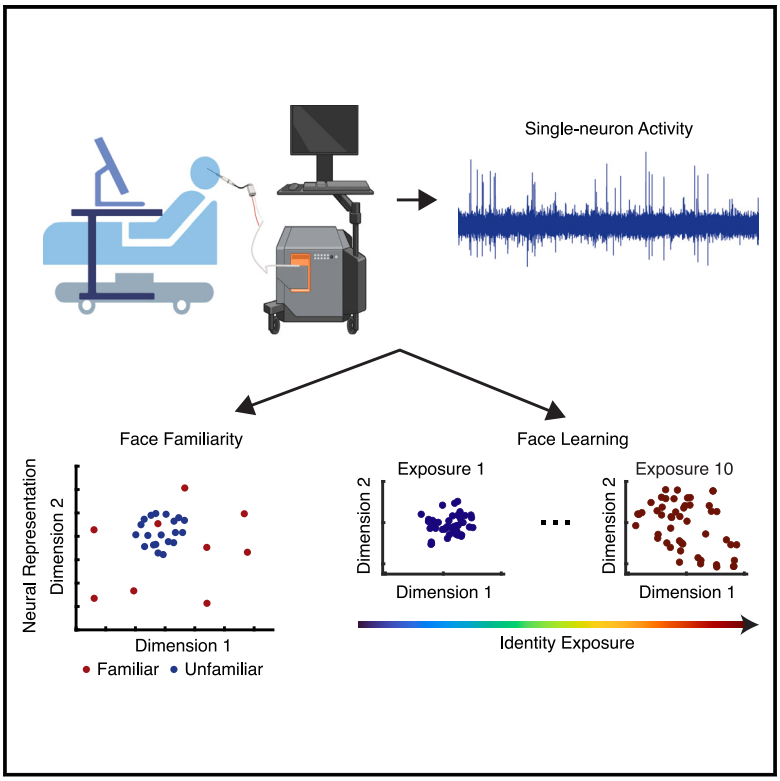
Cell Reports

Article

Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus

Graphical abstract



Authors

Runnan Cao, Jinge Wang, Peter Brunner, ..., Ueli Rutishauser, Nicholas J. Brandmeir, Shuo Wang

Correspondence

r.cao@wustl.edu (R.C.), shuowang@wustl.edu (S.W.)

In brief

Cao et al. show that the neuronal population geometry in the human amygdala and hippocampus, quantified by the representational distance, encodes face familiarity, similarity, and learning. The neuronal representational distance can be a generic code to explain neural face representations.

Highlights

- Greater neuronal representational distance between familiar than unfamiliar faces
- Neuronal representational distance increases with face learning and familiarization
- Representational distance correlates with visual dissimilarity between faces
- Exposure to visually similar faces increases neuronal representational distance

Cao et al., 2023, Cell Reports ■■, 113520
■■, 2023
https://doi.org/10.1016/j.celrep.2023.113520

CellPress

Please cite this article in press as: Cao et al., Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus, Cell Reports (2023), https://doi.org/10.1016/j.celrep.2023.113520

Cell Reports

CellPress
OPEN ACCESS

Article

Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus

Runnan Cao,[1,2,*] Jinge Wang,[2] Peter Brunner,[3] Jon T. Willie,[3] Xin Li,[2] Ueli Rutishauser,[4] Nicholas J. Brandmeir,[5] and Shuo Wang[1,2,3,6,*]

[1]Department of Radiology, Washington University in St. Louis, St. Louis, MO 63110, USA
[2]Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV 26506, USA
[3]Department of Neurosurgery, Washington University in St. Louis, St. Louis, MO 63110, USA
[4]Departments of Neurosurgery and Neurology, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA
[5]Department of Neurosurgery, West Virginia University, Morgantown, WV 26506, USA
[6]Lead contact
*Correspondence: r.cao@wustl.edu (R.C.), shuowang@wustl.edu (S.W.)
https://doi.org/10.1016/j.celrep.2023.113520

SUMMARY

Recognizing familiar faces and learning new faces play an important role in social cognition. However, the underlying neural computational mechanisms remain unclear. Here, we record from single neurons in the human amygdala and hippocampus and find a greater neuronal representational distance between pairs of familiar faces than unfamiliar faces, suggesting that neural representations for familiar faces are more distinct. Representational distance increases with exposures to the same identity, suggesting that neural face representations are sharpened with learning and familiarization. Furthermore, representational distance is positively correlated with visual dissimilarity between faces, and exposure to visually similar faces increases representational distance, thus sharpening neural representations. Finally, we construct a computational model that demonstrates an increase in the representational distance of artificial units with training. Together, our results suggest that the neuronal population geometry, quantified by the representational distance, encodes face familiarity, similarity, and learning, forming the basis of face recognition and memory.

INTRODUCTION

Understanding face familiarity is essential to our understanding of face recognition. For many years, it has been recognized that there are notable distinctions in the perception of familiar and unfamiliar faces (for reviews, see Hancock et al.,[1] Johnston and Edmonds,[2] and Young and Burton[3]), and this point has been emphasized in theoretical models of face recognition.[4,5] Familiar faces, of which participants can recognize the identity, demonstrate processing advantages. For example, across experiments requiring participants to match unfamiliar faces, performance is highly error prone, especially when matches vary in viewpoint and expression.[6] It has been shown that a brief period of familiarization with faces can improve internal feature matching performance beyond that observed with completely novel faces.[7] In studies of recognition memory, familiar faces are recognized faster and more accurately than unfamiliar faces.[8,9] A recent view even argues that expertise in face recognition is limited to familiar faces, while perceptual performance with unfamiliar faces does not meet the criteria for expertise.[10] However, the neural basis of the transition from error-prone and inflexible recognition of unfamiliar faces[1] to highly accurate and robust recognition of familiar faces[2] remains unclear.

Faces that vary in their degree of visual familiarity elicit neural responses with different spatial and temporal characteristics in multiple regions of the ventral-temporal and parietal cortex.[11] Electroencephalogram (EEG) experiments using flicker steady state visually evoked potential (SSVEP) argue for distinct neural processes for the perception of familiar vs. unfamiliar faces along the visual hierarchy.[12] In analysis of the time course of neural responses to faces using magnetoencephalography (MEG), it has been shown that the representations of identity and gender in familiar faces undergo early enhancement, indicating that the behavioral advantage associated with familiar faces arises from the tuning of early feedforward processing mechanisms.[13] Different patterns of neural activity are also elicited in response to seeing visually familiar vs. unfamiliar people in motion.[14] Furthermore, EEG experiments suggest how face familiarity coding and identity coding vary as a function of levels of familiarization (brief perceptual exposure vs. extensive media familiarization vs. real-life personal familiarization).[15] In addition, primate studies have suggested a transition that turns face perception into face memory in the temporal pole.[16]

The human amygdala and hippocampus play a key role in both the processing of faces and memory for faces.[17] Using faces with experimentally induced visual familiarity that carries no biographical information or emotional content, it has been shown

Please cite this article in press as: Cao et al., Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus, Cell Reports (2023), https://doi.org/10.1016/j.celrep.2023.113520

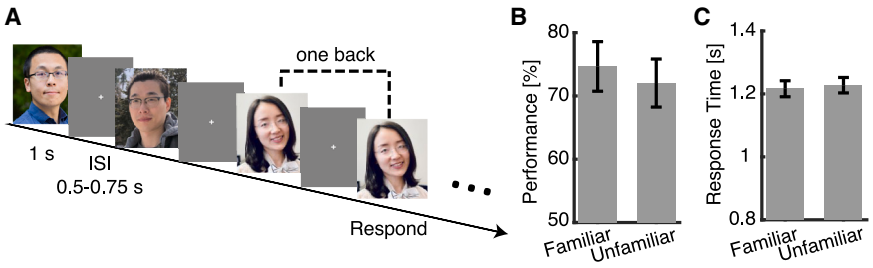


Figure 1. Behavior

(A) Task. We employed a one-back task in which patients responded whenever an identical face was repeated. Each face was presented for 1 s, followed by a jittered inter-stimulus interval (ISI) of 0.5–0.75 s. For copyright reasons, we replaced the original stimulus images with similar pictures.

(B) One-back detection accuracy for familiar vs. unfamiliar faces.

(C) One-back detection reaction time for familiar vs. unfamiliar faces.

Error bars denote ±SEM across sessions. See also Figure S1.

that familiar faces evoke a differential amygdala blood-oxygen-level-dependent (BOLD) response relative to novel faces.[18] In contrast to the ventral occipitotemporal face-preferential regions, where BOLD activity reflects visual information irrespective of face familiarity, the amygdala and hippocampus exhibit an abrupt increase in the BOLD signal when sufficient information is provided to identify a face as familiar.[19] At the single-neuron level, "concept" neurons in the human amygdala and hippocampus that demonstrate identity-specific coding are primarily probed using familiar faces,[20–22] and selective cells are more likely for familiar faces than unfamiliar faces.[23] Together, these findings suggest that a potential mechanism for the emergence of highly selective cells in the amygdala and hippocampus is that familiarization leads to the emergence and/or sharpening of tuning of cells responsive to small subsets of faces of specific individuals (i.e., concept cells).

In this study, we focused on the neural mechanisms underlying face familiarization and learning. In particular, motivated by theories of pattern separation,[24,25] we investigated whether neuronal representational distance changed as a function of face familiarity, similarity, and learning, and we further investigated whether the neuronal population geometry of face representations changed during face learning. We finally employed a high-performing deep neural network (DNN)-based computational face model (e.g., Blauch et al.[26]) to examine the computational principles underlying the neural face learning process and face representation.

RESULTS

Behavior

Nine neurosurgical patients undergoing single-neuron recordings (Table S1) viewed 500 unique natural face images of 50 celebrities from the CelebA dataset (10 images per celebrity) while performing a one-back task (Figure 1A; accuracy = 72.45% ± 20.54% [mean ± SD across sessions]). Patients completed a questionnaire after they completed their recordings of whether they recognized (i.e., could tell the name from a picture) each face identity (one of the 10 faces was randomly selected from each identity for this questionnaire).[27,28] We refer to the face identities that a patient recognized as "familiar" and the face identities that a patient did not recognize as "unfamiliar." On average, patients rated 40.67% ± 21.52% (mean ± SD across patients) of the identities shown as familiar. One-back detection performance (Figure 1B; hit rate: 74.66% ± 21.47% for familiar faces and 72.04% ± 20.79% for unfamiliar faces; two-tailed paired t test: $t(30) = 1.02$, $p = 0.32$) and reaction time (Figure 1C; relative to image onset; familiar: 1.22 ± 0.14 s; unfamiliar: 1.23 ± 0.13 s; $t(30) = 1.17$, $p = 0.25$) were similar for familiar faces vs. unfamiliar faces (see Figure S1 for additional behavioral results).

Neurons encoding face familiarity

In the CelebA dataset, we recorded from 1,402 neurons in the amygdala and hippocampus (31 sessions in total; overall firing rate greater than 0.15 Hz; all sessions were recorded on different days, and neurons from each individual recording session were considered independent even when they were from the same patient), which included 623 neurons from the amygdala, 478 neurons form the anterior hippocampus, and 401 neurons from the posterior hippocampus. In this section, we identified and analyzed the neurons that encoded face familiarity by contrasting their response between familiar vs. unfamiliar faces. Using the ratings provided by each patient, we found that the response of 97 neurons (6.92%, binomial $p = 0.0007$) differed significantly between all familiar vs. unfamiliar faces (see Figures 2A and 2B for examples and Figures 2C–2F for group results). This suggests that a subset of amygdala and hippocampal neurons encodes face familiarity, and we focus on this subset of neurons for further analyses. Among these familiarity-selective neurons, 79 neurons had a greater response to familiar faces (i.e., increasing activity for familiar faces; Figures 2A–2C), and 18 neurons had a greater response to unfamiliar faces (i.e., decreasing activity for familiar faces; Figure 2D; see also Figure S2A for group analysis using single-trial response index and Figure S2B for receiver operating characteristic [ROC] analysis). The proportion of familiarity-selective neurons that increased activity for familiar faces was higher (χ^2 test: $p < 10^{-20}$).

We next considered the response of all familiarity-selective neurons ($n = 97$, including neurons increasing and decreasing activity for familiar faces) as a population using a pairwise distance metric (i.e., representational distance). We found that the neuronal representational distance (STAR Methods) between pairs of familiar faces was greater than that of pairs of unfamiliar faces (Figure 2E; two-tailed paired t test: $t(96) = 4.30$, $p = 4.13 \times 10^{-5}$; linear mixed effect model [representational distance ~ familiarity + (1|subject:session)]: $\beta = 0.38 \pm 0.12$, $t(168.71) = 3.21$, $p = 0.002$; see Figure 2F for temporal dynamics of individual neurons), suggesting that pairs of familiar faces were more neurally distinct compared with pairs of unfamiliar faces.

