorrected for multiple comparisons): $^{***}P < 0.001$.

https://doi.org/10.1038/s41562-025-02218-1

that has revealed neural coding of association between concepts (for example, neural correlates of conceptual associations between celebrities and between personally relevant people)[18,19] (Supplementary Fig. 8a). They subsequently also rated how visually similar they thought each pair of identities was (Supplementary Fig. 8b). Conceptual association strength was positively correlated with visual similarity (Pearson's correlation: $P < 0.001$ for all patients; two-tailed paired t-test of r against 0 for the group: $r = 0.40 \pm 0.18$ (mean ± s.d.), $t(4) = 4.93$, $P = 0.008$, $d = 1.76$, 95% CI: 0.17, 0.62; Fig. 2f). This was also the case for faces familiar to a particular patient. To further confirm this finding, we repeated our analyses with data from 40 control participants from the general population. Again, we found that participants' conceptual association ratings were correlated with their ratings of visual similarity ($r = 0.52 \pm 0.19$, $t(39) = 17.6$, $P < 10^{-19}$, $d = 2.72$, 95% CI: 0.46, 0.57; Fig. 2f). While the notable correlation suggests that conceptual associations can be explained by the measure of visual similarity to a



considerable degree (note that conceptual associations were always rated first to prevent patients from using visual similarity as a strategy to judge conceptual associations), a sizable degree of the variance (48–60%) in one cannot be explained by the other. Compatible with this finding, we found separate populations of M-ID neurons encoding conceptual associations and visual similarities (Fig. 2g; see Fig. 2h for a scatterplot of conceptual associations vs visual similarities; note that here visual similarity was indexed using DNN feature distance, but similar results were derived using visual similarity ratings; Supplementary Information). We confirmed that neurons encoding visual similarities (a subset of feature M-ID neurons) had a smaller feature distance (Fig. 2i; two-tailed paired *t*-test: $t(15) = 6.00$, $P = 2.44 \times 10^{-5}$, $d = 1.77$, 95% CI: 0.05, 0.11) and neurons encoding conceptual associations had a greater conceptual association ratings (Fig. 2j; $t(5) = 2.90$, $P = 0.034$, $d = 1.07$, 95% CI: 0.09, 1.55) for the pairs of encoded/selected identities compared with the other pairs, as expected. Interestingly, we found that neurons encoding visual similarities had even lower conceptual association ratings for the pairs of encoded identities (Fig. 2j; $t(15) = 4.36$, $P = 5.59 \times 10^{-4}$, $d = 1.21$, 95% CI: 0.29, 0.85), whereas neurons encoding conceptual associations had even larger feature distance (Fig. 2i; $t(5) = 5.11$, $P = 0.0037$, $d = 1.54$, 95% CI: 0.04, 0.11). We obtained a consistent result in neurons encoding both visual similarities and conceptual associations (Fig. 2i,j) and in neurons encoding neither visual similarities nor conceptual associations (Fig. 2i,j). Furthermore, we replicated our results using visual similarity ratings from the patients (Supplementary Fig. 8g–n) and the average (that is, consensus) ratings acquired from the general controls (Supplementary Fig. 8c–f,k–n). Together, although conceptual association strength was correlated with visual similarity, we found separate populations of M-ID neurons encoding conceptual associations and visual similarities. In particular, there were M-ID neurons whose response could only be explained by visual similarity.

Second, we used a web-association metric[18], in which the names of celebrities were paired in internet searches to determine the degree of their association based on search results (Supplementary Fig. 9a; see Methods for details). We restricted our analysis to the identities each patient rated as familiar, but similar results were found when we included all identities (both familiar and unfamiliar). We found that the web-association scores between pairs of encoded identities were not significantly greater than between the other pairs (Fig. 2k left; two-tailed paired *t*-test: $t(37) = 0.80$, $P = 0.43$, $d = 0.17$, 95% CI: −0.31, 0.13). This was also the case for non-feature M-ID neurons (Fig. 2k right; $t(53) = 0.76$, $P = 0.45$, $d = 0.09$, 95% CI: −0.07, 0.15). Web-association scores were correlated with conceptual association ratings of the patients (Supplementary Fig. 8o; $P < 0.05$ for all patients; $t(4) = 5.53$, $P = 0.0052$, $d = 1.97$, 95% CI: 0.06, 0.17) and general controls (Supplementary Fig. 8o; $t(39) = 14.8$, $P < 10^{-13}$, $d = 2.29$, 95% CI: 0.09, 0.11; Supplementary Fig. 8p; mean rating: $r(1,255) = 0.16$, $P = 4.5 \times 10^{-8}$), indicating that this web-based metric captured variance meaningful for our participants. Therefore, the encoding of visually similar identities by M-ID neurons was not likely explained by conceptual associations as measured by web-association scores.

Third, visual similarity ratings from both patients and general controls were correlated with DNN feature similarity (that is, the negative of the DNN feature distance; Fig. 2l and Supplementary Fig. 8q; $n = 1{,}000$ permutation runs), especially in the later layers. We also found that the DNN feature similarity (the negative of the full feature distance) was not correlated with the web-association scores in the later layers we analysed (Fig. 2m), and the pairwise distance in the *t*-SNE feature space was not correlated with the web-association scores ($P > 0.05$ for all layers; Supplementary Fig. 9b). This suggests that the organization of the feature space could not be explained by conceptual associations measured by the web-association scores.

Fourth, 6 of our recorded patients provided social trait judgements of the stimuli using a comprehensive set of social traits[24] (see Methods). We found that the pairs of identities encoded by the feature M-ID neurons were not more similar in social trait judgements than the other pairs (Fig. 2n; all $P > 0.05$ except for 'feminine' because the feature space was organized by gender (Fig. 1c), but feature M-ID neurons did not encode gender per se as shown above). Furthermore, we found that the absolute difference in social trait judgements between identity pairs was largely uncorrelated with visual similarity (Supplementary Fig. 8r) and conceptual association (Supplementary Fig. 8s), suggesting that our results could not be explained by the concept of social traits (that is, relatedness in social traits).

Fifth, we found that M-ID neurons encoding visual similarity (that is, feature M-ID neurons; 264 ms relative to stimulus onset) responded earlier than M-ID neurons encoding conceptual associations (387 ms; permutation $P = 0.007$; see Methods for differential latency analysis). We also observed that feature M-ID neurons (264 ms) responded earlier than S-ID neurons (438 ms; permutation $P < 0.001$; see also Supplementary Fig. 2g). Therefore, our results indicate that feature M-ID neurons are involved in an earlier stage of the processing hierarchy (see Discussion).

Lastly, using data from a publicly available dataset containing well-characterized identity neurons recorded in response to famous and/or familiar faces[19], we replicated the findings from our own dataset by again identifying distinct subsets of neurons that encoded visual facial features only, conceptual associations only, or both (Supplementary Information, and Supplementary Figs. 10 and 11).

Together, our results reveal two types of M-ID neurons: one type that responds to visually related identities (our new finding), and another that responds to conceptually related identities (extending previous work; see Discussion for a comprehensive summary).

A broader category of feature neurons

Different faces of the same identity were not clustered in the feature spaces formed by earlier and intermediate layers of the DNN (Supplementary Figs. 3c and 4a). We therefore next asked whether there were neurons that coded for sets of faces with similar visual features—as assessed by earlier DNN layers—independent of face identity. If so, such neurons would, by definition, be feature coding neurons. Using the same method to select identity neurons based on later DNN layers, we identified a broader category of 'feature neurons' that were tuned to specific regions of the feature space from each DNN layer (see Fig. 3a,b for examples and Fig. 3c for a summary; see Supplementary Fig. 7b,e for a breakdown of amygdala and hippocampal neurons), regardless of whether the neuron was an identity neuron.

Feature neurons mostly appeared in DNN layers where faces started to become clustered by identities. Therefore, feature neurons primarily encode higher-level visual information related to identification rather than lower-level image characteristics. Six of the later DNN layers (Conv5_2, Conv5_3, Pool5, FC6, FC7 and FC8) contained an above-chance number of feature neurons at the population level (Fig. 3c), and we restricted our analysis to these feature neurons. The number of identities (Fig. 3d) and faces (Fig. 3e) covered by the tuning region of feature neurons indicated the size of the 'receptive field' (in the feature space) of these feature neurons. The tuning region of each feature neuron covered approximately 2.5–11% of the feature space (Fig. 3f), and the total observed neuronal population covered approximately 61–87% of the feature space (Fig. 3g). With increasing levels of abstraction, tuning regions in later layers Pool5, FC6, FC7 and FC8 contained fewer identities (Fig. 3d; two-tailed two-sample *t*-test: $P < 0.001$ for all comparisons) but more faces (Fig. 3e; $P < 0.001$) compared with the preceding convolutional layers. Encoded faces were also more widely distributed in the FC layers than the preceding layers (Fig. 3h; two-tailed two-sample *t*-test on feature distance: $P < 0.0001$).