Please cite this article in press as: Cao et al., Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus, Cell Reports (2023), https://doi.org/10.1016/j.celrep.2023.113520

CellPress
OPEN ACCESS

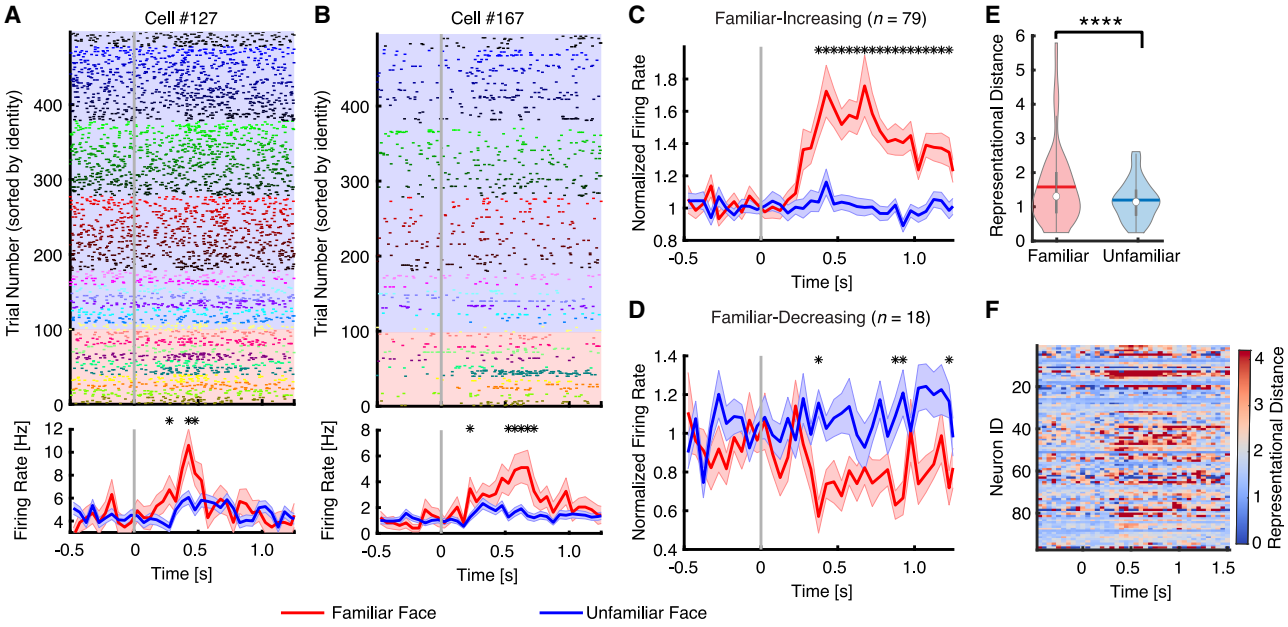


Figure 2. Neurons that differentiate familiar vs. unfamiliar faces

(A and B) Example neurons that had a greater firing rate for familiar faces (selection by two-tailed t test in a time window of 250–1,250 ms after trial offset; both $p < 0.01$). Trials are aligned to face stimulus onset (gray line) and grouped by individual identities. Shaded areas in the raster plot indicate familiar (red) and unfamiliar (blue) identities. Asterisks indicate a significant difference between familiar vs. unfamiliar faces in that bin ($p < 0.01$, two-tailed t test, uncorrected; bin size = 50 ms). Note that the selection of neurons was based on the entire time window.

(C) Average normalized firing rate of neurons with a greater response to familiar faces (n = 79; i.e., increasing activity for familiar faces).

(D) Average normalized firing rate of neurons with a greater response to unfamiliar faces (n = 18; i.e., decreasing activity for familiar faces). Shaded areas denote ±SEM across neurons. Asterisks indicate a significant difference between the conditions in that bin ($p < 0.05$, two-tailed t test, corrected by false discovery rate [FDR][29] for $q < 0.05$).

(E) Representational distance between familiar vs. unfamiliar faces. Representational distance was calculated between faces using absolute difference in firing rate. In the violin plots, the white dot represents the median, the thick gray bar in the center represents the interquartile range, and the thin gray line represents the rest of the distribution, except for points that are determined to be outliers using a method that is a function of the interquartile range. On each side of the gray line is a kernel density estimation to show the distribution shape of the data. Asterisks indicate a significant difference between familiar vs. unfamiliar faces using two-tailed paired t test. ****$p < 0.0001$.

(F) Representational distance for each familiarity-selective neuron with a greater response to familiar faces.

See also Figures S2 and S3.

Notably, our results remained robust even after accounting for differences in signal-to-noise ratio (SNR) related to changes in firing rate (SNR = [(across trial variance)/(baseline variance)]−1; analysis of covariance [ANCOVA] of familiarity effect: $t = 2.75$, $p = 0.007$). The results were similar when we computed the representational distance between pairs of identities (Figure S2F; $t(96) = 11.21$, $p = 3.72 \times 10^{-19}$; linear mixed effect model: $\beta = 6.27 \pm 0.59$, $t(172) = 10.72$, $p = 7.68 \times 10^{-21}$; ANCOVA controlling for SNR: $t = 10.93$, $p = 6.73 \times 10^{-22}$), showing that the increase in representational distance is not restricted to specific visual features but extends to the more abstract concept of identity of a face. Such a difference in neural representation is in line with the widely accepted idea that distinctive faces undergo deeper or more elaborate processing than typical faces and are thus better recognized.[2] Furthermore, we observed a comparable encoding of familiarity in both the amygdala and hippocampus (Table S1; Figures S3A, S3B, S3G, and S3H). We show response latency (Figures S2D and S2E), representational similarity with DNN units (Figure S2H), and relationship with identity coding (Figures S2G and S2I).

Face learning and familiarization through identity exposures within a session

Why does the representational distance between pairs of familiar faces increase? Familiarity with a human face typically develops via multiple visual exposures to a person during social interactions. We next investigated the neural basis of how faces become familiar (i.e., how the neuronal familiarity effect arises) and how people learn to recognize a face identity through visual exposures. In each session, we presented 10 different (unique, each only seen once) pictures of the same person (identity) to the patients, so we were able to examine the change in neural response as a function of the number of exposures to a given identity (see Figure S4 for behavioral results). We then tested, using a linear regression, whether neurons changed their firing rate as a function of the number of times an image of the same identity was seen by the participant. Due to both physiological and non-physiological factors, such as changes in attention or electrode drift, there may be changes in firing rate as a function of time, which could confound the analysis of face learning over identity exposure within a session. To isolate the response

Please cite this article in press as: Cao et al., Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus, Cell Reports (2023), https://doi.org/10.1016/j.celrep.2023.113520

specific to the stimulus from potential slow drifts in neurons' firing over time, we calculated a mean baseline firing rate every 30 trials over a time window of 500 ms (−500 to 0 ms relative to stimulus onset) and divided the response to the stimulus by the corresponding baseline. Furthermore, we combined familiar and unfamiliar identities for analysis, but we show below that face learning and familiarization occurred similarly for both familiar faces and unfamiliar faces (i.e., face learning still happened for familiar faces; Figure S5).

We identified 225 neurons (16.05%, binomial $p < 10^{-20}$) that exhibited a significant change in response to identity exposure within a session (see Figures 3A–3C and 3E for examples and Figures 3G, 3H, 3J, and 3K for group summary). Among these neurons, 165 (73.3%) linearly increased firing rate with identity exposure (Figures 3A, 3C, 3E, 3G, and 3H), while only 60 (26.7%) linearly decreased firing rate with identity exposure (Figures 3B, 3J, and 3K; χ^2 test: $p < 10^{-20}$). As a consequence, the representational distance between faces increased over identity exposure (see Figures 3D and 3F for examples; Figures 3I and 3L for group results; and Figures S3C, S3D, S3I, and S3J for a breakdown of amygdala and hippocampal neurons). Furthermore, the mean linear regression coefficient of the entire neuronal population was significantly greater than 0 (linear mixed effect model [regression coefficient ~ 1 + (1|subject:session)]: $\beta = 0.074 \pm 0.029$, $t(7.13) = 2.58$; $p = 0.036$). Because the majority of the neurons increased their response, neural face representations tended to be more distinct after exposure to more faces of an identity. Notably, we found a similar pattern of results for both familiar faces and unfamiliar faces; i.e., face learning still happened for familiar faces with a similar strength as for unfamiliar faces (Figure S5; regression coefficient: familiar: 0.82 ± 0.13, unfamiliar: 0.81 ± 0.14; two-tailed two-sample t test: $t(265) = 0.56$, $p = 0.58$), suggesting that amygdala and hippocampal neurons signal face learning and familiarization regardless of whether an identity is already familiar to the patient. Face familiarization is thus a continuous process that also applies to known faces, and familiar faces can get further familiarized.

Were our results specific to face identities? Although time cells have been described in the hippocampus,[30] and we controlled for temporal change of overall neural response, it is still possible that the response change over exposures was due to adaptation or sensitization of faces rather than learning about face identities per se. To address this potential confounding factor, we conducted two control analyses. First, when we grouped the trials by time bins (10 bins, 50 trials per bin) rather than identity exposure, we did not find an above-chance population of neurons showing a significant linear correlation with time bins (75 neurons, 5.35%, binomial $p = 0.25$). Second, we conducted a control experiment using FaceGen model faces,[31] which contained only facial feature information but no face identity information. We recorded from 938 neurons (overall firing rate greater than 0.15 Hz) in 28 sessions (10 patients; see Cao et al.[32] for a detailed analysis of behavior). Again, we grouped the trials into 10 consecutive time bins (30 trials per bin), and we did not find an above-chance population of neurons showing a significant linear correlation with time bins (37 neurons, 3.94%, binomial $p = 0.92$). Therefore, our results were specific to face identities; i.e., neurons linearly changed firing rate as a function of identity exposure.

Neuronal population geometry for identity exposures

Above, we analyzed face learning for individual neurons. Would the population geometry of the neural face space (including both representational distance and angle) change as a function of identity exposure? Specifically, if all neurons change their response proportionally, then the angle between the neuronal vectors will not change; otherwise, a change in the angle will suggest a change in the population geometry. To answer this question, we calculated the representational distance and angle for the population of neurons (STAR Methods) and used a linear regression to test whether the population geometry changed linearly as a function of identity exposure. We combined all neurons showing a significant linear change in response to identity exposure (see Figures S6A–S6F for separate analyses of neurons with increased or decreased firing rate as well as temporal dynamics; we also derived similar results with all recorded neurons). As expected, changes in firing rate were translated into changes in representational distance, and we found that the population representational distance increased as a function of identity exposure (Figure 4A; Pearson correlation: $r(10) = 0.88$, $p = 0.00083$; see Figures 4E and 4F for illustration). We also obtained similar results when accounting for the SNR (partial linear correlation: $r(10) = 0.88$, $p = 0.0017$). However, we found that the angle between the neuronal vectors did not change (Figure 4B; $r(10) = 0.21$, $p = 0.56$), suggesting that individual neurons changed their response proportionally so that the population geometry remained constant (unchanged) across identity exposures. This indicates that the primary effect of familiarization was a scaling of the response rather than a change in tuning (which would result in angle changes). Separate analyses within the amygdala and hippocampus derived similar results (Figures S3E, S3F, S3K, and S3L).

Can face learning happen across sessions? Specifically, can representational distance further increase across sessions? Indeed, we found that the representational distance increased from the first session to the second session for all neurons showing a significant linear change in response to identity exposure (Figure 4C; two-tailed paired t test: $t(1,224) = 40.50$, $p = 3.30 \times 10^{-228}$), suggesting that face learning and familiarization continued across sessions. Interestingly, the neuronal population geometry also changed as the angle between neuronal vectors changed (Figure 4D; $t(1224) = 25.43$, $p = 6.76 \times 10^{-115}$; for both representational distance and angle, we derived the same results when we used the same number of neurons between sessions). Here, we compared representational distance and angle by identity pairs, but we derived similar results when we compared face pairs (Figures S6G and S6H; both $p < 10^{-50}$). We also derived similar results with all recorded neurons (Figures S6I and S6J; both $p < 10^{-231}$).

Together, our results show that representational distance increases over identity exposures, an effect that can be carried over to the following session. Therefore, neural face representations are sharpened with identity exposures. Our results also show that the neuronal population geometry is constant within a session but changes over sessions.

Please cite this article in press as: Cao et al., Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus, Cell Reports (2023), https://doi.org/10.1016/j.celrep.2023.113520

CellPress
OPEN ACCESS

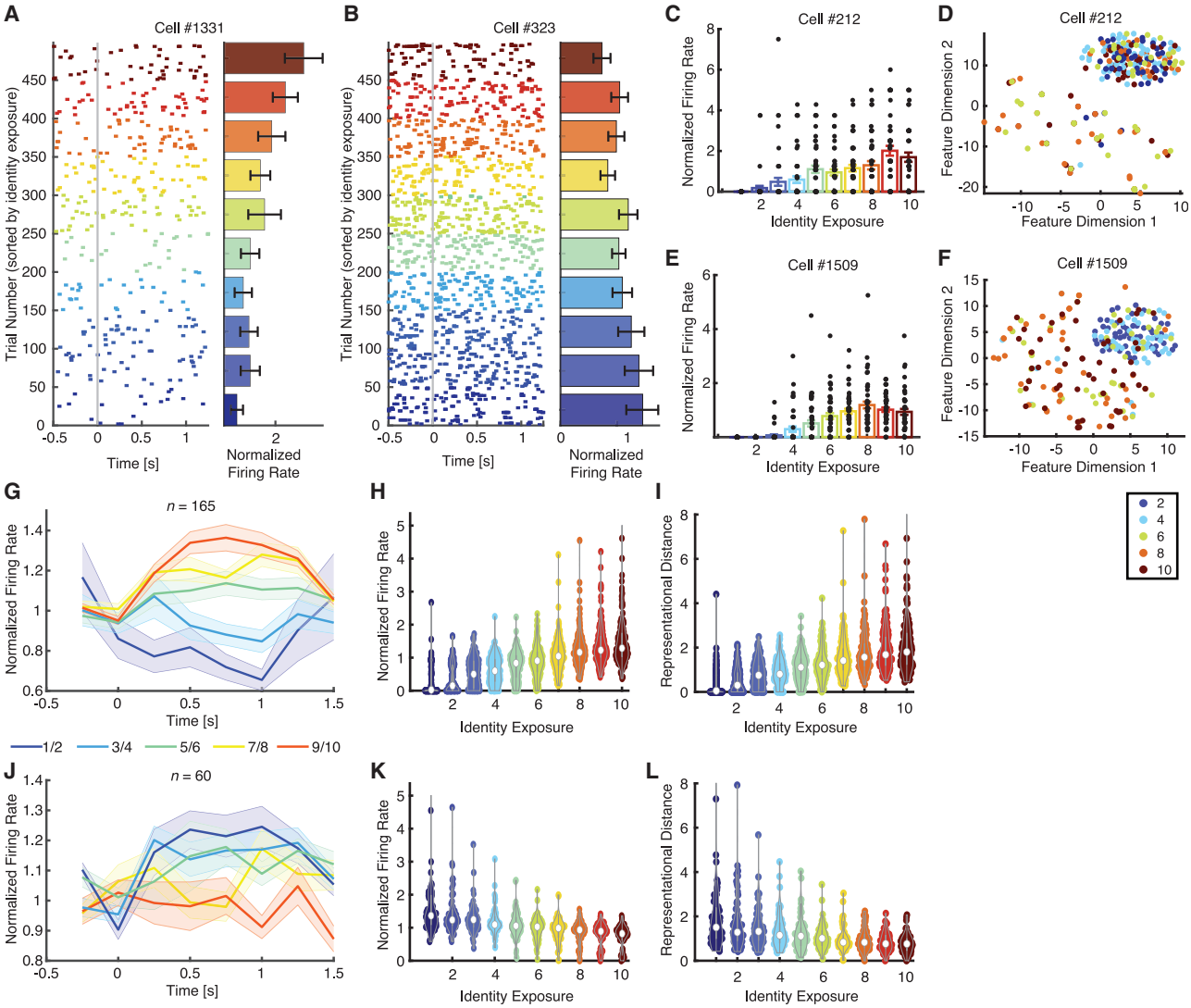


Figure 3. Face learning within a session

(A and B) An example neuron that linearly increased firing rate with identity exposure (A) and an example neuron that linearly decreased firing rate with identity exposure (B). Trials are aligned to face stimulus onset (gray line) and grouped by identity exposures. Error bars denote ±SEM across trials. We calculated a mean baseline firing rate every 30 trials in a time window of −500 to 0 ms relative to stimulus onset, and firing rate was normalized to the baseline.

(C–F) Example neurons whose representational distance of faces increased as a function of identity exposure.

(C and E) Normalized firing rate for each identity exposure. Each dot represents a face/trial, and error bars denote ±SEM across faces/trials. These neurons changed firing rate as a linear function of identity exposure.

(D, F) Distribution of faces in the neuronal feature space. Each dot represents a face. Color coding shows the number of identity exposures.

(G and H) Mean normalized firing rate for neurons that linearly increased firing rate with identity exposure (n = 165).

(J and K) Mean normalized firing rate for neurons that linearly decreased firing rate with identity exposure (n = 60).

(G and J) Group peristimulus time histogram (PSTH). Shaded areas denote ±SEM across neurons. Here, we averaged adjacent identity exposures for illustration purposes.

(I) Mean representational distance for neurons that linearly increased firing rate with identity exposure.

(L) Mean representational distance for neurons that linearly decreased firing rate with identity exposure. Representational distance was calculated by faces (STAR Methods).

See also Figure S3, S4, and S5.

Representational distance encodes face similarity/distinctiveness

We have shown above that population representational distance is increased by face familiarity and learning and that, thus, neural face representations become more distinct as faces become familiar. Along this line of reasoning, representational distance may, in addition, also reflect levels of face similarity; pairs of visually similar faces will have a smaller representational distance. We

Please cite this article in press as: Cao et al., Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus, Cell Reports (2023), https://doi.org/10.1016/j.celrep.2023.113520

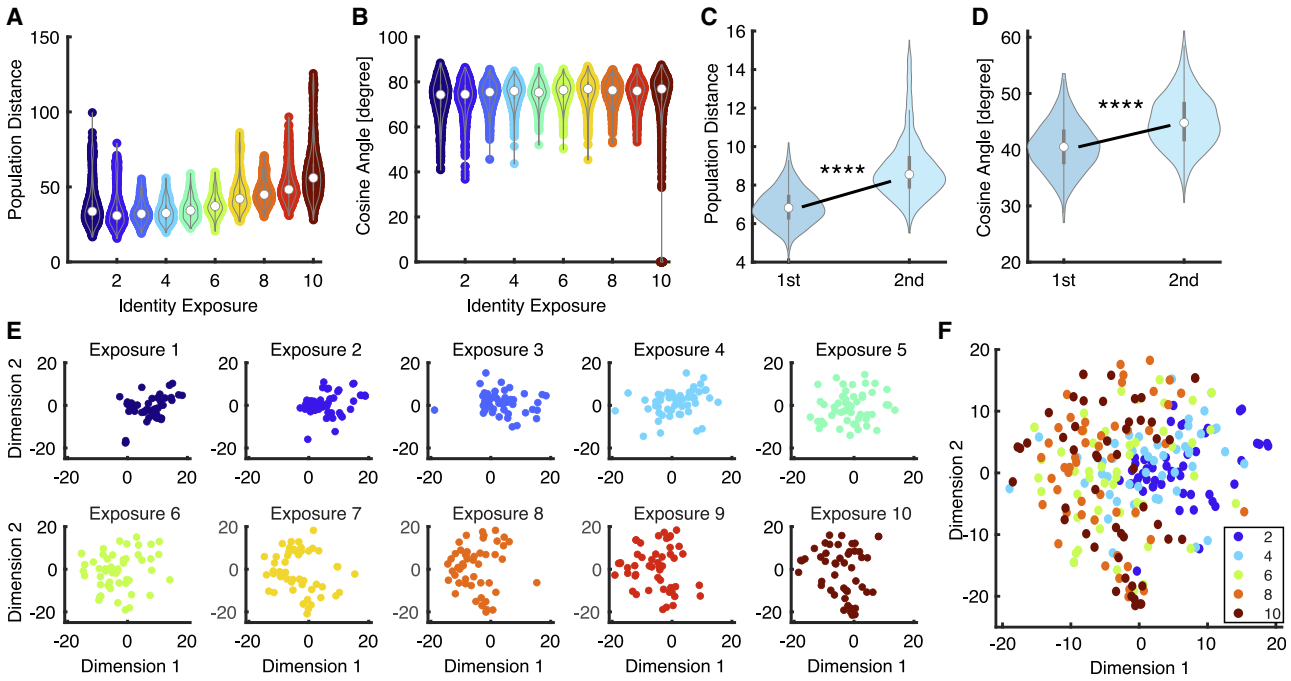


Figure 4. Neuronal population geometry

(A and C) Representational distance for the population of neurons.

(B and D) Angle between the neuronal vectors.

(A and B) Comparison across identity exposures.

(C and D) Comparison between the first vs. second session. Asterisks indicate a significant difference between sessions using two-tailed paired t test. ****$p < 0.0001$. Legend conventions are as in Figure 2 for violin plots.

(E and F) Distribution of faces in the neuronal space (constructed by t-distributed stochastic neighbor embedding [t-SNE] dimension reduction of the neuronal vector for each face). Each dot represents a face. Color coding shows the number of identity exposures.

See also Figures S3, S4, S5, and S6.

have shown that neurons in the human amygdala and hippocampus encode visually similar faces (i.e., faces sharing similar visual features are neurally more similar as well).[27] Here, we investigated whether representational distance was related to visual face similarity/distinctiveness by comparing pairwise distances between face identities assessed for similarity by human raters, neuronal representational distance, and DNN representational distance.

With the CelebA stimuli (Figure 5A), 5 patients from whom we recorded provided ratings for how visually similar each pair of face identities looked to them (Figure 5B). We then correlated this pairwise similarity with the population representational distance for each pair of face identities. We found a significant correlation for all recorded neurons (Figures 5D and 5E; permutation test: $p < 0.001$), face-responsive neurons (Figures S7A and S7C; permutation test: $p = 0.001$), and identity-selective neurons (Figures S7B and S7D; permutation test: $p < 0.001$) with visual dissimilarity, suggesting that population representational distance was influenced by visual similarity/distinctiveness between face identities; the more visually similar two faces were rated, the smaller the representational distance. Similarly, we replicated this finding using DNN feature distance instead of human ratings (Figures 5C, 5F, and S7).

We next used the FBI Twins stimuli (Figure 5G) to validate our findings and further demonstrate that representational distance is influenced by face similarity in addition to face familiarity. The FBI stimuli contained faces that were all unfamiliar to the patients with various levels of visual similarity, including identical twins (ITs), mirror twins (MTs), fraternal twins (FT), mother-child (MC), father-child (FC), and spouses (SPs). We first employed a Siamese neural network (i.e., a convolutional neural network that assigns a similarity score between input images)[33] to estimate the visual similarity between faces and confirmed the levels of visual similarity in the stimuli (Figure 5H). We recorded from 837 neurons in the amygdala and hippocampus (overall firing rate greater than 0.15 Hz) in 27 sessions (10 patients; accuracy = 75.7% ± 23.0% [mean ± SD across sessions]), and we identified 84 face-responsive neurons (10.0%, binomial $p = 9.74 \times 10^{-10}$). We found that the representational distance of face-responsive neurons (Figure 5I; one-way ANOVA: $p = 0.048$) and all neurons (Figure 5J; $p = 0.039$) was related to visual similarity in a graded manner; twin faces (including all IT, MT, and FT pairs) were most visually similar and had the smallest representational distance, parent-child faces (including MC and FC pairs) were visually similar and had an intermediate representational distance, and SP faces (including SP pairs) were not visually similar and had the largest representational distance. Therefore, our results again suggest that representational distance is related to face similarity (here for faces that are all unfamiliar).

Together, our results suggest that representational distance is correlated significantly with face similarity/distinctiveness,

Please cite this article in press as: Cao et al., Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus, Cell Reports (2023), https://doi.org/10.1016/j.celrep.2023.113520

CellPress
OPEN ACCESS

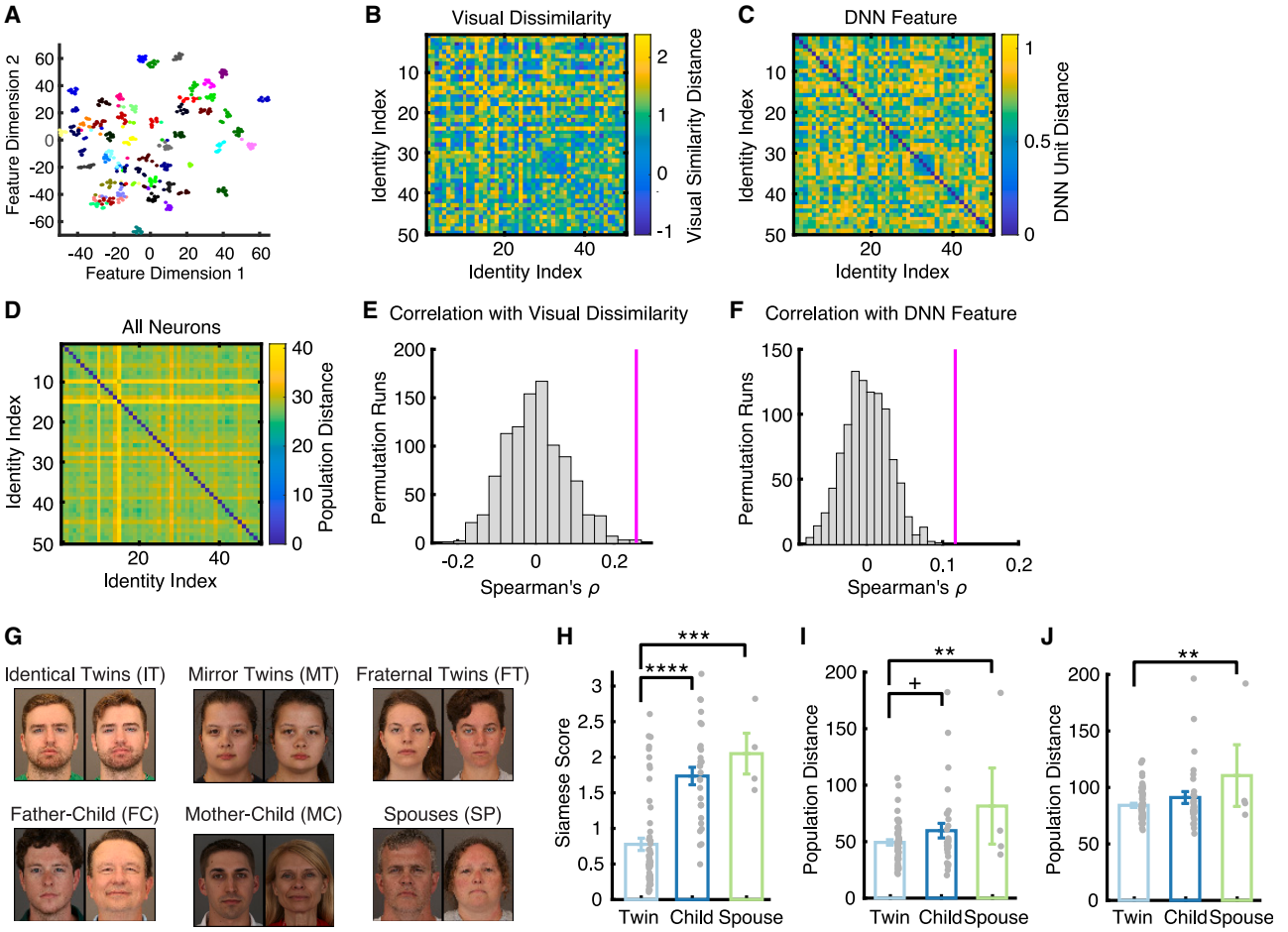


Figure 5. Face similarity

(A) The CelebA face feature space constructed by t-SNE for the deep neural network (DNN) layer Pool5. We applied t-SNE to convert high-dimensional DNN features into a two-dimensional feature space. Each dot represents a face image, and each color represents an identity.