Although an appreciable proportion of feature neurons were also identity neurons, some feature neurons were not identity neurons (that is, neither S-ID nor M-ID neurons; red bars in Fig. 3c; in particular in

https://doi.org/10.1038/s41562-025-02218-1

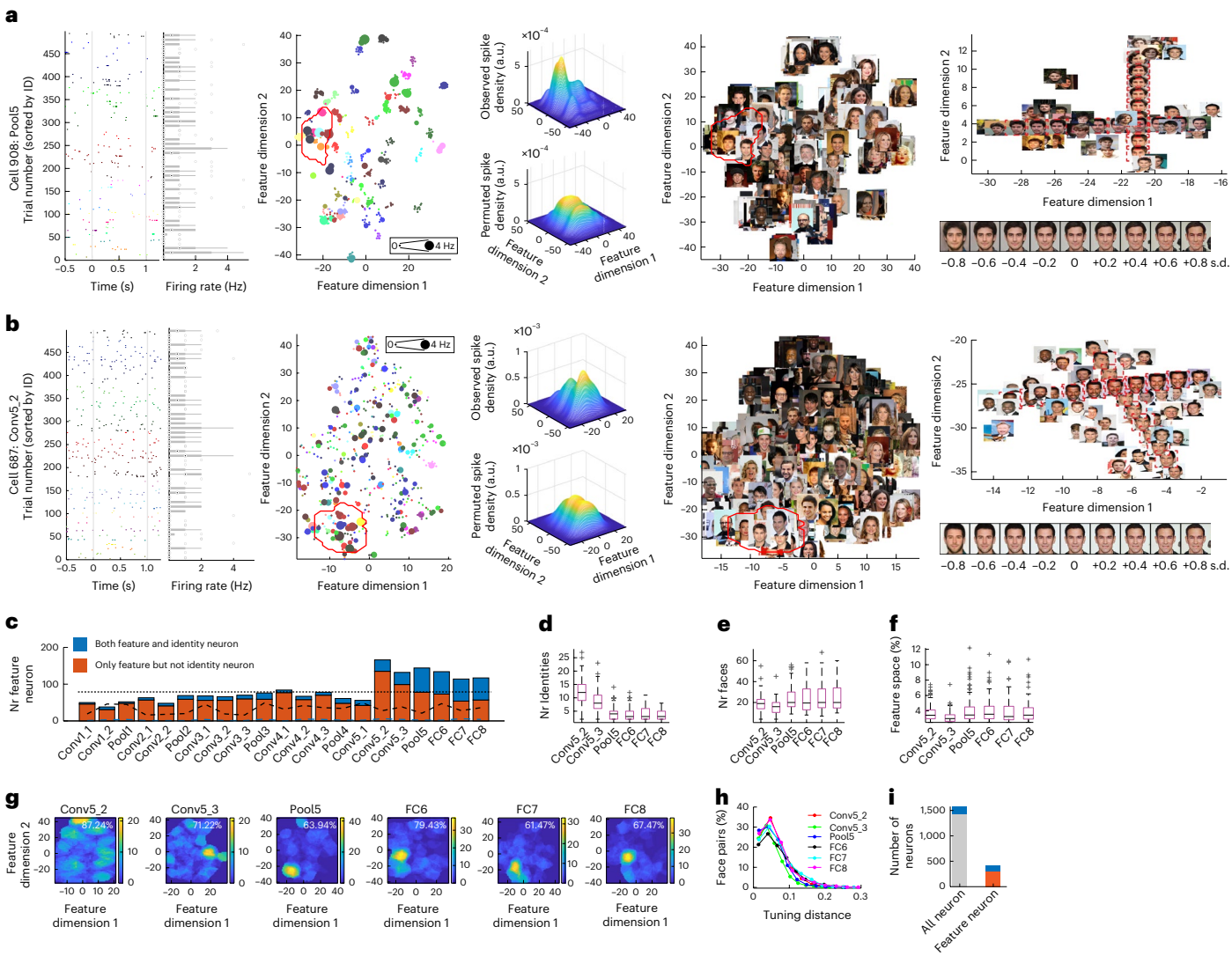


Fig. 3 | Characterization of feature neurons. a,**b**, Two example feature neurons that encoded visually similar faces. **a**, Cell 908. **b**, Cell 687. Legend conventions as in Fig. 1. **c**, The number of (Nr) feature neurons identified from each DNN layer. Blue denotes feature neurons that were also identity neurons. The black dashed line denotes the chance level for all feature neurons (estimated using a DNN with the same structure but random activations), and the blue dashed line denotes the chance level for feature neurons that were also identity neurons. The black fine-dashed line shows our cut-off threshold (set at 5%). **d**, The number of identities encoded by feature neurons. **e**, The number of faces encoded by feature neurons (that is, the number of faces that fell within the tuning region of a feature neuron). **d**–**f**, For each box, the central line is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points the algorithm considers to be not outliers, and the outliers are plotted individually. $n = 166, 132, 144, 134, 114, 117$ for Conv5_2, Conv5_3, Pool5, FC6, FC7 and FC8, respectively. **f**, Percentage of feature space covered by the tuning regions of feature neurons. **g**, The aggregated tuning regions of the neuronal population. **h**, Distribution of pairwise distance between faces in each feature neuron's tuning region. Legend conventions as in Fig. 2. **i**, The number of identity neurons in the whole population (left) and among feature neurons (right). Blue refers to the number of identity neurons ($n = 155$ for the whole population and $n = 104$ for feature neurons). Red denotes the number of non-identity feature neurons ($n = 312$). Grey is the number of non-identity neurons ($n = 1{,}422$).

convolutional layers; see Fig. 3b for an example) because they covered a region in the face space where faces from the same identity were not clustered. Therefore, identity selectivity was not necessary for feature-based coding. In other words, feature neurons can respond to faces that were adjacent in the feature space but were not from the same identity.

Were feature neurons more likely to be identity neurons (that is, either S-ID or M-ID neurons) than other neurons? Feature neurons had a higher proportion (104/417, 24.94%) of identity neurons compared with the entire neuronal population (155/1,577, 9.83%; χ^2-test: $P = 3.33 \times 10^{-16}$; Fig. 3i; note that feature neurons here included those from layer Conv5_2 and Conv5_3 even though identity neurons could in principle only emerge in layers with clustering of faces; see Supplementary Fig. 7c,f for a breakdown of amygdala and hippocampal neurons), suggesting that region-based coding is a key component in identity selectivity.

Notably, we assessed whether the response of MTL neurons could be understood as axis based[7,9] or norm based[25], as observed in non-human primate studies in the IT cortex. The axis-based code assumes that each neuron encodes one or a few specific feature dimensions (for example, texture, shape), while the norm-based code hypothesizes that neurons encode the relative distance between a given face and the prototype or average representation of faces stored in the brain. In contrast to the macaque IT cortex[7,9], only the response of a small and non-significant proportion of MTL neurons could be explained by an axis code (Supplementary Information and Supplementary Fig. 13). Similarly, in contrast to the macaque IT cortex[25], the response of MTL neurons could not be explained by