(B) Distance matrix for visual similarity ratings. Color coding shows the negative of *Z*-scored similarity ratings for each identity pair.

(C) Distance matrix for DNN features (i.e., DNN units). Color coding shows dissimilarity values (1 – Pearson's *r*) between each identity pair.

(D) Population representational distance matrix. Color coding shows the Euclidean distance of neurons between each identity pair.

(E and F) Observed vs. permuted correlation coefficient between distance matrices. The correspondence between distance matrices was assessed using permutation tests with 1,000 runs. The magenta line indicates the observed correlation coefficient between distance matrices. The null distribution of correlation coefficients (shown in the gray histogram) was calculated by permutation tests of shuffling the face identities.

(G) Example stimuli from the FBI Twins dataset.

(H) Siamese score for each category of faces.

(I and J) Population representational distance for each category of faces.

(I) Face-responsive neurons (n = 84).

(J) All neurons (n = 837).

Each dot represents a face pair, and error bars denote ±SEM across face pairs. Asterisks indicate a significant difference using two-tailed two-sample t test. +$p < 0.1$, **$p < 0.01$, ***$p < 0.001$, ****$p < 0.0001$. See also Figure S7.

another important aspect of neural face representation (see Discussion for a comprehensive summary).

Face learning through visually similar faces

An important aspect of face learning is to generalize recognition to similar faces (e.g., recognize a person's sibling who looks alike). Above, we have shown that neural face representations become more distinct after exposure to faces of the same identity. Does population representational distance also increase after exposure to visually similar faces? In other words, can visually similar faces play the same role in sharpening the neural face representations during face learning as faces of the same identity? To address this question, we calculated the population representational distance and angle between one of the twin faces (including ITs, MTs, and FTs; STAR Methods) or one of the non-twin faces (including MC, FC, and SPs), and contrasted the population representational distance and angle between the first vs. second exposures to these faces (see Figures 6A and 6D

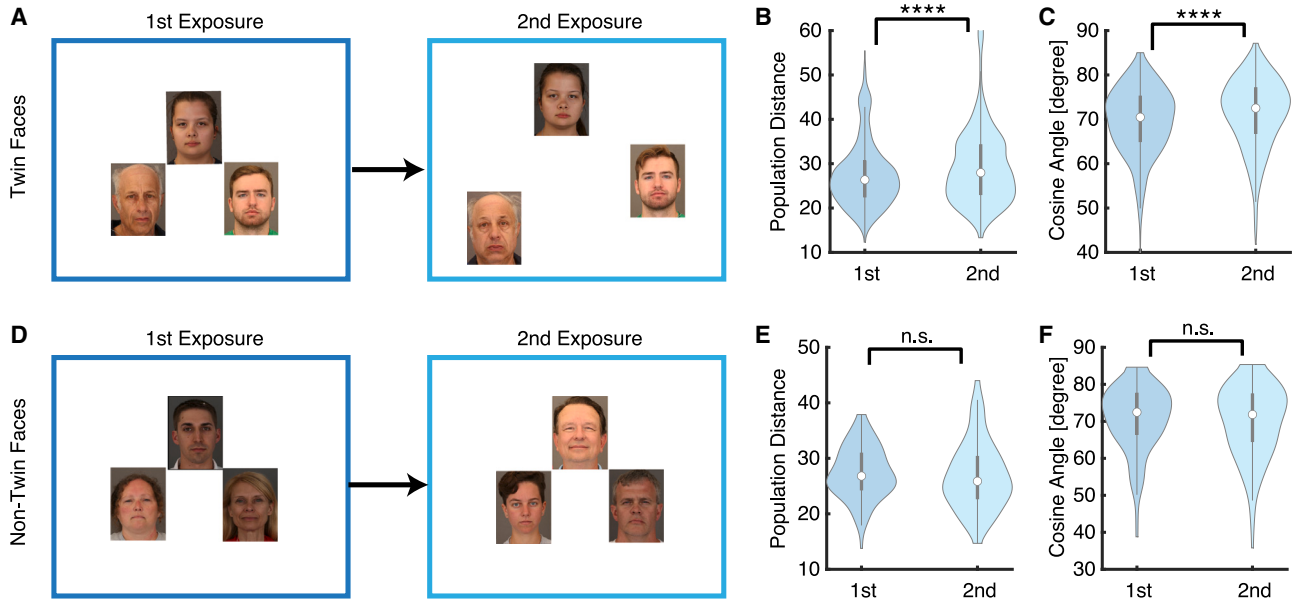


Figure 6. Face learning through visually similar faces

(A–C) Exposure to twin faces.

(D–F) Exposure to non-twin faces.

(A and D) Schematics of changes in representational distance following exposure to (A) twin faces and (D) non-twin faces.

(B and E) Representational distance for the population of neurons.

(C and F) Angle between the neuronal vectors. Here, we focused on the face-responsive neurons (n = 84).

Asterisks indicate a significant difference using two-tailed paired t test. ****p < 0.0001. n.s., not significant. Legend conventions are as in Figure 2 for violin plots.

for schematics; see also Figures 5H–5J for quantification of visual similarity).

Indeed, when we compared population geometry between the first vs. second exposures to a twin face (Figure 6A), we found that population representational distance of face-responsive neurons increased between face pairs (Figure 6B; two-tailed paired t test: $t(1,539) = 6.01$, $p = 2.26 \times 10^{-9}$), consistent with the increased representational distance after exposure to faces of the same identity (Figures 4A and 4C). Interestingly, we found that the neuronal population geometry also changed as the angle between the neuronal vectors increased (Figure 6C; n = 84 neurons for each group; $t(1,539) = 6.14$, $p = 1.06 \times 10^{-9}$), a result similar to face learning across sessions (Figure 4D; Discussion). This result included all three categories of twins (ITs, MTs, and FTs), but it could be replicated without ITs (population representational distance: $t(629) = 3.91$, $p = 1.03 \times 10^{-4}$; angle: $t(629) = 3.75$, $p = 1.96 \times 10^{-4}$). Furthermore, as a control, we compared population geometry between the first vs. second exposures to a non-twin face (Figure 6D). We found that face learning was abolished with non-twin faces (Figures 6E and 6F); neither the representational distance (Figure 6E; $t(119) = 0.89$, $p = 0.37$) nor the angle between the neuronal vectors (Figure 6F; n = 84 neurons for each group; $t(119) = 0.61$, $p = 0.54$) changed between exposures, suggesting that the change in neural face representations through face exposure was specific to visually similar faces.

Together, our results show that neural face representations can be sharpened by exposure to visually similar faces, which, in turn, suggest that face learning can be generalized to visually similar faces. The learning through visually similar faces supports the notion that individuals possess distinct areas of specialized knowledge associated with each familiar face.[34] These knowledge "islands" enable us to enhance our performance when dealing with faces that bear resemblance to those with which we are already acquainted.

A computational model for face learning and similarity

Last, we employed a DNN model to investigate potential computational mechanisms that could give rise to the effects of face learning and similarity we described. In our previous study,[35] we showed that DNN units can discriminate faces that were not involved in the training, indicating that DNNs generalize to unfamiliar faces. The faces not included in the original training of the DNN are equivalent to unfamiliar faces to humans. Therefore, DNNs may offer a computational model to understand face learning.

We first fine-tuned a VGG-Face model step by step to simulate the face learning process (Figure 7A; STAR Methods). We updated the DNN each time with 3 images of each of the 50 identities from the same CelebA dataset. The representational distance between faces of DNN units increased with updating (Figures 7B and 7C; output layer: 1,225 of 1,225 face pairwise distances increased from iteration 1 to 2, 1,140 of 1,225 face pairwise distances increased from interaction 2 to 3), suggesting that face discriminability was sharpened by training. This result was consistent with the increasing representational distance

Please cite this article in press as: Cao et al., Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus, Cell Reports (2023), https://doi.org/10.1016/j.celrep.2023.113520

CellPress
OPEN ACCESS

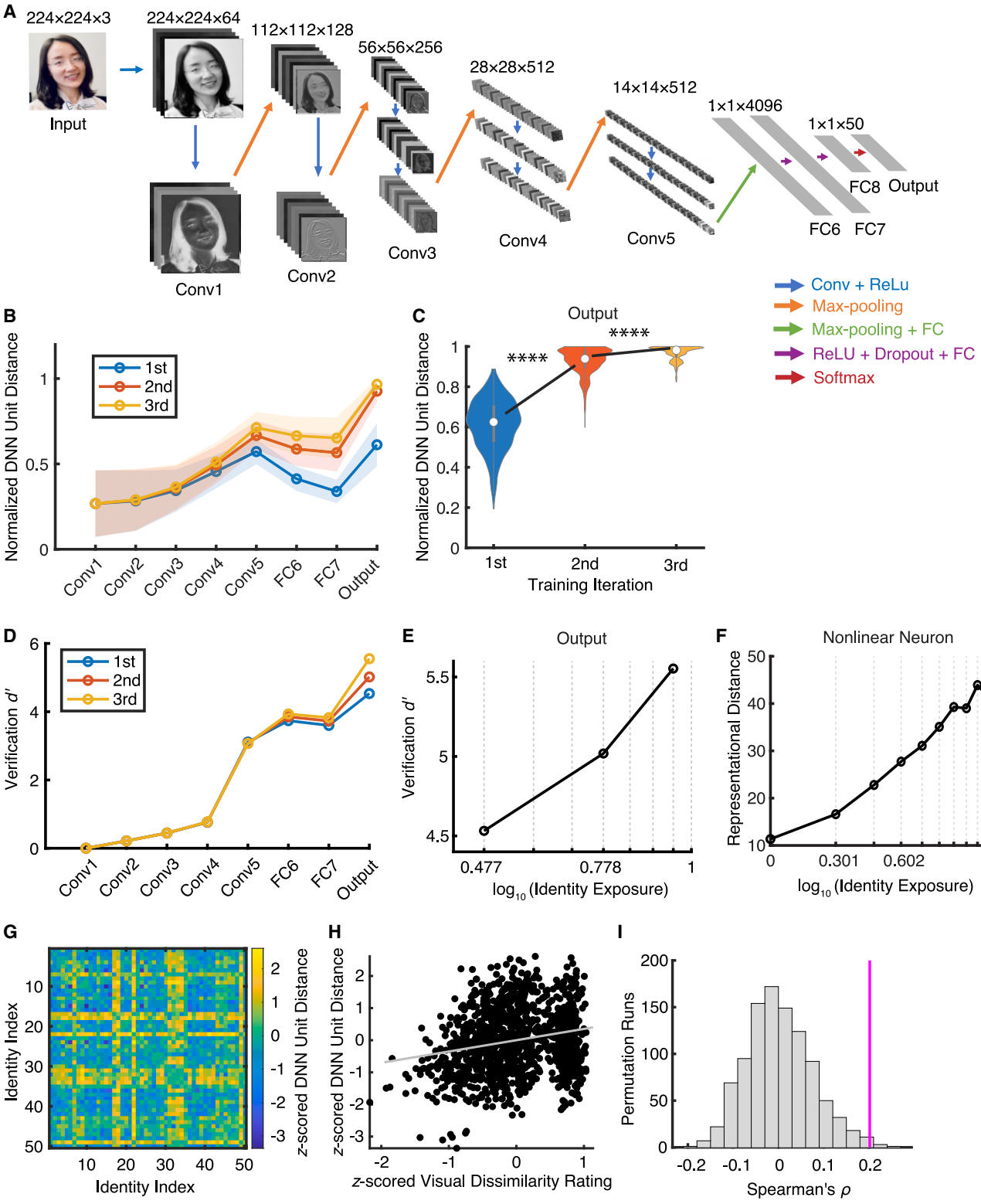


Figure 7. Computational modeling for face learning

(A) Structure of the DNN (see STAR Methods for details). For copyright reasons, we replaced the original stimulus images with similar pictures.

(B and C) Representational distance of DNN units as a function of training iteration (STAR Methods).

(B) Across DNN layers. Error shade denotes ±SD across identity pairs.

(legend continued on next page)

Please cite this article in press as: Cao et al., Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus, Cell Reports (2023), https://doi.org/10.1016/j.celrep.2023.113520

with learning (Figure 4A), suggesting a possible computational mechanism for face learning. Furthermore, the larger the representational distance, the more distinct the neural face representations, leading to a better face recognition performance (accuracy [mean ± SD]: iteration 1: 0.94 ± 0.020; iteration 2: 0.98 ± 0.020; iteration 3: 0.993 ± 0.0115). In addition, we observed a very similar pattern of verification performance (including a sharp increase in the output layer) for familiar faces (Figure 7D) as well as a similar, roughly log-linear learning curve (Figure 7E), as reported in Blauch et al.[26] Notably, we observed a population of neurons in the human amygdala and hippocampus that exhibited a nonlinear response to face learning (n = 151, binomial $p < 10^{-20}$; STAR Methods), and these neurons had a similar, roughly log-linear learning curve in representational distance (Figure 7F).

Using a pre-trained VGG-Face model, we found that DNN unit representational distance between face pairs correlated with human similarity ratings (Figures 7G–7I; $r(1,255) = 0.20$, $p = 1.2 \times 10^{-12}$; permutation p = 0.007), suggesting that representational distance can be the computational basis for visual similarity. While visual similarity is a distinct aspect of face representations, our modeling results indicate that fine-tuning an existing network can enhance the distinctiveness of face representations by reducing their visual similarity.

Together, our results suggest computational mechanisms underlying face learning and similarity.

DISCUSSION

In this study, we identified neurons in the human amygdala and hippocampus that encoded face familiarity. Familiar faces had a greater representational distance between identities compared with unfamiliar faces. We further showed that, with face learning and familiarization, the representational distance increased as a function of identity exposure within a session, and it was the case for both familiar and unfamiliar faces. Representational distance further increased across sessions, but the neuronal population geometry also changed. Moreover, representational distance encoded visual similarity of faces, and notably, face learning could be acquired with visually similar faces. We finally constructed a computational model to account for the neural findings and elucidated the computational principles underlying face learning and similarity. We revealed a neural computational mechanism based on representational distance for face familiarity, familiarization, and similarity.

Representational distance explains three aspects of neural face representation

Our present results can be interpreted in the framework of pattern separation,[24,25] the process of transforming similar representations or memories into dissimilar, non-overlapping representations. Specifically, we showed that representational distance changed as a function of three separate but related aspects of neural face representation (i.e., familiarity, learning, and similarity). First, the representational distance was larger between pairs of familiar faces compared with unfamiliar faces, suggesting that familiar faces were more neurally distinct and had a greater pattern separation compared with unfamiliar faces. Recent theories suggest that familiar faces have a more robust representation in memory because they have been encountered over a wide variety of contexts and image changes. In contrast, unfamiliar faces are encountered only once, and so they do not benefit from such richness of experience and are represented based on image-specific details.[36] Second, the representational distance increased during face familiarization, indicating that the neural representation of faces was amplified and that, as a result, pattern separation was enhanced during learning, similar to the tuning sharpening in the primate inferior temporal cortex.[37,38] The greater representational distance suggests that individual neurons can better distinguish different identities, which is also reflected by the stronger correlation with the DNN pre-trained to distinguish face identities. Our results are consistent with the idea that deeper or more elaborate processing of faces leads to a better recognition of faces,[2] which has also been shown in our simulation results (Figure 7). Third, the representational distance between similar faces was smaller compared with that between visually distinct faces (notably, this was also the case for unfamiliar FBI faces). The neuronal representational distance varies as a function of face familiarity, visual similarity, and learning, suggesting that the representation of face space in the human amygdala and hippocampus is shaped by pattern separation processes.[24,25]

It is worth noting that these three aspects of neural face representation are consistent with each other. Faces become familiar after learning and familiarization, and the representational distance between faces increases. Furthermore, visually similar faces are more difficult to be discriminated because the representational distance is smaller, but they can be better discriminated when they are familiar or familiarized because the representational distance becomes greater. Interestingly, exposure to visually similar faces can also induce learning. Therefore, representational distance can be a generic code to explain neural face representation.

(C) The output layer of the DNN. Asterisks indicate a significant difference using two-tailed paired t test. ****$p < 0.0001$. Legend conventions are as in Figure 2 for violin plots.
(D) Verification *d'* estimated with an ROC-based analysis. We followed the same procedure as described in Blauch et al.[26]
(E) Verification performance as a function of identity exposure in the output (i.e., explicit probability) layer.
(F) Neuronal representational distance as a function of identity exposure. Shown are neurons that had a significant logistic response to identity exposures (n = 151). A $\log_{10}$ x scale is plotted against a linear y scale.
(G) Distance matrix for DNN features (i.e., DNN units). Color coding shows *Z*-scored Euclidean distance between each identity pair.
(H) Correlation between distance in visual dissimilarity rating (the negative of *Z*-scored similarity ratings) and DNN unit distance (*Z* scored) across pairs of face identities. Ratings from the general controls were averaged across participants. Each dot represents a pair of face identities (n = 1,225), and the gray line denotes the linear fit ($r(1,255) = 0.20$, $p = 1.2 \times 10^{-12}$).
(I) Observed vs. permuted correlation coefficient between distance matrix for visual similarity ratings and distance matrix for DNN units. Legend conventions are as in Figure 5.

Please cite this article in press as: Cao et al., Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus, Cell Reports (2023), https://doi.org/10.1016/j.celrep.2023.113520

Neuronal population geometry

The population representational distance increased as a function of identity exposure (i.e., face learning), even across sessions and for visually similar face identities. Interestingly, the angle between the neuronal vectors did not change for identity exposures *within* a session, but changed for identity exposures *across* sessions and for learning of visually similar faces. The change in representational geometry is consistent with a recent study showing that familiar faces are represented in a distinct subspace from unfamiliar faces in the monkey face patch AM and perirhinal cortex and that the familiar face subspace is distorted to increase neural distances between faces.[39] In addition, analyzing the temporal sequence of face presentations demonstrates an early separation of neural patterns between highly familiar and unfamiliar faces in the face-selective regions fusiform face area (FFA) and occipital face area (OFA),[11] consistent with our present results.

The observed disparity in representational geometry changes between within-session identity exposure and across-session identity exposure indicates the involvement of distinct underlying mechanisms. One possible explanation for the consistent representational geometry during within-session identity exposure is the maintenance of a stable internal representation or template for familiar faces. When a face becomes familiar within a session, the neural system can establish a consistent encoding of its unique identity characteristics, resulting in a uniform representational geometry across repeated exposures. This stability is likely crucial for robust recognition and efficient processing of familiar faces within a short time frame. Conversely, the changes in representational geometry during across-session identity exposure suggest a flexible adaptation process that occurs over longer periods. As faces are encountered repeatedly across different sessions, the neural system may gradually adjust to optimize the representation and discrimination of distinct identities. These adjustments might involve fine-tuning of neural responses to better capture the specific features that differentiate individual faces, leading to alterations in the representational geometry. To gain a comprehensive understanding of these patterns, further research is warranted to investigate the precise neural mechanisms at play.

Computational modeling of face learning

Inspired by the primate visual system, DNNs have not only made impressive progress in recognizing faces[40] but also contributed significantly to the understanding of neural face coding[27,35] (see O'Toole et al.[41] and O'Toole and Castillo[42] for reviews). In this study, we constructed a DNN-based computational model that explained the process of face familiarization and learning. A similar computational approach using the same VGG-16 neural network has been applied in a previous study that simulated familiarization with face identities by fine-tuning the network on images of unfamiliar identities.[26] It has been shown that familiarization leads to a sharp improvement in verification performance, but it reaches near-optimal performance levels only in networks that have received extensive training on faces.[26] Notably, consistent with our present study and the fact that amygdala and hippocampal neurons correspond to the response of top DNN layers,[27] the sharp familiarity benefit is observed solely at the identity-based output probability layer, independent of alterations in perceptual representations.[26] Furthermore, consistent with our findings, a linear discriminant analysis (LDA) model on identity recognizes unseen ambient images of familiar (trained) faces but not unknown (untrained) faces.[43] The LDA model also reveals facial features relevant to face learning[43] and indicates that the concept of face familiarity can be better understood as the development of more robust statistical descriptions of unique within-person variations.[44] However, the DNN-based approach is a much more powerful tool to study neural face representation.[27] It is also worth noting that the VGG-16 DNN can discriminate face identities that are not involved in the training.[35] In addition, in line with our present results, research has shown that the similarity structure found in face-trained DNNs is consistent with human similarity ratings.[45] Finally, our recent simulation study systematically investigated the critical period in the development of face processing.[46] It demonstrated the computational mechanisms involved and proposed restoration strategies for early face learning.

Relationship with repetition suppression and perceptual learning

Our present study differs from previous studies that have focused on the repetition suppression phenomenon, which has been observed in the visual cortex in non-human primates at the single-neuron level[47] and in humans using fMRI.[48] Our study found both neurons that increased response and neurons that decreased response as a function of identity exposure within a session. This difference may be because previous studies used familiar stimuli, such as photos of celebrities and familiar individuals, landmark buildings, animals, and objects, and repeated the same images.[49] In contrast, we used unfamiliar faces and different pictures of the same identity to avoid adaptation. Consistent with our results, previous neuroimaging studies have shown concurrent repetition enhancement and suppression responses in the extrastriate visual cortex, which supports predictive coding models that involve the computation of both predictions (which are enhanced by repetition) and prediction errors (which are suppressed by repetition).[50] Our study also showed that the population representational distance between face identities increased over identity exposures both within and across sessions, consistent with previous studies showing that repeated image exposures (e.g., hundreds to thousands) sharpen the tuning of inferior temporal neurons.[37,38] Moreover, our results are consistent with the finding that spaced learning enhances face recognition memory by reducing neural repetition suppression.[51] It is important to note that a mix of repetition suppression and enhancement is typical in the brain, and it may be that repetition warps representational spaces in a general way that is relevant to perceptual learning. Therefore, while some of the effects we observed may be related to repetition, it is also possible that they reflect a general mechanism that could be relevant to both repetition suppression and enhancement.

Our present study is related to perceptual learning,[52,53] where face identities were learned and familiarized with visual exposures. In particular, it has been shown that face learning occurs through prediction errors, which update the level of stimulus familiarity in accordance with predictive coding principles.[54] On

Please cite this article in press as: Cao et al., Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus, Cell Reports (2023), https://doi.org/10.1016/j.celrep.2023.113520

one hand, a ubiquitous property of perceptual learning is specificity; i.e., performance improvement obtained during training applies only to the trained stimulus features,[52] which is consistent with our observed specificity to identity for face exposures. On the other hand, perceptual learning (using low-level stimuli) can be generalized in the absence of sensory adaptation,[55] which is consistent with our face learning through different images (including learning with visually similar faces). In addition, context and expectation may modulate neural response in perceptual learning (see Summerfield and de Lange[56] for a review). A future study is needed to understand the role of expectation in face learning.

Limitations of the study

The task and stimuli employed in this study exhibit intrinsic differences compared with the classic behavioral and simulation studies on face familiarity.[10,26] Given these disparities, it is crucial to note the following limitations associated with interpreting our behavioral and neural findings.

First, in our main experiment, we utilized celebrity faces, and it is important to acknowledge that prior knowledge of familiar faces can influence face perception, particularly in social trait judgment.[57–59] However, we discovered face learning not only for both familiar and unfamiliar celebrity faces but also for visually similar faces using the FBI stimuli (consisting of entirely unfamiliar faces). On the other hand, although we observed differential neural responses to familiar and unfamiliar faces, it is crucial to recognize that the level of familiarity may vary significantly among stimuli and across participants. Notably, the celebrity faces used in our study, while familiar, may not have been personally relevant to the participants. It is possible that the familiarity level associated with celebrity faces was weaker compared with that of personally relevant and familiar faces, such as photographs of the participants themselves, their families, or the experimenters. This disparity in familiarity levels may explain the absence of a familiarity effect in behavior and the presence of learning effects with both familiar and unfamiliar faces (indicating ample room for participants to improve their familiarity). Supporting this notion, previous research has shown that personally relevant faces elicit the strongest selective responses.[23] To further substantiate the learning effect, it is necessary for a future study to control for the level of familiarity and potentially employ personally familiar faces instead of visually familiar ones.