https://doi.org/10.1038/s41562-025-02218-1

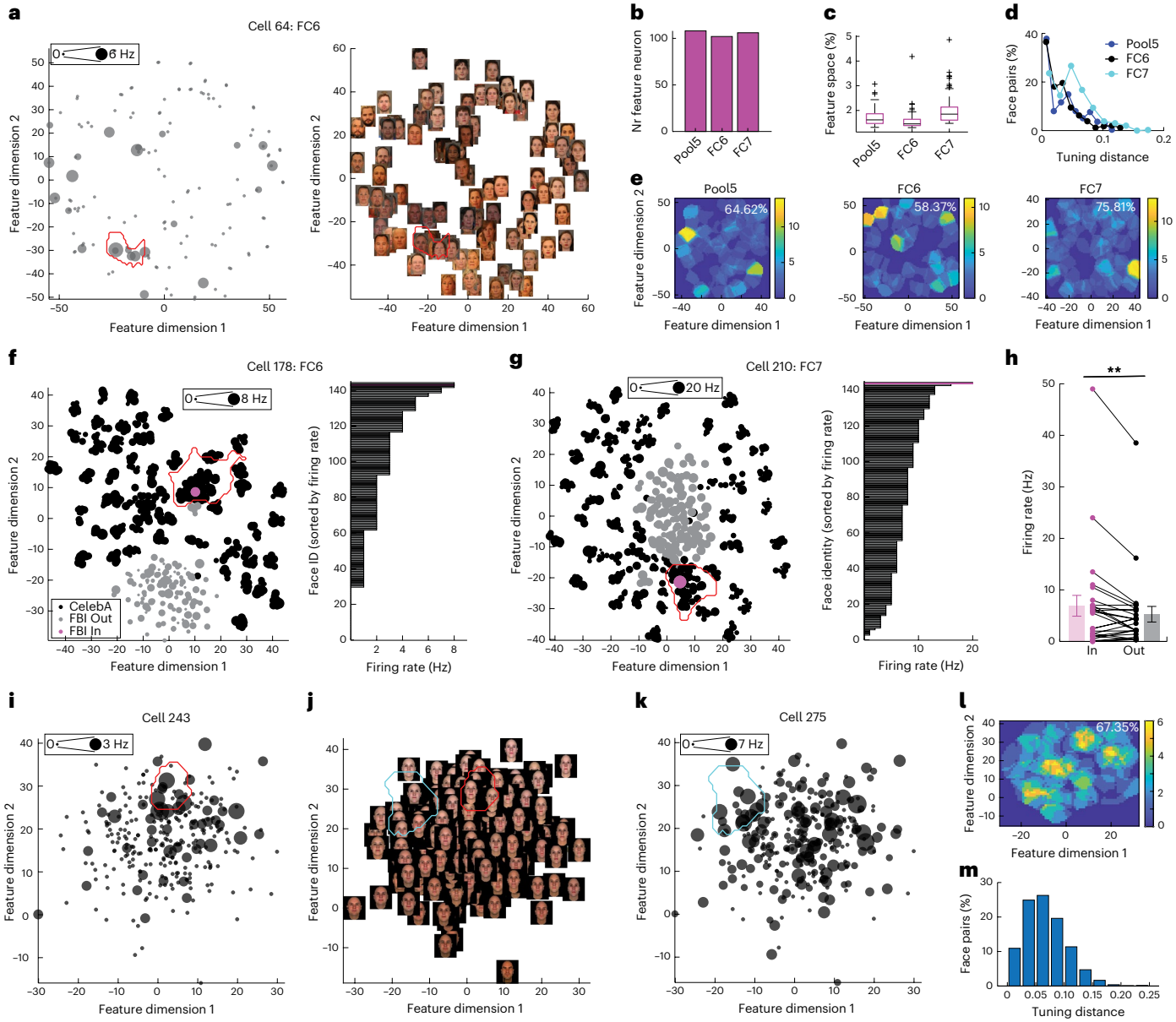


Fig. 4 | Validation and generalization of region-based feature coding with unfamiliar and model faces. a–**h**, Results from the FBI Twins dataset. **i**–**m**, Results from the FaceGen dataset. **a**, An example neuron demonstrating region-based feature coding. The size of the dot indicates the firing rate. The red outline delineates the tuning region of the neuron in the feature space. **b**, The number of identified feature neurons using the FBI stimuli. Only DNN layers with an above-chance number of feature neurons (based on our simulations) are shown. **c**, Percentage of feature space covered by the tuning regions of feature neurons. **d**, Distribution of pairwise distance between faces in each feature neuron's tuning region. **e**, The aggregated tuning regions of the neuronal population. Legend conventions as in Fig. 2. **f**,**g**, Example CelebA feature neurons showing elevated responses for FBI stimuli falling in their tuning regions. The feature spaces were constructed for combined CelebA and FBI stimuli. The size of the dot indicates the firing rate. The red outline delineates the tuning region of the neuron (identified by the CelebA stimuli). Black dots are faces from the CelebA stimuli. Grey dots are faces from the FBI stimuli. Magenta dots are FBI stimuli falling in the tuning region of the neuron. Note that in the face feature space combining the CelebA and FBI stimuli, the clustering based on gender and skin colour was retained and similar to the feature spaces using the CelebA stimuli only. **f**, Cell 178. **g**, Cell 210. **h**, Population results comparing neuronal responses to the FBI stimuli falling in vs out of the tuning region ($n = 26$). Each dot represents a neuron. Error bars denote ±s.e.m. across neurons. Asterisks indicate a significant difference between In vs Out responses using a one-tailed paired t-test (**$P < 0.01$). **i**–**k**, Two example neurons demonstrating region-based feature coding. Note that the feature space was constructed using parameters (that is, features) used to synthesize the faces rather than DNN features. The dimensions of the feature space are the first shape and tone/texture principal components (PCs) used to generate the stimuli. Note that face shape varied along Feature Dimension 1, and skin colour varied along Feature Dimension 2. **i**, Cell 243. **j**, Feature space. **k**, Cell 275. The red and cyan outlines delineate the tuning regions. **l**,**m**, Population summary ($n = 61$). **l**, The aggregated tuning regions of the neuronal population. **m**, Distribution of pairwise distance between faces in each feature neuron's tuning region. Legend conventions as in Fig. 2.

a norm-based code either (Supplementary Information and Supplementary Fig. 14a–c). Furthermore, responses in the human MTL and macaque IT were best explained by different processing stages in our DNN (Supplementary Fig. 13n,o), and region-based feature coding could be observed in the norm polar face space as well (Supplementary Fig. 14d–i).

Lastly, we tested the dependence of the feature space and stimuli in identifying feature neurons (Supplementary Information). Feature

neurons could be identified using other face recognition DNNs (Supplementary Information). Although to some extent region-based feature coding could still be observed in feature spaces constructed using DNNs trained for object recognition (for example, AlexNet (Supplementary Fig. 15a,b) and ResNet (Supplementary Fig. 15e–h)), there were fewer feature neurons (Supplementary Information), suggesting that MTL neurons were sensitive to the organization of the feature space; hence the organization of the feature space played a critical role in identifying feature neurons. In addition, by projecting non-face stimuli onto the AlexNet (Supplementary Fig. 15c,d), ResNet (Supplementary Fig. 15i–l) or the original VGG-Face (Supplementary Fig. 15m–p) feature spaces, we found that faces had unique visual features, and feature neurons encoded higher-level visual features related to faces rather than lower-level visual features that might be common between face and non-face stimuli.

Together, by surveying each DNN layer for region-based feature coding independent of identity coding, we revealed a broader category of feature neurons. In particular, feature neurons derived from the Conv5_2 and Conv5_3 feature spaces, where faces of the same identity are not clustered, do not qualify as identity neurons. This suggests that the encoding of visual features is independent of the identity-based encoding of conceptual associations.

Region-based feature coding predicts responses to novel stimuli

We conducted two additional experiments to assess whether region-based feature coding generalized to new sets of novel images. First, we recorded from 837 neurons in the same 10 patients (27 sessions; firing rate > 0.15 Hz; accuracy = 75.7% ± 23.0%, mean ± s.d. across sessions) using face stimuli from the FBI Twins dataset (Fig. 4a), which were all novel to our patients. We applied the same DNN to extract facial features and construct feature spaces. We again found region-based feature coding by single neurons in this experiment (see Fig. 4a and Supplementary Fig. 16a,b for examples, and Fig. 4b–e for group results). Notably, we recorded the responses of a subset of the same 699 neurons to both the CelebA stimulus set (which we used in our principal experiment) and the FBI Twins dataset to directly investigate the generalizability of region coding across these two tasks. In the common feature space for the CelebA and FBI stimuli, the tuning region of 26 (out of 146) CelebA feature neurons overlapped with identities from the FBI stimuli (Fig. 4f,g and Supplementary Fig. 16c–e). Across all 26 neurons, FBI stimuli within the CelebA feature neurons' tuning regions elicited significantly greater responses than FBI stimuli outside those regions (right-tailed paired t-test: $t(25) = 2.61$, $P = 0.0076$, $d = 0.18$, 95% CI: 0.56, ∞; see Fig. 4f,g for examples and Fig. 4h for group results). This demonstrates that region-based feature coding generalized across different image sets and to novel stimuli unfamiliar to the participants (see also Supplementary Information).

Second, we recorded from a separate population of 658 neurons (25 sessions from 6 patients; firing rate > 0.15 Hz) while patients performed a trustworthiness or dominance judgement task using FaceGen model faces (Fig. 4i–m)[26], which contained only feature information but little identity information. Behaviourally, the ratings from our patients were consistent with the consensus ratings[26] (Pearson correlation: $r = 0.23 \pm 0.14$ (mean ± s.d.) across sessions for trustworthiness and $r = 0.39 \pm 0.18$ for dominance; two-tailed t-test against 0: both $P < 0.001$), indicating that they were performing the task as instructed. Again, we found region-based feature coding (61 neurons, above chance; each neuron covered 2.50% ± 0.66% (mean ± s.d.) of the feature space; see Fig. 4i–k for examples and Fig. 4l,m for group results). FaceGen model faces contain little conceptual information about face identities, suggesting that region-based feature coding likely does not depend on conceptual associations between face identities.

Together, these additional experiments suggest that region-based feature coding can generalize to new and unfamiliar face stimuli and may operate independently of conceptual associations.

Discussion

Our results reveal that a subset of identity neurons in the human amygdala and hippocampus encodes face identities that are related visually (that is, faces sharing similar features) rather than conceptually (for example, Bill and Hillary Clinton). We further show a broader category of feature neurons exhibiting region-based feature coding. The response of feature neurons depends neither on identity selectivity nor on face familiarity, gender, race or low-level features, and their tuning regions can be validated using new face stimuli. Lastly, we show that in contrast to the macaque IT cortex, feature-based coding in the human MTL is primarily region based rather than axis based.

The amygdala and hippocampus are downstream of the face-selective regions in the higher visual cortex, where feature-based coding for faces is first evident[11–13]. Despite being downstream from face-selective areas, in the human MTL no feature-based encoding of faces has been reported so far. Instead, only exemplar-based coding has been demonstrated[16,17]. This indicates that the format of the neural representation is fundamentally different in the MTL compared with the higher visual cortex. The key finding of our present study is that visual feature-based coding of faces persists in the MTL, a brain region located downstream of the visual ventral stream, where high-level visual processing is thought to be completed. In particular, our results suggest a possible mechanism for how the brain transitions from a perception-driven representation of features in the higher visual cortex to a memory-driven representation of semantics in the MTL: MTL neurons receive processed visual information from the higher visual cortex (specifically, axes of the face feature space encoded by axis-coding cells[7]) and encode a region in the high-level feature space such that they become selective to the identities that fall into this region, thereby providing a bridge between the feature-based and exemplar-based coding mechanisms. We hypothesize that region-coding neurons serve as a basis for semantic representations in the MTL, which in turn are the basis for declarative memory[27]. Indeed, our latency analysis supports the proposed processing hierarchy: the encoding of visual features precedes the encoding of concepts (feature M-ID neurons responded earlier than M-ID neurons encoding conceptual associations) and feature-based coding is earlier than exemplar-based coding (feature M-ID neurons responded earlier than S-ID neurons). Therefore, our findings bridge the two extreme hypotheses by revealing an intermediate region-based feature code in the MTL.