Second, in this study, we employed a one-back task where participants were presented with naturalistic faces and required to judge whether the current stimulus matched the immediately preceding one. This task aimed to mimic real-world scenarios where individuals perceive faces and form instant impressions. As a result, we did not assess face recognition or matching performance to demonstrate improved processing of familiar faces.[1–3] Consequently, we did not observe a behavioral familiarity effect. Specifically, our findings revealed that face familiarity did not modulate attention, as indicated by one-back task accuracy and reaction time (Figures 1B and 1C), nor did it significantly impact social trait judgment (Figure S1A). Eye tracking results showed only weak modulation of eye movements (Figures S1B–S1E). Previous research has shown that, although the number of fixations does not differ between familiar and unfamiliar faces, the fixation locations vary,[60] which aligns with our eye tracking outcomes. Furthermore, in the one-back catch trials, participants were required to press a button after the stimulus disappeared, which resulted in missed responses from some patients. Notably, when we eliminated this requirement, patients exhibited a considerably higher response rate of 93.29% ± 6.82% (mean ± SD across 14 sessions from 11 patients). This may explain why the behavioral performance did not reach its maximum potential.

Finally, while the amygdala and hippocampus are crucial components of the face-processing network, our current study was restricted by clinical limitations, preventing us from examining the broader neural circuitry involved in face processing. Given the distributed nature of face processing in the temporal cortex,[61,62] future studies should explore the neuronal coding of face familiarity and recognition at the network level. We propose the hypothesis that amygdala and hippocampal neurons exhibit categorical responses as opposed to the graded responses observed in the temporal cortex when encoding face familiarity.[19,26] Furthermore, in contrast to the previous finding that familiarity coding and identity coding emerge independently,[15] here we show that identity coding varies as a function of face familiarity, which is likely due to different spatial resolution (EEG vs. single unit) and brain areas recorded (scalp vs. amygdala/hippocampus).

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT PARTICIPANT DETAILS
- METHOD DETAILS
  - Stimuli
  - Experimental procedure
  - Social trait judgment ratings of the CelebA stimuli
  - Electrophysiology
  - Eye tracking
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Response index for single trials
  - Single-neuron ROC analysis
  - Differential latency
  - Representational distance
  - Representational distance in a deep neural network (DNN)
  - Identity-selective neuron
  - Population decoding of face identities
  - Model comparison
  - Computational modeling of face learning
  - Statistics

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2023.113520.

ACKNOWLEDGMENTS

We thank all patients for their participation, staff from WVU Ruby Memorial Hospital for support with patient testing, Jeremy Dawson for contributing the FBI Twins dataset, and Marcus Raichle for discussions and valuable comments. This research was supported by the McDonnell Center for Systems Neuroscience, AFOSR (FA9550-21-1-0088), NSF (BCS-1945230 and IIS-2114644), NIH (R01MH129426, R01MH120194, R01EB026439, U24NS109103, U01NS108916, U01NS128612, R21NS128307, and P41EB018783), and Fondazione Neurone. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

R.C., X.L., and S.W. designed research. R.C., P.B., and S.W. performed experiments. N.J.B. performed surgery. R.C., J.W., U.R., and S.W. analyzed data. R.C., P.B., J.T.W., X.L., U.R., and S.W. wrote the paper. All authors discussed the results and contributed to the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 19, 2022
Revised: September 12, 2023
Accepted: November 14, 2023
Published: December 26, 2023

REFERENCES

1. Hancock, P., Bruce, V., and Burton, A.M. (2000). Recognition of unfamiliar faces. Trends Cognit. Sci. *4*, 330–337.
2. Johnston, R.A., and Edmonds, A.J. (2009). Familiar and unfamiliar face recognition: A review. Memory *17*, 577–596.
3. Young, A.W., and Burton, A.M. (2017). Recognizing Faces. Curr. Dir. Psychol. Sci. *26*, 212–217.
4. Bruce, V., and Young, A. (1986). Understanding face recognition. Br. J. Psychol. *77*, 305–327.
5. Burton, A., Bruce, V., and Hancock, P.J.B. (1999). From pixels to people: A model of familiar face recognition. Cognit. Sci. *23*, 1–31.
6. Bruce, V., Henderson, Z., Greenwood, K., Hancock, P.J.B., Burton, A.M., and Miller, P. (1999). Verification of face identities from images captured on video. J. Exp. Psychol. Appl. *5*, 339–360.
7. Clutterbuck, R., and Johnston, R.A. (2005). Demonstrating how unfamiliar faces become familiar using a face matching task. Eur. J. Cognit. Psychol. *17*, 97–116.
8. Ellis, H.D., Shepherd, J.W., and Davies, G.M. (1979). Identification of Familiar and Unfamiliar Faces from Internal and External Features: Some Implications for Theories of Face Recognition. Perception *8*, 431–439.
9. Klatzky, R.L., and Forrest, F.H. (1984). Recognizing familiar and unfamiliar faces. Mem. Cognit. *12*, 60–70.
10. Young, A.W., and Burton, A.M. (2018). Are We Face Experts? Trends Cognit. Sci. *22*, 100–110.
11. Natu, V.S., and O'Toole, A.J. (2015). Spatiotemporal changes in neural response patterns to faces varying in visual familiarity. Neuroimage *108*, 151–159.
12. Collins, E., Robinson, A.K., and Behrmann, M. (2018). Distinct neural processes for the perception of familiar versus unfamiliar faces along the visual hierarchy revealed by EEG. Neuroimage *181*, 120–131.
13. Dobs, K., Isik, L., Pantazis, D., and Kanwisher, N. (2019). How face perception unfolds over time. Nat. Commun. *10*, 1258.
14. Hahn, C.A., and O'Toole, A.J. (2017). Recognizing approaching walkers: Neural decoding of person familiarity in cortical areas responsive to faces, bodies, and biological motion. Neuroimage *146*, 859–868.
15. Ambrus, G.G., Eick, C.M., Kaiser, D., and Kovács, G. (2021). Getting to Know You: Emerging Neural Representations during Face Familiarization. J. Neurosci. *41*, 5687–5698.
16. Landi, S.M., Viswanathan, P., Serene, S., and Freiwald, W.A. (2021). A fast link between face perception and memory in the temporal pole. Science *373*, 581–585.
17. Squire, L.R., Stark, C.E.L., and Clark, R.E. (2004). The Medial Temporal Lobe. Annu. Rev. Neurosci. *27*, 279–306.
18. Gobbini, M.I., and Haxby, J.V. (2006). Neural response to the visual familiarity of faces. Brain Res. Bull. *71*, 76–82.
19. Ramon, M., Vizioli, L., Liu-Shuang, J., and Rossion, B. (2015). Neural microgenesis of personally familiar face recognition. Proc. Natl. Acad. Sci. USA *112*, E4835–E4844.
20. Quiroga, R.Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. Nature *435*, 1102–1107. http://www.nature.com/nature/journal/v435/n7045/suppinfo/nature03687_S1.html.
21. De Falco, E., Ison, M.J., Fried, I., and Quian Quiroga, R. (2016). Long-term coding of personal and universal associations underlying the memory web in the human brain. Nat. Commun. *7*, 13408.
22. Rey, H.G., Gori, B., Chaure, F.J., Collavini, S., Blenkmann, A.O., Seoane, P., Seoane, E., Kochen, S., and Quian Quiroga, R. (2020). Single Neuron Coding of Identity in the Human Hippocampal Formation. Curr. Biol. *30*, 1152–1159.e3.
23. Viskontas, I.V., Quiroga, R.Q., and Fried, I. (2009). Human medial temporal lobe neurons respond preferentially to personally relevant images. Proc. Natl. Acad. Sci. USA *106*, 21329–21334.
24. Yassa, M.A., and Stark, C.E.L. (2011). Pattern separation in the hippocampus. Trends Neurosci. *34*, 515–525.
25. Leal, S.L., and Yassa, M.A. (2018). Integrating new findings and examining clinical applications of pattern separation. Nat. Neurosci. *21*, 163–173.
26. Blauch, N.M., Behrmann, M., and Plaut, D.C. (2021). Computational insights into human perceptual expertise for familiar and unfamiliar face recognition. Cognition *208*, 104341.
27. Cao, R., Wang, J., Lin, C., Rutishauser, U., Todorov, A., Li, X., Brandmeir, N., and Wang, S. (2020). Feature-based encoding of face identity by single neurons in the human medial temporal lobe. Preprint at bioRxiv.
28. Cao, R., Lin, C., Brandmeir, N.J., and Wang, S. (2022). A human single-neuron dataset for face perception. Sci. Data *9*, 365.
29. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. Roy. Stat. Soc. B *57*, 289–300.
30. Umbach, G., Kantak, P., Jacobs, J., Kahana, M., Pfeiffer, B.E., Sperling, M., and Lega, B. (2020). Time cells in the human hippocampus and entorhinal cortex support episodic memory. Proc. Natl. Acad. Sci. USA *117*, 28463–28474.
31. Oosterhof, N.N., and Todorov, A. (2008). The functional basis of face evaluation. Proc. Natl. Acad. Sci. USA *105*, 11087–11092.
32. Cao, R., Todorov, A., Brandmeir, N.J., and Wang, S. (2022). Task Modulation of Single-Neuron Activity in the Human Amygdala and Hippocampus. eneuro *9*. ENEURO.0398, 21.2021, ENEURO.
33. Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese Neural Networks for One-Shot Image Recognition (France: held in Lille).

34. Hancock, P.J.B. (2021). Familiar faces as islands of expertise. Cognition *214*, 104765.

35. Wang, J., Cao, R., Brandmeir, N.J., Li, X., and Wang, S. (2022). Face identity coding in the deep neural network and primate brain. Commun. Biol. *5*, 611.

36. Chapman, A.F., Hawkins-Elder, H., and Susilo, T. (2018). How Robust Is Familiar Face Recognition? A Repeat Detection Study of More than 1000 Faces, *5* (Royal Society Open Science), pp. 170634.

37. Freedman, D.J., Riesenhuber, M., Poggio, T., and Miller, E.K. (2006). Experience-Dependent Sharpening of Visual Shape Selectivity in Inferior Temporal Cortex. Cereb. Cortex *16*, 1631–1644.

38. Anderson, B., Mruczek, R.E.B., Kawasaki, K., and Sheinberg, D. (2008). Effects of Familiarity on Neural Activity in Monkey Inferior Temporal Lobe. Cereb. Cortex *18*, 2540–2552.

39. She, L., Benna, M.K., Shi, Y., Fusi, S., and Tsao, D.Y. (2021). The Neural Code for Face Memory. Preprint at bioRxiv.

40. Phillips, P.J., Yates, A.N., Hu, Y., Hahn, C.A., Noyes, E., Jackson, K., Cavazos, J.G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. Proc. Natl. Acad. Sci. USA *115*, 6171–6176.

41. O'Toole, A.J., Castillo, C.D., Parde, C.J., Hill, M.Q., and Chellappa, R. (2018). Face Space Representations in Deep Convolutional Neural Networks. Trends Cognit. Sci. *22*, 794–809.

42. O'Toole, A.J., and Castillo, C.D. (2021). Face Recognition by Humans and Machines: Three Fundamental Advances from Deep Learning. Annu. Rev. Vis. Sci. *7*, 543–570.

43. Kramer, R.S.S., Young, A.W., Day, M.G., and Burton, A.M. (2017). Robust social categorization emerges from learning the identities of very few faces. Psychol. Rev. *124*, 115–129.

44. Kramer, R.S.S., Young, A.W., and Burton, A.M. (2018). Understanding face familiarity. Cognition *172*, 46–58.

45. Dobs, K., Martinez, J., Kell, A.J.E., and Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. Sci. Adv. *8*, eabl8913.

46. Wang, J., Cao, R., Chakravarthula, P.N., Li, X., and Wang, S. (2023). A critical period for developing face recognition. Patterns *5*, 100895.

47. Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. Proc. Natl. Acad. Sci. USA *93*, 13494–13499.

48. Henson, R.N.A., and Rugg, M.D. (2003). Neural response suppression, haemodynamic repetition effects, and behavioural priming. Neuropsychologia *41*, 263–270.

49. Pedreira, C., Mormann, F., Kraskov, A., Cerf, M., Fried, I., Koch, C., and Quiroga, R.Q. (2010). Responses of Human Medial Temporal Lobe Neurons Are Modulated by Stimulus Repetition. J. Neurophysiol. *103*, 97–107.

50. de Gardelle, V., Waszczuk, M., Egner, T., and Summerfield, C. (2013). Concurrent Repetition Enhancement and Suppression Responses in Extrastriate Visual Cortex. Cereb. Cortex *23*, 2235–2244.

51. Xue, G., Mei, L., Chen, C., Lu, Z.-L., Poldrack, R., and Dong, Q. (2011). Spaced Learning Enhances Subsequent Recognition Memory by Reducing Neural Repetition Suppression. J. Cognit. Neurosci. *23*, 1624–1633.

52. Sasaki, Y., Nanez, J.E., and Watanabe, T. (2010). Advances in visual perceptual learning and plasticity. Nat. Rev. Neurosci. *11*, 53–60.

53. Sagi, D. (2011). Perceptual learning in Vision Research. Vision Res. *51*, 1552–1566.

54. Apps, M.A.J., and Tsakiris, M. (2013). Predictive codes of familiarity and context during the perceptual learning of facial identities. Nat. Commun. *4*, 2698.