Although it may not be possible to entirely discount any sort of conceptual association in explaining our results, our results could not be simply attributed to conceptual associations of identities. First, feature neurons encode unfamiliar identities (shown by both a subset of CelebA stimuli and all FBI stimuli that were unfamiliar to the patients), for which patients cannot have formed much conceptual knowledge; feature coding exists in DNN layers where faces of the same identity are not clustered; and feature coding could not be explained by simple conceptual knowledge of race, gender or age (that is, cross-race, cross-gender or cross-age effects could not explain our results), or social trait judgements. Feature-based coding of identities was even observed for synthetic model faces, of which patients had little conceptual knowledge. Second, when patients had a considerable knowledge of the identities (Supplementary Figs. 10 and 11; all familiar stimuli), we observed a dissociation of coding of visually similar identities and coding of conceptually associated identities, consistent with our results using the CelebA stimuli (Fig. 2; including both familiar and unfamiliar stimuli). Furthermore, conceptual association measured using web-association scores neither explained feature-based coding nor correlated with DNN features.

On the other hand, coding visual features in the human MTL is not mutually exclusive from but may complement the well-known coding of conceptual associations[18,19]. Indeed, we found separate subsets of neurons that encoded either visual features or conceptual associations, as well as neurons in which both measures explained response variance (Fig. 2g,h; neurons encoding both measures may represent an intermediate step between feature-only and concept-only neurons). Furthermore, when we analysed an independent dataset with well-characterized identity neurons[19] (Supplementary Figs. 10 and 11), we also found feature-based coding in addition to the coding of conceptual associations, highlighting an under-appreciated role of MTL neurons in these previous studies[18,19]. Therefore, MTL neurons embody two forms of coding of face identities that complement each other. Moreover, concept neurons and encoding of conceptual associations have also been found to be prominent in non-face stimuli[17–19]; thus, it is necessary to investigate feature-based coding in a broader object space (see ref. 28). Furthermore, identity coding interacts with face familiarity coding[29], a topic that requires further investigation.

Neurons in the human MTL have been shown to demonstrate prominent categorical responses to visual objects (that is, visual selectivity)[30] and facial expressions of emotions (that is, emotion selectivity)[31,32]. Region-based feature coding may also provide an account for visual and emotion selectivity: objects or emotions falling within the coding region of a neuron may elicit an elevated response. A future direction will be to construct the feature space for objects in general (for example, using the convolutional neural network AlexNet) and investigate region-based feature coding in this feature space. A face feature space in the MTL also supports the hypothesis that cognitive maps in the hippocampus[33] may generalize across diverse dimensions in life experiences[34].

Previous research on identity neurons primarily used familiar faces[16,18,21]. In the present study, we found that region-based feature coding of face identity was independent of face familiarity, similar to feature coding by primate IT neurons that even encode computer-generated faces[7,9]. It is also worth noting that in contrast to the traditional axis-based face spaces where axes of the space and coordinates of faces are fixed[7,25], the feature space constructed by *t*-SNE in the present study varies as a function of the input stimuli because it models the similarity between all input stimuli (but note that our results were robust to the construction of the feature space and could be replicated using Euclidean distance of full DNN features). Therefore, our observed feature neurons in the human MTL may demonstrate a form of similarity-based or manifold-based coding (that is, finding meaningful low-dimensional structures hidden in the high-dimensional observations using nonlinear dimensionality reduction)[35,36], which may in turn support the MTL's critical role in face recognition, classification and memory.

Methods

Patients

There were 38 sessions across 12 patients in total (Supplementary Tables 1 and 2). All participants provided written informed consent using procedures approved by the Institutional Review Boards of West Virginia University (WVU) and Washington University in St. Louis (WUSTL). Patients were not financially compensated for participating in this study.

No statistical methods were used to pre-determine sample sizes, but our sample sizes are larger than those reported in previous publications[32,37–40]. This study was not preregistered. Participants were not grouped and hence no randomization was performed.

Stimuli

We used faces of celebrities from the CelebA dataset[41]. We selected 50 identities with 10 images for each identity, totalling 500 face images (Fig. 1a). To avoid the confounding effects of cross-race or cross-gender influences, our selected stimuli included both genders (33 of the 50 identities were male) and multiple races (40 identities were Caucasian, 9 identities were African-American and 1 identity was biracial). We also ensured sufficient variance in both visual appearance (for example, gender, race, age) and semantic information (for example, level of fame, familiarity) of the selected identities. We used the same stimuli for all patients.

All images had the same resolution, and the faces had a similar size and position in the images[41]. The DNN (see below for details) could well recognize the faces (98.20 ± 4.38% accuracy, mean ± s.d. across identities); however, the DNN was not able to recognize faces based only on the background (3.84 ± 11.12% accuracy; in contrast, the accuracy for cropped faces was 97.40 ± 5.27%). We also confirmed that the DNN had a similar accuracy in recognizing feature M-ID neurons' selected identities (98.05 ± 3.33%) vs non-selected identities (98.89 ± 4.59%; two-tailed two-sample *t*-test: $t(48) = 0.52$, $P = 0.61$, $d = 0.19$, 95% CI: −0.04, 0.02). Furthermore, although African-American faces had a lower brightness, within each race group there was no significant difference between selected identities (Caucasian: 117.18 ± 39.00; African-American: 97.53 ± 43.40) vs non-selected identities (Caucasian: 119.60 ± 37.15; African-American: 105.51 ± 39.91; Caucasian: $t(398) = 0.47$, $P = 0.10$, $d = 0.06$, 95% CI: −12.42, 7.59; African-American: $t(88) = 0.55$, $P = 0.73$, $d = 0.18$, 95% CI: −36.68, 20.72). Therefore, the response of feature M-ID neurons could not simply be explained by image resolution, brightness or background. Lastly, we asked patients to indicate whether they were familiar with each identity in a follow-up survey. We confirmed with the patients that they were able to visually distinguish different identities and recognize those with which they were familiar. It is also worth noting that in the present study we focused on the coding of 'identities' rather than 'concepts'. Therefore, we did not expect the patients to be familiar with all identities and we did not inquire about their depth of knowledge regarding the identities, simply whether they recognized them (that is, face familiarity).

We further used two validation datasets (Fig. 4). First, we used a newly collected FBI Twins Dataset that included pairs of coloured photos with the following relationships: identical twins (IT), mirror twins (MT), fraternal twins (FT), mother–child (MC), father–child (FC) and spouses (SP). Therefore, this dataset contained faces with various levels of similarity, and all faces from this dataset were unfamiliar to the patients. The photographing conditions were well controlled to ensure similar background and lighting, and all photos were of high resolution (3,840 × 5,760 pixels). There was one face per identity and a total of 144 faces.

Second, we used a FaceGen Dataset with model faces, which notably contains only feature information but no real identity information. We used the FaceGen Modeller programme (v.3.1, http://facegen.com) to randomly generate 300 faces (see ref. 26 for detailed procedures). FaceGen constructs face space models using information extracted from 3D laser scans of real faces. To create the face space model, the shape of a face was represented by the vertex positions of a polygonal model of fixed mesh topology. With the vertex positions, a principal component analysis (PCA) was used to extract the components that accounted for most of the variance in face shape. Each principal component (PC) thus represented a different holistic non-localized set of changes in all vertex positions. The first 50 shape PCs were used to construct faces that had a symmetric shape. Similarly, because skin texture is also important for face perception, 50 texture PCs based on PCA of the red, green and blue (RGB) values of the faces were also used to represent faces. The resulting 300 faces were randomly generated from the 50 shape and 50 skin texture components with the constraint that all faces were set to be Caucasian. It is worth noting that each PC is a feature dimension of the face space.

Conceptual association and visual similarity ratings of our stimuli

After the neural recordings, patients were asked to rate the conceptual association and visual similarity between each pair of the 50 CelebA

identities (n = 1,225 in total) using a 7-point Likert scale. Ratings of conceptual association and visual similarity were collected in separate sessions lasting 20 to 30 min each. In each trial, two different identities were presented to the participants side by side. For the rating of conceptual associations, patients were instructed to assess "How conceptually related do you think each pair of identities is?". For visual similarity ratings, patients were instructed to evaluate "How visually similar do you think each pair of identities is?". We also collected conceptual association and visual similarity ratings from 45 participants from the general population via Prolific to obtain normative data.

To evaluate the internal consistency and reliability of the rating measures[42], we estimated the intraclass correlation coefficient (ICC) using a two-way random-effects model for consistency across mean ratings. The ICC was calculated using the 'ICC(2k)' function in the 'psych' package in R. We found that both the ratings of visual similarity (ICC = 0.93, d.f. = 1,223, lower bound = 0.91, upper bound = 0.94) and conceptual association (ICC = 0.95, d.f. = 1,223, lower bound = 0.94, upper bound = 0.95) were highly consistent across participants.

Social trait judgement ratings of our stimuli

To understand whether social trait judgements could explain our results, we used a set of social traits that most comprehensively characterize social trait judgements[24], including warm, critical, competent, practical, feminine, strong, youthful and charismatic (Fig. 2n and Supplementary Fig. 8r–t). These social traits represent the four core psychological dimensions of comprehensive trait judgements of faces (warmth, competence, femininity and youth; 2 traits per dimension), and they were well validated in the previous study[24]. Patients were asked to rate the faces on 8 social traits using a 7-point Likert scale in an online rating task. We further acquired social trait ratings of the faces from 500 participants from the general population (age: 26.20 ± 7.11 years (mean ± s.d.), 180/500 females). Identical to the neurosurgical patients from the current study, online participants rated the faces on 8 social traits using a 7-point Likert scale.

Experimental procedure

We used a one-back task for CelebA and FBI stimuli. In each trial, a single face was presented at the centre of the screen for a fixed duration of 1 s, with a uniformly jittered interstimulus interval (ISI) of 0.5–0.75 s (Fig. 1b). Each image subtended a visual angle of approximately 10°. Patients pressed a button if the present face image was identical to the immediately previous image. Approximately 9% of the trials were one-back repetitions. Each face was shown once, except when repeated in one-back trials. We excluded all one-back trials to have an equal number of trials for each face. This task kept patients attending to the faces but avoided potential biases from focusing on a particular facial feature (for example, compared to asking patients to judge a particular facial feature). The order of faces was randomized for each patient. This task procedure has been shown to be effective to study face representation in humans[43].

For FaceGen stimuli, patients performed two face judgement tasks. In each task, there was a judgement instruction, that is, patients judged how trustworthy or how dominant a face was. We used a 1–4 scale: '1': not trustworthy/dominant at all, '2': somewhat trustworthy/dominant, '3': trustworthy/dominant and '4': very trustworthy/dominant. Each image was presented for 1.5 s at the centre of the screen. One patient performed an additional passive-viewing task. We combined data from all tasks for analysis.

Although judgement instructions might potentially impact the response of MTL neurons (but note that all responses of each neuron were under the same instruction), it provided an opportunity to investigate whether feature-based coding could be generalized to tasks with explicit judgement instructions rather than passive viewing only. Notably, we further used the consensus ratings of the stimuli from ref. 26 to dissociate the coding of facial trustworthiness/dominance from feature-based coding of visual similarity. Although faces within a feature neuron's tuning region were visually similar and thus had similar facial trustworthiness/dominance, we found that other non-selected faces could be similarly trustworthy/dominant in comparison to the selected faces. In other words, faces with similar facial trustworthiness/dominance were not all clustered within a feature neuron's tuning region but distributed in the feature space. Therefore, feature neurons did not encode facial trustworthiness or dominance per se, and feature-based coding could be generalized to tasks with explicit judgement instructions.

Stimuli were presented using MATLAB with Psychtoolbox 3 (ref. 44) (http://psychtoolbox.org) (screen resolution: 1,600 × 1,280 pixels).

Feature extraction and construction of feature space

We used the well-known DNN implementation based on the VGG-16 convolutional neural network (CNN) architecture[45] to extract features for each face image (see Supplementary Fig. 3a for details). The same network was also used in recent work[43] as the computational model for deep face feature extraction. The VGG-16 model was pre-trained with the VGG-Face dataset, and we refer to it here as the VGG-Face model.

To confirm the model's ability to discriminate between different identities and to ensure its suitability as a feature extractor, we performed fine tuning of the network. Initially, we fine tuned the FC8 layer exclusively. During this fine-tuning process, we utilized all images of the 50 identities in the CelebA dataset, with each identity having 16–30 images. In addition, we modified the output layer of our model to have 50 units, corresponding to the number of identities in the dataset. For the training process, we divided the stimuli, with two-thirds serving as the training set and the remaining stimuli as the testing set. We utilized the Adam optimizer with an initial learning rate of 5×10^{-4} for the training process, which consisted of 10 epochs in total. To facilitate the convergence of the loss function, we applied a learning rate scheduler after each epoch, updating the learning rate by multiplying a decreasing factor ($\Upsilon = 0.9$). During the fine-tuning process, we updated the weights by computing the cross-entropy loss on random batches of four face images, which were scaled to 224 × 224 pixels for back propagation. To evaluate performance, we employed 5-fold cross-validation, which yielded an accuracy of approximately 95%. The high accuracy in identity discrimination suggested that the network effectively extracted all relevant features and was suitable for further analysis. Therefore, we focused solely on fine tuning the FC8 layer while keeping all other layers frozen, utilizing this transferred model for subsequent feature extraction. Overall, our fine-tuning approach allowed us to validate the discriminative ability of the pre-trained model, leverage its extracted features and consider the sufficiency of the network's top layer for our analysis.

It is worth noting that our DNN was trained on both Caucasian and African-American identities, and our DNN could recognize both races equally well (Caucasian: 98.05% ± 4.59%; African-American: 98.89% ± 3.33%; two-tailed two-sample t-test: $t(48) = 0.52$, $P = 0.61$, $d = 0.19$, 95% CI: −0.04, 0.02). Importantly, we used the same parameters for all identities, and the DNN had no knowledge about the race of the identities. Therefore, the organization of the feature space and clustering of identities with similar brightness or race was entirely derived from the input images and learning by the DNN.

We subsequently applied the t-SNE algorithm to project high-dimensional features into a two-dimensional feature space (Fig. 1a,c). t-SNE is a variation of stochastic neighbour embedding (SNE)[46], a commonly used method for multiple-class high-dimensional data visualization[47]. We applied t-SNE for each layer, with the cost function parameter (Perp) of t-SNE representing the perplexity of the conditional probability distribution induced by a Gaussian kernel. Perplexity determines the balance between preserving local structure and capturing global structure in the low-dimensional embedding. The formula for perplexity is the exponent of the average log likelihood

of the probabilities of each face neighbouring relative to each of the other faces. Because a sparse distribution of faces could lead to a larger tuning region, we adjusted the distribution of faces using the *t*-SNE perplexity parameter so that the faces were distributed approximately homogeneously (Supplementary Fig. 4). Notably, we conducted a robustness analysis by varying the perplexity parameter and examining the resulting number of feature neurons (Supplementary Fig. 4b–d). We found that the number of selected feature neurons in the layer FC7 was very stable across perplexity parameters, confirming that our results were robust to the perplexity parameter. In addition, we conducted a robustness analysis by varying the size of the Gaussian kernel and examining the resulting number of feature neurons (Supplementary Fig. 4e). We found that the number of selected feature neurons was stable across kernel sizes and did not drop except at extreme values. This analysis confirmed that our results were robust to the size of the Gaussian kernel.

It is worth noting that neither feature extraction nor construction of the feature spaces utilized any information from neurons. Therefore, clustering of neurally encoded identities in the feature spaces was not by construction.

Electrophysiology

We recorded, using implanted depth electrodes in the amygdala and hippocampus, from patients with pharmacologically intractable epilepsy. Target locations in the amygdala and hippocampus were determined by the neurosurgeon solely on the basis of clinical need and verified using post-implantation computed tomography (Supplementary Table 1). At each site, we recorded from eight 40-µm microwires inserted into a clinical electrode as described previously[48,49]. Efforts were always made to avoid passing the electrode through a sulcus and its attendant sulcal blood vessels; hence the location varied but was always well within the body of the targeted brain area. Microwires projected medially out at the end of the depth electrode, and examination of the microwires after removal suggests a spread of approximately 20–30 degrees. Bipolar wide-band recordings (0.1–9,000 Hz), referenced to one of the eight microwires, were sampled at 32 kHz and stored continuously for offline analysis using a Neuralynx system. The raw signal was filtered with a zero-phase lag 300–3,000 Hz bandpass filter, and spikes were sorted using a semi-automatic template matching algorithm as described previously[50]. Units were carefully isolated, and recording and spike sorting quality were assessed quantitatively (Supplementary Fig. 1). Although there might be dependency between neurons to some extent, as a convention in the vast literature of human single-neuron recordings[51], neurons from each individual recording session were considered independent even if they were from the same patient.

Consistent with our previous studies[32,37–40], only single units with an average firing rate of at least 0.15 Hz throughout the entire task were considered. Trials were aligned to stimulus onset. For CelebA and FBI stimuli, we used the mean firing rate in a time window of 250–1,250 ms after stimulus onset as the response to each face. For FaceGen stimuli, we used the mean firing rate in a time window of 250–1,750 ms after stimulus onset as the response to each face.

Neural recordings from a monkey

One male rhesus macaque (*Macaca mulatta*) was used in this study. All procedures conformed to local and US National Institutes of Health guidelines, including the US National Institutes of Health Guide for Care and Use of Laboratory Animals. All experiments were performed with the approval of the MIT Institutional Animal Care and Use Committee.