55. Harris, H., Gliksberg, M., and Sagi, D. (2012). Generalized Perceptual Learning in the Absence of Sensory Adaptation. Curr. Biol. *22*, 1813–1817.

56. Summerfield, C., and de Lange, F.P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. Nat. Rev. Neurosci. *15*, 745–756.

57. Gordon, I., and Tanaka, J.W. (2011). The role of name labels in the formation of face representations in event-related potentials. Br. J. Psychol. *102*, 884–898.

58. Schwartz, L., and Yovel, G. (2016). The roles of perceptual and conceptual information in face recognition. J. Exp. Psychol. Gen. *145*, 1493–1511.

59. Oh, D., Walker, M., and Freeman, J.B. (2021). Person knowledge shapes face identity perception. Cognition *217*, 104889.

60. Van Belle, G., Ramon, M., Lefèvre, P., and Rossion, B. (2010). Fixation patterns during recognition of personally familiar and unfamiliar faces. Front. Psychol. *1*, 20.

61. Haxby, J.V., Hoffman, E.A., and Gobbini, M.I. (2000). The distributed human neural system for face perception. Trends Cognit. Sci. *4*, 223–233.

62. Tsao, D.Y., Freiwald, W.A., Tootell, R.B.H., and Livingstone, M.S. (2006). A Cortical Region Consisting Entirely of Face-Selective Cells. Science *311*, 670–674.

63. Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep Learning Face Attributes in the Wild.

64. Grossman, S., Gaziv, G., Yeagle, E.M., Harel, M., Mégevand, P., Groppe, D.M., Khuvis, S., Herrero, J.L., Irani, M., Mehta, A.D., and Malach, R. (2019). Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. Nat. Commun. *10*, 4934.

65. Brainard, D.H. (1997). The Psychophysics Toolbox. Spat. Vis. *10*, 433–436.

66. Lin, C., Keles, U., and Adolphs, R. (2021). Four dimensions characterize attributions from faces using a representative set of English trait words. Nat. Commun. *12*, 5168.

67. Rutishauser, U., Mamelak, A.N., and Schuman, E.M. (2006). Single-Trial Learning of Novel Stimuli by Individual Neurons of the Human Hippocampus-Amygdala Complex. Neuron *49*, 805–813.

68. Rutishauser, U., Ross, I.B., Mamelak, A.N., and Schuman, E.M. (2010). Human memory strength is predicted by theta-frequency phase-locking of single neurons. Nature *464*, 903–907. http://www.nature.com/nature/journal/v464/n7290/suppinfo/nature08860_S1.html.

69. Rutishauser, U., Schuman, E.M., and Mamelak, A.N. (2006). Online detection and sorting of extracellularly recorded action potentials in human medial temporal lobe recordings, in vivo. J. Neurosci. Methods *154*, 204–224.

70. Wang, S., Tudusciuc, O., Mamelak, A.N., Ross, I.B., Adolphs, R., and Rutishauser, U. (2014). Neurons in the human amygdala selective for perceived emotion. Proc. Natl. Acad. Sci. USA *111*, E3110–E3119.

71. Wang, S., Yu, R., Tyszka, J.M., Zhen, S., Kovach, C., Sun, S., Huang, Y., Hurlemann, R., Ross, I.B., Chung, J.M., et al. (2017). The human amygdala parametrically encodes the intensity of specific facial emotions and their categorical ambiguity. Nat. Commun. *8*, 14821. https://www.nature.com/articles/ncomms14821#supplementary-information.

72. Cao, R., Li, X., Brandmeir, N.J., and Wang, S. (2021). Encoding of facial features by single neurons in the human amygdala and hippocampus. Commun. Biol. *4*, 1394.

73. Cao, R., Lin, C., Hodge, J., Li, X., Todorov, A., Brandmeir, N.J., and Wang, S. (2022). A neuronal social trait space for first impressions in the human amygdala and hippocampus. Mol. Psychiatry *27*, 3501–3509.

74. Wang, S., Mamelak, A.N., Adolphs, R., and Rutishauser, U. (2018). Encoding of Target Detection during Visual Search by Single Neurons in the Human Brain. Curr. Biol. *28*, 2058–2069.e4.

75. van der Maaten, L., and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

76. Hutcheon, J.A., Chiolero, A., and Hanley, J.A. (2010). Random measurement error and regression dilution bias. BMJ *340*, c2289.

77. Parkhi, O.M., Vedaldi, A., and Zisserman, A. (2015). Deep Face Recognition.

78. Meyers, E.M. (2013). The Neural Decoding Toolbox. Front. Neuroinform. *7*, 8–12.

79. Rutishauser, U., Ye, S., Koroma, M., Tudusciuc, O., Ross, I.B., Chung, J.M., and Mamelak, A.N. (2015). Representation of retrieval confidence by single neurons in the human medial temporal lobe. Nat. Neurosci. *18*, 1041–1050.

80. Wang, S., Mamelak, A.N., Adolphs, R., and Rutishauser, U. (2019). Abstract goal representation in visual search by neurons in the human pre-supplementary motor area. Brain *142*, 3530–3549.

81. Burnham, K.P., and Anderson, D.R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd Edition (Springer-Verlag).

82. Fried, I., Rutishauser, U., Cerf, M., and Kreiman, G. (2014). Single Neuron Studies of the Human Brain: Probing Cognition (MIT Press).

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and Algorithms		
MATLAB R2022a	MathWorks	https://www.mathworks.com/products/MATLAB.html
OSort	N/A	http://www.urut.ch/new/serendipity/index.php?/pages/osort.html
Psychophysics toolbox PTB3	N/A	http://psychtoolbox.org
Other		
Neuralynx Neurophysiology System	Neuralynx (https://neuralynx.com)	Cat# ATLAS 128
EyeLink Eye Tracker	SR Research (https://www.sr-research.com)	Cat# 1000 Plus Remote
Ad-Tech 8-contact Microelectrode	Ad-Tech (https://adtechmedical.com/subdural-electrodes)	Cat# WB09R-SP00X-014

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Shuo Wang (shuowang@wustl.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

(1) Data for the main experiment has been published as an open-access dataset[28] with detailed descriptions of methods.
(2) The code specific for this study is available on OSF (https://doi.org/10.17605/OSF.IO/ZHQKB).
(3) Any additional information required to analyze the data reported in this paper is available from the lead contact on request.

EXPERIMENTAL MODEL AND SUBJECT PARTICIPANT DETAILS

There were 31 sessions from 9 patients in total (Table S1). We only included patients who indicated whether they recognized each face identity in a follow-up questionnaire.[27] All participants provided written informed consent using procedures approved by the Institutional Review Board of West Virginia University (WVU; protocol #1709745061).

METHOD DETAILS

Stimuli

We used three sets of face stimuli, and we used the same stimuli for all patients.

First, in the main experiment, we used faces of celebrities from the CelebA dataset.[27,63] We selected 50 identities (individuals). For each, we picked 10 different images, resulting in a total of 500 images. The 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). The same stimuli were used for all sessions. It is worth noting that we used different pictures of the same identity to investigate face familiarization. This not only provided a better generalization but also mimicked real world conditions where there is no repeated viewing of an identical face but people perceive the same face identity with different appearances. Using different pictures of the same identity may also reduce stimulus suppression and adaptation. For illustration purposes, alternative facial representations were used in Figure 1 and subjects of the images provided consent for usage of their facial images in this publication.

Second, we used a newly-collected FBI Twins dataset that included pairs of colored 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 are high resolution (3840 × 5760). There

Please cite this article in press as: Cao et al., Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus, Cell Reports (2023), https://doi.org/10.1016/j.celrep.2023.113520

was one face per identity and a total of 144 faces. FBI IRB approval for re-use of images was obtained from select participants in the FBI Twins dataset that provided data during the collections and gave permission for image publication at the time of collection.

Third, we used a FaceGen dataset with model faces, which notably contained only feature information but no real identity information. We used the FaceGen Modeller program (http://facegen.com; version 3.1) to randomly generate 300 faces (see[31] 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 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.

Experimental procedure

We used a 1-back task for the CelebA and FBI stimuli. In each trial, a single face was presented at the center of the screen for a fixed duration of 1 s, with uniformly jittered inter-stimulus-interval (ISI) of 0.5–0.75 s (Figure 1A). 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. 10% of trials were one-back repetitions. Each face was shown once unless repeated in one-back trials; and we excluded responses from one-back trials to have an equal number of responses for each face. This task kept patients attending to the faces, but avoided potential biases from focusing on a particular facial feature (e.g., 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.[64]

For FaceGen stimuli, patients performed two face judgment tasks. In each task, there was a judgment instruction, i.e., 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 center of the screen. One patient performed an additional passive-viewing task. In our previous studies,[27,32] we have shown that we are able to combine data from all tasks for analysis.

Stimuli were presented using MATLAB with the Psychtoolbox 3[65] (http://psychtoolbox.org) (screen resolution: 1600 × 1280).

Social trait judgment ratings of the CelebA stimuli

To examine whether face familiarity modulates social trait judgments (Figure S1A), we acquired trait ratings from patients using a set of social traits that most comprehensively characterize social trait judgments,[66] including *warm*, *critical*, *competent*, *practical*, *feminine*, *strong*, *youthful*, and *charismatic*. These social traits represent the four core psychological dimensions of comprehensive trait judgments of faces (warmth, competence, femininity, and youth; 2 traits per dimension), and they were well validated in the previous study.[66] Patients were asked to rate the faces on eight social traits using a 7-point Likert scale through an online rating task.

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 based solely on clinical need and verified using post-implantation CT. At each site, we recorded from eight 40 μm microwires inserted into a clinical electrode as described previously.[67,68] Efforts were always made to avoid passing the electrode through a sulcus, and its attendant sulcal blood vessels, and thus the location varied but was always well within the body of the targeted area. Microwires projected medially out at the end of the depth electrode and examination of the microwires after removal suggests a spread of about 20–30°. The amygdala electrodes were likely sampling neurons in the mid-medial part of the amygdala and the most likely microwire location is the basomedial nucleus or possibly the deepest part of the basolateral nucleus. Bipolar wide-band recordings (0.1–9000 Hz), using one of the eight microwires as reference, were sampled at 32 kHz and stored continuously for offline analysis with a Neuralynx system. The raw signal was filtered with a zero-phase lag 300–3000 Hz bandpass filter and spikes were sorted using a semi-automatic template matching algorithm as described previously.[69] Units were carefully isolated and recording and spike sorting quality were assessed quantitatively.[28]

Consistent with our previous studies,[27,28,70–73] 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 the CelebA and FBI stimuli, we used the mean firing rate in a time window 250 ms–1250 ms after stimulus onset as the response to each face. For FaceGen stimuli, we used the mean firing rate in a time window 250 ms–1750 ms after stimulus onset as the response to each face.[32]

Eye tracking

Patients were recorded with a remote non-invasive infrared Eyelink 1000 system (SR Research, Canada) (Figures S1B–S1I). One of the eyes was tracked at 500 Hz. The eye tracker was calibrated with the built-in 9-point grid method at the beginning of each block. Fixation extraction was carried out using software supplied with the Eyelink eye tracking system. Saccade detection required a deflection greater than 0.1°, with a minimum velocity of 30°/s and a minimum acceleration of 8000°/s^2, sustained for at least

Please cite this article in press as: Cao et al., Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus, Cell Reports (2023), https://doi.org/10.1016/j.celrep.2023.113520

4 ms. Fixations were defined as the complement of a saccade, i.e., periods without saccades. Analysis of the eye movement record was carried out offline after completion of the experiments.

To quantitatively compare the fixation densities within certain parts of the face, we defined three rectangular ROIs: eyes, mouth, and nose. The fixation density was calculated for each session during the entire 1 s stimulus period, and was normalized within each session. Fixation locations were smoothed using a 2D Gaussian kernel (kernel size = 0.04 * image height by 0.04 * image width) with a standard deviation of 10 pixels.

We excluded sessions that had fewer than 10 fixations onto each facial region of interest (ROI) due to a substantial amount of missing eye tracking data, resulting in a total of 21 sessions for the CelebA stimuli and 16 sessions for the FBI stimuli for further analysis.

QUANTIFICATION AND STATISTICAL ANALYSIS

Response index for single trials

To combine both types of familiarity-selective neurons (i.e., neurons with a greater response to familiar faces or unfamiliar faces) for group analysis, we calculated a single-trial response index (Figure S2A). For each neuron we quantified whether its response differed between familiar faces and unfamiliar faces using a single-trial index, R_i (Equation 1; Equation 2). The R_i facilitates group analysis and comparisons between different types of neurons, as motivated by previous studies.[70,74] The R_i quantifies the response during trial i relative to the mean response to all unfamiliar faces and baseline (the interval right before stimulus onset). The mean response and baseline were calculated individually for each neuron.