The monkey passively viewed the original CelebA stimuli. In each trial, the monkey first viewed a white central fixation point (0.2 degrees of visual angle (DVA)) on a grey background for 300 ms to initiate a trial. Then, 8 faces were presented for 100 ms each, each followed by a blank (grey) screen for an ISI of 100 ms. The central fixation point persisted through the trial, and fluid reward was given if the monkey successfully fixated through the entire trial. The intertrial interval of blank grey screen was at least 500 ms. We recorded 4,155 trials in total, and we rejected 666 trials where the monkey broke the fixation (±2 DVA). For each round of presentation, we generated a random sequence for the 500 faces, and we used different sequences for different rounds of presentation. On average, each face was presented 55.7 ± 1.49 (mean ± s.d.) times. Note that we randomly inserted one grey image in each round of presentation as a control stimulus for baseline subtraction and normalization.

The monkey was chronically implanted with two Utah arrays (Blackrock Microsystems) in the anterior and central IT cortex (see refs. 52,53 for details). Each array consisted of one 10 × 10 electrode grid with 96 active iridium oxide electrodes. Each electrode was 1.5 mm long with an inter-electrode distance of 400 µm. During each recording session, bandpass filtered (0.1–7,500 Hz) neural activity was recorded continuously at a sampling rate of 20 kHz using an Intan Recording Controller (Intan Technologies). We detected the multi-unit spikes after the raw data were zero-phase bandpass filtered between 300–6,000 Hz (Matlab ellip function, fourth order with 0.1 decibel pass-band ripple and 40 dB stop-band attenuation), and we used multi-unit activity (MUA) for analyses. A multi-unit spike event was defined as the threshold crossing when voltage (falling edge) deviated by more than 3 s.d. of the raw voltage values. We estimated internal consistency for each channel using a standardized image set that was run before the recording session on the same day and we accepted 53 MUA channels (from two arrays) that showed sufficient internal consistency (>0.6). Consistent with previous studies[52,53], we used the mean firing rate in a time window of 70–180 ms after stimulus onset as the response to each face. We averaged the response from repeated presentations for each face.

Selection of identity neurons

To select identity neurons (Figs. 1 and 2), we first used one-way ANOVA to identify neurons with significantly unequal responses to different identities. We next imposed an additional criterion to identify the 'selected identities': the neural response of an identity was 2s.d. above the mean of neural response from all identities (note that the mean neural response could be considered as a baseline for the epochs when images were shown, even if we did not subtract a baseline). These identified identities whose response stood out from the global mean were the encoded identities. We refer to the neurons that encoded a single identity as S-ID neurons and those that encoded multiple identities as M-ID neurons.

We were able to select a similar set of neurons using the same criteria from previous studies[18,19]. Note that because identity neurons may respond to only a few stimuli, an overall response to stimulus onset might not be observed. Therefore, given such sparseness of firing of MTL neurons, we did not impose face responsiveness (overall change of activity in response to stimulus onset compared to baseline) as a criterion for neuron selection.

Selection of feature neurons

To select feature neurons (Figs. 3 and 4), we first estimated a continuous spike density map in the feature space by smoothing the discrete firing rate map using a 2D Gaussian kernel (kernel size = feature dimension range × 0.2, s.d. = 4). We then estimated statistical significance for each pixel by permutation testing: in each of the 1,000 runs, we randomly shuffled the labels of faces. We calculated the *P* value for each pixel by comparing the observed spike density value to those from the null distribution derived from permutation. We applied a mask to exclude pixels from the edges and corners of the spike density map where there were no faces because these regions were susceptible to false positives given our procedure. We lastly selected the region with significant pixels (permutation $P < 0.01$, cluster size > 2.5% of the pixels within the mask; note that in Fig. 2a we did not apply the threshold for minimum

cluster size for S-ID and non-feature M-ID neurons to compare across different categories of identity neurons). If a neuron had a region with significant pixels, the neuron was defined as a 'feature neuron' and demonstrated 'region-based feature coding'. We selected feature neurons for each individual DNN layer. Because the distribution of faces was more sparse for the FBI stimuli, we used a larger kernel (kernel size = feature value range × 0.3) and a lower threshold for cluster size (>1.6% of the pixels within the mask). Note that when we constructed the FBI face space, we also included additional faces for each identity in different viewpoints (5 faces in total) to stabilize the *t*-SNE projection. Lastly, given the configuration of the FaceGen face space, we considered clusters whose size was greater than 1.8% of the total number of pixels of the face space. Qualitatively the same results were derived when we used different thresholds to define feature neurons.

Conceptual association neurons and visual similarity neurons

For each neuron, we calculated the mean score between pairs of identities to which the neuron was selective (S–S), and between pairs in which the neuron was selective to one identity but not the other (S–NS). We selected conceptual association neurons or visual similarity neurons from M-ID neurons by comparing conceptual association ratings or DNN feature distances between the S–S pair vs all S–NS pairs using a two-tailed paired *t*-test ($P < 0.01$). Among a total of 62 neurons from 5 patients who provided conceptual association and visual similarity ratings, we found a subset of 16 neurons that only encoded visual features (note that to be comparable with the selection of conceptual association neurons, we here selected visual feature neurons by comparing feature distance between S–S vs S–NS pairs, but the selected visual feature neurons substantially overlapped with the feature M-ID neurons as we described above), a subset of 6 neurons that only encoded conceptual associations, a subset of 21 neurons that encoded both visual features and conceptual associations, and a subset of 19 neurons that encoded neither visual features nor conceptual associations. Notably, in Fig. 2g, the feature-only neurons (blue) were part of the feature M-ID neurons (blue and yellow), which also included neurons encoding both visual similarities and conceptual associations (yellow).

Identity selectivity index

To assess each neuron's selectivity to different identities, we defined an identity selectivity index (Supplementary Fig. 2b) as the *d′* between the most-preferred and least-preferred identities:

$$\text{Identity selectivity index} = \frac{\mu_{\text{best}} - \mu_{\text{least}}}{\sqrt{\frac{1}{2}\left(\sigma^2_{\text{best}} + \sigma^2_{\text{least}}\right)}} \quad (1)$$

where μ_{best} and μ_{worst} denote the mean firing rate for the most-preferred and least-preferred identities, respectively, and σ^2_{best} and σ^2_{worst} denote the variance of firing rate for the most-preferred and least-preferred identities, respectively. A similar index was used in previous studies to assess the level of selectivity to different faces[43]. It is worth noting that the identity selectivity index was not used to select identity neurons or estimate the number of neurons that were identity selective. Instead, the identity selectivity index was used to quantify the degree of identity selectivity for the identity and non-identity neurons that had already been selected.

Response ratio

Response ratio (Supplementary Fig. 2c,d) was calculated for each identity by first dividing by the response of the most-preferred identity and then ranking the identities from the most preferred to the least preferred. The response ratio of the most-preferred identity is thus 1. We compared the response ratios for each ordered identity between S-ID/M-ID vs non-identity neurons using a two-tailed two-sample *t*-test (corrected for multiple comparisons using false discovery rate (FDR)[54]). A steeper change from the best to the worst identity indicates a stronger identity selectivity.

Depth of selectivity index

To summarize the response of identity neurons, we quantified the depth of selectivity (DOS) for each neuron (Supplementary Fig. 2e): $\text{DOS} = \frac{n-(\sum_{j=1}^{n} r_j)/r_{\max}}{n-1}$, where n is the number of identities ($n = 50$), r_j is the mean firing rate to identity j and $r_{\max}$ is the maximal mean firing rate across all identities. DOS varies from 0 to 1, with 0 indicating an equal response to all identities and 1 an exclusive response to one identity, but not to any of the other identities. Thus, a DOS value of 1 is equal to maximal sparseness of identity coding. The DOS index has been used in many previous studies investigating visual selectivity[38,55,56].

Population decoding of face identities

For population decoding of face identities (Supplementary Fig. 2f,g), we pooled all recorded neurons into a large pseudo-population. Firing rates were *z*-scored individually for each neuron to give an equal weight to each unit regardless of baseline firing rate. We used a maximal correlation coefficient (MCC) classifier as implemented in the MATLAB Neural Decoding Toolbox (NDT)[57]. The MCC estimates a mean template for each class *i* and assigns a class to each test trial. We used 8-fold cross-validation, that is, all trials were randomly partitioned into 8 equal-sized subsamples, of which 7 subsamples were used as the training data and the remaining single subsample was retained as the validation data for assessing the accuracy of the model, and this process was repeated 8 times, with each of the 8 subsamples used exactly once as the validation data. We then repeated the cross-validation procedure 50 times for different random train/test splits. Statistical significance of the decoding performance for each group of neurons against chance was estimated by calculating the percentage of bootstrap runs (50 in total) that had an accuracy below chance (that is, 2% when decoding all identities). Statistical significance for comparing between groups of neurons was estimated by calculating the percentage of bootstrap runs (50 in total) where one group of neurons had a greater accuracy than the other. Spikes were counted in 500 ms bins with a step size of 50 ms. The first bin started at −500 ms relative to trial onset (with its centre at −250 ms), and we tested 31 consecutive bins, ending with the final bin spanning 1,000 to 1,500 ms after trial onset. For each bin, a different classifier was trained and tested. For both tests, we used FDR[54] to correct for multiple comparisons across time points. The same decoding approach was used in our previous studies[58,59] and has been shown to be very effective in studying neural population activity.

Web-association score

We employed a web-based association metric (Fig. 2k) to study the relationship between different identities. To estimate the degree of relationship between the 50 celebrity identities, we used an internet search engine (Google) and compared the number of hits to the joint searches with the number of hits to the individual searches. The rationale is that the name of associated concepts will often appear together in web pages. The web-association score for each identity pair was calculated as $a_{ij} = \log_2\left(\frac{\text{hits(identity}_i\text{ANDidentity}_j\text{)}}{\text{hits(identity}_i\text{)}\times\text{hits(identity}_j\text{)}}\right)$. Because we used celebrity faces, all identities were well searchable from the internet and could thus give a reasonable number of hits to calculate web-association values. We lastly normalized the web-association value using *z*-scoring. Similar results were derived using the search engine Bing. The web-association score has been shown to be effective in revealing 'universal associations' between identities in a previous study[18].

Differential latency

We conducted a spike train analysis to estimate differential response latency for each group of neurons (that is, feature M-ID, M-ID neurons encoding conceptual associations, S-ID)[38]. We binned spike trains into 1-ms bins and computed the cumulative sum. We then averaged the normalized cumulative sums for encoded identities and non-encoded

identities, respectively, and compared, at every point of time, whether the normalized cumulative sums were different between encoded identities and non-encoded identities (two-tailed paired *t*-test; $P < 0.005$; FDR-corrected). The first time point of the significant cluster (cluster size > 20 time points) was used as the estimate of differential latency. Note that this method is not sensitive to differences in baseline firing rates between neurons, as the latency estimate is pairwise for each neuron individually.

To assess statistical significance, we applied a bootstrap procedure with 1,000 runs. In each run, to ensure a fair comparison between groups, we randomly selected 27 neurons from both feature M-ID neurons ($n = 42$) and S-ID neurons ($n = 53$) to match the number of M-ID neurons encoding conceptual associations ($n = 27$). We then compared the average latency of one group with the bootstrap distribution of the other group to obtain *P* values.

Regression analyses

To identify neurons that encoded a linear combination of facial features, we employed both partial least squares (PLS) regression with VGG-Face DNN feature maps and linear regression with the two dimensions of the *t*-SNE feature space (Supplementary Fig. 13). The PLS method has been shown to be effective in studying the neural response to DNN features[9,60]. For PLS, we used 4 components for each layer (explaining at least 80% of the variance; we selected the number of components with a 10-fold cross-validation to minimize the prediction error). For both approaches, we used a permutation test with 1,000 runs to determine whether a neuron encoded a significant face model. In each run, we randomly shuffled the face labels and used 50% of the faces as the training dataset. We used the training dataset to construct a model (that is, deriving regression coefficients), predicted responses using this model for the remaining 50% of faces (that is, test dataset), and computed the Pearson correlation between the predicted and actual response in the test dataset. The distribution of correlation coefficients computed with shuffling (that is, null distribution) was eventually compared to the one without shuffling (that is, observed response). If the correlation coefficient of the observed response was greater than 95% of the correlation coefficients from the null distribution, this face model was considered significant. This procedure has been shown to be very effective in selecting units with significant face models[7]. The correlation coefficient could also indicate the model's predictability and thus be compared between different neurons (Supplementary Fig. 13). Similar results were derived using face models from refs. 7,26.

Pairwise distances in the face space

We employed a pairwise distance metric[43] to compare neural coding of face identities between human/monkey neurons and VGG-Face DNN units (Supplementary Fig. 13n,o). For each pair of identities, we used the dissimilarity value (1 − Pearson's *r*)[61] as a distance metric. The human/monkey neuronal distance metric was calculated between firing rates of all recorded neurons, and the DNN distance metric was calculated between feature weights of all DNN units. For human neurons, we used the mean firing rate in a time window of 250–1,250 ms after stimulus onset as the response to each face[62]. For monkey neurons, we used the mean firing rate in a time window of 70–180 ms after stimulus onset as the response to each face[52,53]. We then correlated the human/monkey neuronal distance metric and the DNN distance metric. To determine statistical significance, we used a non-parametric permutation test with 1,000 runs. In each run, we randomly shuffled the face labels and calculated the correlation between the human/monkey neuronal distance metric and the DNN distance metric. The distribution of correlation coefficients computed 'with' shuffling (that is, null distribution) was eventually compared to the one 'without' shuffling (that is, observed response). If the correlation coefficient of the observed response was greater than 95% of the correlation coefficients from the null distribution, it was considered 'significant'. A significant correlation indicated that the DNN face space had some correspondence to the human/monkey neuronal face space[43]. We computed the correlation for each DNN layer so that we could determine the specific layer to which the neuronal population most closely corresponded.

Norm-based coding and norm polar face space

To test norm-based coding (Supplementary Fig. 14), we first averaged all 500 CelebA faces shown to patients to derive the average face (that is, the origin of the face space). We still used the same VGG-Face DNN to extract features from the average face. We next calculated the Euclidean distance from each face to the average face using full DNN features and correlated the Euclidean distance to the firing rate of a neuron. If a neuron showed a significant correlation, it embodied norm-based coding.

To construct the norm polar face space, we used *t*-SNE to extract 25 dimensions from full DNN features for each face (including the average face). Using *t*-SNE features, we calculated the radial coordinate (that is, radius; Euclidean distance to the average face) and the angular coordinate (that is, azimuth; vector angle with the average face) for each face. Similar results were derived using full DNN features.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data supporting the findings of this study, including the CelebA dataset, FBI dataset, FaceGen dataset and monkey dataset, are publicly available on OSF at https://doi.org/10.17605/OSF.IO/36KZC (ref. 63). We also analysed identity neurons from a publicly available human single-neuron dataset (https://doi.org/10.25392/leicester.data.8796335.v1).

Code availability

The source code for this study is publicly available on OSF at https://doi.org/10.17605/OSF.IO/36KZC (ref. 63).

References

1. Freeman, W. J. *Mass Action in The Nervous System* (Academic, 1975).
2. Hinton, G. E., McClelland, J. L. & Rumelhart, D. E. in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* Vol. 1 (eds Rumelhart, D. E. & McClelland, J. L.) 77–109 (MIT Press, 1986).
3. Rolls, E. T., Treves, A. & Tovee, M. J. The representational capacity of the distributed encoding of information provided by populations of neurons in primate temporal visual cortex. *Exp. Brain Res.* **114**, 149–162 (1997).
4. Churchland, P. S. & Sejnowski, T. J. *The Computational Brain* (MIT Press, 2016).
5. Turk, M. A. & Pentland, A. P. Face recognition using eigenfaces. In *Proc. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 586–591 (IEEE, 1991).
6. Freiwald, W. A., Tsao, D. Y. & Livingstone, M. S. A face feature space in the macaque temporal lobe. *Nat. Neurosci.* **12**, 1187–1196 (2009).
7. Chang, L. & Tsao, D. Y. The code for facial identity in the primate brain. *Cell* **169**, 1013–1028.e14 (2017).
8. Bashivan, P., Kar, K. & DiCarlo, J. J. Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019).
9. Ponce, C. R. et al. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell* **177**, 999–1009.e10 (2019).
10. Bao, P. et al. A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
11. Loffler, G. et al. fMRI evidence for the neural representation of faces. *Nat. Neurosci.* **8**, 1386–1391 (2005).

12. Carlin, J. D. & Kriegeskorte, N. Adjudicating between face-coding models with individual-face fMRI responses. *PLoS Comput. Biol.* **13**, e1005604 (2017).
13. Cao, R. et al. A flexible neural representation of faces in the human brain. *Cereb. Cortex Commun.* **1**, tgaa055 (2020).
14. Barlow, H. B. Single units and sensation: a neuron doctrine for perceptual psychology? *Perception* **1**, 371–394 (1972).
15. Valentine, T. A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Q. J. Exp. Psychol. A* **43**, 161–204 (1991).
16. Quian Quiroga, R. et al. Invariant visual representation by single neurons in the human brain. *Nature* **435**, 1102–1107 (2005).
17. Quian Quiroga, R. Concept cells: the building blocks of declarative memory functions. *Nat. Rev. Neurosci.* **13**, 587–597 (2012).
18. De Falco, E. et al. Long-term coding of personal and universal associations underlying the memory web in the human brain. *Nat. Commun.* **7**, 13408 (2016).
19. Rey, H. G. et al. Encoding of long-term associations through neural unitization in the human medial temporal lobe. *Nat. Commun.* **9**, 4372 (2018).
20. Bausch, M. et al. Concept neurons in the human medial temporal lobe flexibly represent abstract relations between concepts. *Nat. Commun.* **12**, 6164 (2021).
21. Rey, H. G. et al. Single neuron coding of identity in the human hippocampal formation. *Curr. Biol.* **30**, 1152–1159.e3 (2020).
22. Tyree, T. J., Metke, M. & Miller, C. T. Cross-modal representation of identity in the primate hippocampus. *Science* **382**, 417–423 (2023).
23. Reber, T. P. et al. Representation of abstract semantic knowledge in populations of human single neurons in the medial temporal lobe. *PLoS Biol.* **17**, e3000290 (2019).
24. Lin, C., Keles, U. & Adolphs, R. Four dimensions characterize attributions from faces using a representative set of English trait words. *Nat. Commun.* **12**, 5168 (2021).
25. Leopold, D. A., Bondar, I. V. & Giese, M. A. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature* **442**, 572–575 (2006).
26. Oosterhof, N. N. & Todorov, A. The functional basis of face evaluation. *Proc. Natl Acad. Sci. USA* **105**, 11087–11092 (2008).
27. Rutishauser, U. Testing models of human declarative memory at the single-neuron level. *Trends Cogn. Sci.* **23**, 510–524 (2019).
28. Cao, R. et al. A neuronal code for object representation and memory in the human amygdala and hippocampus. *Nat. Commun.* **16**, 1510 (2025).
29. Cao, R. et al. Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus. *Cell Rep.* **43**, 113520 (2024).
30. Kreiman, G., Koch, C. & Fried, I. Category-specific visual responses of single neurons in the human medial temporal lobe. *Nat. Neurosci.* **3**, 946–953 (2000).
31. Fried, I., MacDonald, K. A. & Wilson, C. L. Single neuron activity in human hippocampus and amygdala during recognition of faces and objects. *Neuron* **18**, 753–765 (1997).
32. Wang, S. et al. Neurons in the human amygdala selective for perceived emotion. *Proc. Natl Acad. Sci. USA* **111**, E3110–E3119 (2014).
33. O'keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Oxford Univ. Press, 1978).
34. Behrens, T. E. J. et al. What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* **100**, 490–509 (2018).
35. Roweis, S. T. & Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323–2326 (2000).
36. Tenenbaum, J. B., Silva, V. D. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323 (2000).
37. Wang, S. et al. The human amygdala parametrically encodes the intensity of specific facial emotions and their categorical ambiguity. *Nat. Commun.* **8**, 14821 (2017).
38. Wang, S. et al. Encoding of target detection during visual search by single neurons in the human brain. *Curr. Biol.* **28**, 2058–2069. e4 (2018).
39. Cao, R. et al. Encoding of facial features by single neurons in the human amygdala and hippocampus. *Commun. Biol.* **4**, 1394 (2021).
40. Cao, R. et al. A neuronal social trait space for first impressions in the human amygdala and hippocampus. *Mol. Psychiatry* **27**, 3501–3509 (2022).
41. Liu, Z. et al. Deep learning face attributes in the wild. In *Proc. International Conference on Computer Vision (ICCV)* 3730–3738 (IEEE, 2015).
42. Parsons, S., Kruijt, A.-W. & Fox, E. Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Adv. Methods Pract. Psychol. Sci.* **2**, 378–395 (2019).
43. Grossman, S. et al. Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nat. Commun.* **10**, 4934 (2019).
44. Brainard, D. H. The psychophysics toolbox. *Spat. Vis.* **10**, 433–436 (1997).
45. Parkhi, O. M., Vedaldi, A. & Zisserman, A. *Deep Face Recognition*. In *BMVC 2015—Proceedings of the British Machine Vision Conference 2015* (British Machine Vision Association, 2015).
46. Hinton, G. E. & Roweis, S. T. Stochastic neighbor embedding. *Adv. Neural Inf. Process. Syst.* **15**, 857–864 (2002).
47. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
48. Rutishauser, U., Mamelak, A. N. & Schuman, E. M. Single-trial learning of novel stimuli by individual neurons of the human hippocampus-amygdala complex. *Neuron* **49**, 805–813 (2006).
49. Rutishauser, U. et al. Human memory strength is predicted by theta-frequency phase-locking of single neurons. *Nature* **464**, 903–907 (2010).
50. Rutishauser, U., Schuman, E. M. & Mamelak, A. N. Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. *J. Neurosci. Methods* **154**, 204–224 (2006).
51. Fried, I. et al. *Single Neuron Studies of the Human Brain: Probing Cognition* (MIT Press, 2014).
52. Kar, K. et al. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* **22**, 974–983 (2019).
53. Kar, K. & DiCarlo, J. J. Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron* **109**, 164–176.e5 (2021).
54. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Stat. Methodol.* **57**, 289–300 (1995).
55. Rainer, G., Asaad, W. F. & Miller, E. K. Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature* **393**, 577–579 (1998).
56. Minxha, J. et al. Fixations gate species-specific responses to free viewing of faces in the human and macaque amygdala. *Cell Rep.* **18**, 878–891 (2017).
57. Meyers, E. The neural decoding toolbox. *Front. Neuroinform.* **7**, 8 (2013).
58. Rutishauser, U. et al. Representation of retrieval confidence by single neurons in the human medial temporal lobe. *Nat. Neurosci.* **18**, 1041–1050 (2015).

59. Wang, S. et al. Abstract goal representation in visual search by neurons in the human pre-supplementary motor area. *Brain* **142**, 3530–3549 (2019).
60. Yamins, D. L. K. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
61. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
62. Mormann, F. et al. Latency and selectivity of single neurons indicate hierarchical processing in the human medial temporal lobe. *J. Neurosci.* **28**, 8865–8872 (2008).
63. Cao, R. Data for "Feature-based encoding of face identity by single neurons in the human amygdala and hippocampus". *OSF* https://doi.org/10.17605/OSF.IO/36KZC (2025).

Acknowledgements

We thank all patients for their participation; staff from WVU Ruby Memorial Hospital for support with patient testing; M. Yin and S. Uddenberg for help with analysis; J. Dawson for contributing the FBI Twins dataset; and R. Adolphs, M. Raichle, C. Ponce, L. Chang, D. Tsao, L. She and P. Webster for discussion and valuable comments. This research was supported by the AFOSR (FA9550-21-1-0088, S.W.), NSF (BCS-1945230, S.W. and X.L.; IIS-2114644, X.L. and S.W.), NIH (K99EY036650, R.C.; R01MH129426, S.W. and X.L.; R01MH120194, J.T.W.; R01EB026439, P.B.; U24NS109103, P.B.; U01NS108916, P.B.; U01NS128612, P.B.; R21NS128307, P.B.; P41EB018783, P.B.), the McDonnell Center for Systems Neuroscience (R.C.), Fondazione Neurone (P.B.), and the Dana Foundation (S.W.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

R.C., A.T., X.L. and S.W. designed the research. R.C., A.P., P.B. and S.W. performed experiments. N.J.B. and J.T.W. performed surgery. R.C., J.W., C.L., E.D.F., A.P., H.G.R., X.L. and S.W. analysed data. R.C, J.J.D., A.T., U.R., X.L. and S.W. wrote the paper. All authors discussed the results and contributed to the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-025-02218-1.

Correspondence and requests for materials should be addressed to Runnan Cao or Shuo Wang.

Peer review information *Nature Human Behaviour* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Shuo Wang

Last updated by author(s): Mar 3, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	MATLAB 2018b; Psychtoolbox 3; Neuralynx; Eyelink 1000; Blackrock Microsystems; Intan Recording Controller
Data analysis	Custom code written in MATLAB 2022a; MATLAB neural decoding toolbox (NDT); Deep neural network (DNN) models implemented in MATLAB; psych package in R; FaceGen Modeller program (http://facegen.com; version 3.1); OSort-v4 (https://rutishauserlab.org/osort)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data supporting the findings of this study, including the CelebA dataset, FBI dataset, FaceGen dataset, and monkey dataset, are publicly available on OSF

(https://osf.io/36kzc/). We also analyzed identity neurons from a publicly available human single-neuron dataset (https://doi.org/10.25392/leicester.data.8796335.v1).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex of each participant is reported in Supplementary Table 1.
Reporting on race, ethnicity, or other socially relevant groupings	Race of each participant is reported in Supplementary Table 1.
Population characteristics	Twelve neurosurgical patients with pharmacologically intractable epilepsy participated in the study. Their detailed characteristics have been shown in Supplementary Table 1.
Recruitment	Patients undergoing invasive electrophysiological recordings for clinical purposes were recruited through the neurosurgery clinic. All patients provided informed consent under an approved Institutional Review Board (IRB) protocol. Patients of both sexes and any race were included. Only patients who successfully completed the task were included in the analyses reported in the study. No systematic recruitment bias was expected to alter the results.
Ethics oversight	All participants provided written informed consent using procedures approved by the Institutional Review Boards of West Virginia University (WVU) and Washington University in St. Louis (WUSTL). Patients were not financially compensated for participating in this study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined based on our previous publications (Wang et al. PNAS 2014; Wang et al. Nat Commun 2017).
Data exclusions	No data were excluded.
Replication	The effect reported in the study was consistent and replicated across different subjects. The results obtained using the CelebA stimuli were replicated using the FBI stimuli and FaceGen stimuli in two separate experiments (same patients but different stimuli). Additionally, we replicated the findings using an independent dataset (https://doi.org/10.25392/leicester.data.8796335.v1), which includes different patients and different stimuli from experiments conducted by other researchers.
Randomization	Participants were not grouped and hence no randomization was performed. Trial order was fully randomized for each participant. There was only one monkey subject; therefore, there were no experimental groups or randomization. Trial order was fully randomized for the subject.
Blinding	Data collection was performed automatically and experimenters were blind to the experimental conditions (occurence of catch trials). No blinding was involved during data analysis, as we relied on a fully automatized procedure that did no require data manipulation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	One male rhesus macaque (Macaca mulatta) was used in this study.
Wild animals	The study did not involve wild animals.
Reporting on sex	There was only one monkey subject. Sex was not considered in the study design.
Field-collected samples	The study did not involve field-collected samples.
Ethics oversight	All procedures conformed to local and U.S. National Institutes of Health guidelines, including the U.S. National Institutes of Health Guide for Care and Use of Laboratory Animals. All experiments were performed with the approval of the MIT Institutional Animal Care and Use Committee (IACUC).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*