Neurons with a greater response to familiar faces:

$$R_i = \frac{FR_i - mean(FR_{Unfamiliar})}{mean(FR_{Baseline})} \cdot 100\% \quad \text{(Equation 1)}$$

Neurons with a greater response to unfamiliar faces:

$$R_i = -\frac{FR_i - mean(FR_{Unfamiliar})}{mean(FR_{Baseline})} \cdot 100\% \quad \text{(Equation 2)}$$

For each trial i, R_i is the baseline normalized mean firing rate (FR) during an interval from 250 ms to 1250 ms after stimulus onset (the same time interval as neuron selection). Different time intervals were tested as well, to ensure that results were qualitatively the same and not biased by particular spike bins.

If a neuron distinguishes familiar faces from unfamiliar faces, the mean value of R_i across all trials/faces will be significantly different from 0. Since neurons with a greater response to familiar faces have more spikes for familiar faces, the mean R_i is positive for these neurons (Equation 1). Since neurons with a greater response to unfamiliar faces have more spikes for unfamiliar faces, to get an aggregate measure of activity that pools across neurons, R_i was multiplied by -1 if the *neuron* has a greater response to unfamiliar faces (Equation 2) so that the mean R_i is also positive for these neurons. Therefore, Equation 1 and Equation 2 make the mean R_i positive for both types of neurons. Notice that the factor -1 in Equation 2 depends only on the *neuron* type but not *trial* type (note that each neuron type has both familiar and unfamiliar face trials).

After calculating R_i for every trial, we subsequently averaged all R_i of trials that belong to the same category. By definition, the average value of R_i for *unfamiliar faces* will be equal to zero because the definition of R_i is relative to the response to unfamiliar faces (see Equation 1; Equation 2); and the average value of R_i for *familiar faces* will be greater than zero. Notably, this is the case for both neurons with a greater response to familiar faces and neurons with a greater response to unfamiliar faces. The mean baseline firing rate was calculated across all trials. The same $FR_{Unfamiliar}$ was subtracted for all types of trials.

The cumulative distribution function (CDF) was constructed by calculating for each possible value x of the R_i how many examples are smaller than x (Figure S2A). That is, $F(x) = P(X \leq x)$, where X is a vector of all R_i values. The CDFs of trials between categories were compared using two-tailed two-sample Kolmogorov-Smirnov (KS) tests.

Single-neuron ROC analysis

We conducted another group analysis to describe how neurons discriminated familiar vs. unfamiliar faces (Figure S2B). Neuronal Receiver Operating Characteristics (ROCs) were constructed based on the spike counts in a time window of 250 ms–1250 ms after stimulus offset. We varied the detection threshold between the minimal and maximal spike count observed, linearly spaced in 20 steps. The Area Under the Curve (AUC) of the ROC was calculated by integrating the area under the ROC curve (trapezoid rule). The AUC value is an unbiased estimate for the sensitivity of an ideal observer that counts spikes and makes a binary decision based on whether the number of spikes is above or below a threshold (Figure S2B). We defined the category with a higher overall firing rate as 'true positive' and the category with a lower overall firing rate as 'false positive'. Therefore, the AUC value was always above 0.5 by definition.

Differential latency

We conducted a spike train analysis to estimate differential response latency of neurons for processing familiar vs. unfamiliar faces (Figures S2D, and S2E). We binned spike trains into 1-ms bins and computed the cumulative sum. To facilitate averaging across

Please cite this article in press as: Cao et al., Neural mechanisms of face familiarity and learning in the human amygdala and hippocampus, Cell Reports (2023), https://doi.org/10.1016/j.celrep.2023.113520

neurons with different firing rates, we normalized the cumulative sum by its maximum value for each neuron. We then averaged the normalized cumulative sums for familiar faces and unfamiliar faces, respectively, and compared, at every point of time, whether the normalized cumulative sums were different between familiar and unfamiliar faces (two-tailed paired *t* test; $p < 0.001$; FDR-corrected). The first point of time of the significant cluster (cluster size >20 time points) was used as the estimate of the differential latency.

Representational distance

We calculated representational distance metrics for *individual* neurons and a *population* of neurons, as follows:

For *individual* neurons (Figures 2; S2; and Figures 3; S5), we used the absolute difference in firing rate to indicate the representational distance for each *face* pair (Figures 2E; S2G; and Figures 3I, 3L; S5Q–S5T). Additionally, we calculated the Euclidean distance in firing rate between 10 faces of an identity to indicate the representational distance for each *identity* pair (Figure S2F). Next, we averaged the representational distance for each neuron and compared them across neurons.

For a *population* of neurons (Figures 4A–4D; Figures 5D, 5I, 5J; and S6; Figures 6; and S7) or DNN units (Figure S2H; Figures 5C and 5F; S7E–S7I; Figure 7), we first obtained a mean response for each identity by averaging across faces, and then calculated the Euclidean distance between neurons for each pair of identities. We obtained similar results when we first calculated the Euclidean distance between neurons for each pair of faces and then averaged the representational distance across face pairs. We also obtained similar results when we used the Pearson correlation coefficient (i.e., $1-r$) as the distance metric.

Change in neuronal population *geometry* (Figures 4B and 4D; S6; Figure 6) was described using the cosine angle between the neuronal vectors: $cos\ \theta = \frac{a \cdot b}{|a||b|}$, where a and b are the neuronal vectors for different conditions.

To visualize the change in representational distance, we applied a t-distributed stochastic neighbor embedding (t-SNE) method[75] to convert high-dimensional features (i.e., 15 PSTH bins) into a two-dimensional feature space. We implemented t-SNE in the MATLAB platform.

We also accounted for the potential impact of the signal-to-noise ratio (SNR) on the observed changes in representational distance, which can lead to regression dilution.[76] The signal was approximated by calculating the variance in neural activity during stimulus presentation (the same neural response used for computing representational distance) across trials. The noise was estimated based on the variance in neural activity during the baseline period. As such, we derived a proxy for the SNR that is relevant for representational distance analysis: [(variance across trials)/(variance during baseline)] − 1. We used this metric as a covariate in the analysis of covariance (ANCOVA) for assessing the familiarity effect and in the partial linear regression for assessing the face learning effect. This allowed us to control for any potential influence arising from changes in SNR.

While we estimated the SNR by comparing the variance in neural responses during stimulus presentation and baseline and obtained similar results after accounting for the SNR, it is important for future studies to investigate the potential influence of SNR on data interpretation more thoroughly. For instance, more precise measurements of SNR in neural recordings could help account for any differences between conditions. To gain a deeper understanding, it would be valuable to simulate the data under various assumptions regarding the distribution of activity levels across neurons and the degree of noise in the system. This would allow us to evaluate whether the reported changes in representational distance and correlation measures would still yield significant results under these conditions. Furthermore, a future study should explore alternative explanations for the observed changes in representational distance, such as differences in attentional allocation or processing strategies between the conditions. These avenues of investigation will provide a more comprehensive understanding of the potential influence of SNR and other factors on the observed changes in representational distance.

Representational distance in a deep neural network (DNN)

We calculated the population representational distance between face identity pairs using responses from DNN units (see section above) and explored whether amygdala and hippocampal neurons shared a similar representational structure as DNN units (Figure S2H; Figures 5C and 5F; S7E–S7I; Figure 7). To obtain DNN unit activation for each face image, we used the well-known DNN implementation based on the VGG-16 convolutional neural network (CNN) architecture[77] (Figure 7A), which has been used in recent work[64] as the computational model for deep face feature extraction and has been validated in our previous studies.[27,35] Specifically, the VGG-16 CNN consisted of a feature extraction section (13 convolutional layers) and a classification section (3 fully connected [FC] layers). The feature extraction section was consistent with the typical architecture of a CNN. A 3 × 3 filter with 1-pixel padding and 1-pixel stride was applied to each convolutional layer, which was followed by a Batch Normalization (BatchNorm) and Rectified Linear Unit (ReLU) operation. Some of the convolutional layers were followed by five 2 × 2 max-pool operations with a stride of 2. There were 3 FC layers in each classification section: the first two had 4096 channels each, and the third performed an *n*-way classification. Each FC layer was followed by a ReLU and 50% dropout to avoid overfitting. A nonlinear Softmax operation was applied to the final output of VGG-16 network to make the classification prediction of 50 identities. We also confirmed that the pre-trained model could discriminate the identities used in the present study and thus serve as a suitable feature extractor (see[27,35] for details).

To show representational similarity between human neurons and DNN units (see Figure S2H for comparison between familiar vs. unfamiliar faces; see Figures 5C, 5F; S7E–S7I; Figure 7 for analysis of visual similarity), we correlated the pairwise representational distance between face identities for neurons with that for DNN units. To determine statistical significance, we used a non-parametric permutation test with 1000 runs. In each run, we randomly shuffled the face labels and calculated the correlation between the neuronal representational distance and the DNN unit representational distance. The distribution of correlation coefficients computed

with shuffling (i.e., null distribution) was eventually compared to the one *without* shuffling (i.e., 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*. We computed the correlation for each DNN layer; and we conducted separate analyses for familiar faces and unfamiliar faces or for each neuronal group (i.e., all neurons, face-responsive neurons, and identity-selective neurons).

Identity-selective neuron

We investigated how face identity coding interacted with familiarity coding (Figure S2G) and face learning (Figures S4G, and S4H). We used our previous procedure to select identity-selective neurons.[27] We first used a one-way ANOVA to identify neurons with a significantly unequal response to different identities. We next imposed an additional criterion to identify the *selected identities*: the neural response of an identity was 2 standard deviations (SD) above the mean of neural responses 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. Note that because identity-selective neurons might change their firing rate for only a few stimuli, an overall response to stimulus onset might not be observed. Therefore, given such sparseness of firing of amygdala and hippocampal 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.

To assess each neuron's selectivity to different identities and compare it with familiarity coding, we defined an identity selectivity index as the d' between the most- and least-preferred identities (Figure S4H):

$$IdentitySelectivityIndex = \frac{\mu_{best} - \mu_{least}}{\sqrt{\frac{1}{2}\left(\sigma^2_{best}+\sigma^2_{least}\right)}}$$

where μ_{best} and μ_{worst} denote the mean firing rate for the most- and least-preferred identities, respectively, and σ^2_{best} and σ^2_{worst} denote the variance of firing rate for the most- and least-preferred identities, respectively. A similar index was used in previous studies to assess the level of selectivity to different faces.[64] It is worth noting that the identity selectivity index was not used to select identity-selective 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-selective neurons that had already been selected.

Population decoding of face identities

We employed a population decoding approach to study face identity coding with different familiarity (Figure S2I). We pooled all recorded neurons into a large pseudo-population. Firing rates were *z*-scored individually for each neuron to give equal weight to each unit regardless of firing rate. We used a maximal correlation coefficient classifier (MCC) as implemented in the MATLAB neural decoding toolbox (NDT).[78] The MCC estimates a mean template for each class *i* and assigns the class for test trial. We used 8-fold cross-validation, i.e., 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 (i.e., 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) that one group of neurons had a greater accuracy than the other. Spikes were counted in bins of 500 ms size and advanced by a step size of 50 ms. The first bin started −500 ms relative to trial onset (bin center was thus 250 ms before trial onset), and we tested 31 consecutive bins (the last bin was thus from 1000 ms to 1500 ms after trial onset). For each bin, a different classifier was trained/tested. We used FDR[29] to correct for multiple comparisons across time points. The same decoding approach was used in our prior studies[79,80] and has been shown to be very effective in the study of neural population activity.

Model comparison

Neurons can respond to face learning with different characteristics. To better understand the response profiles of these neurons with respect to face learning, we further investigated the percentage of neurons that changed firing rate gradually (e.g., as a linear function; Figure 3) or abruptly (e.g., as a step function). We first selected neurons that showed a significantly different response across identity exposures (one-way ANOVA: $p < 0.05$; 348 neurons, 24.82%, binomial $p < 10^{-20}$). We then compared three models: linear regression model, logistic model (sigmoidal), and step-function model, using the Akaike Information Criterion (AIC), which measures the relative quality of statistical models for a given set of data.[81] The AIC is founded on information theory and it offers a relative estimate of the information loss when a given model is used to represent the process that generates the data. In doing so, it deals with the trade-off between the goodness of fit of the model and the complexity of the model. Note that the AIC only estimates the quality of each model relative to the other models in comparison, providing a means for model selection, rather than the absolute quality of the model in a sense of testing a null hypothesis.

For each model, we have: $AIC = n \cdot ln\frac{RSS}{n} + 2k + \frac{2k(k+1)}{n-k-1}$, where n is the sample size (i.e., the number of observations; $n = 10$ here for 10 identity exposures), k is the number of parameters of the model ($k_{Linear} = 2$, $k_{Logistic} = 3$, and $k_{Step} = 3$; see below), and RSS is the

residual sum of squares between the observed data and the fitted data. Note that we here corrected the relatively small sample size ($n/k < 40$).

We fitted a *linear function* of $f(x) = ax + b$, where f is the firing rate and x is the number of identity exposure. a and b were fitted from the observed data (f and x); and thus, k_{Linear} is 2.

We fitted a *logistic function* of $f(x) = \frac{F_{inf}}{1+e^{-\alpha(x-x_{half})}}$, where f is the firing rate, x is the number of identity exposure, F_{inf} is the value when x approaches infinity (the curve's maximum value), x_{half} is the symmetric inflection point (the curve's midpoint), and α is the steepness of the curve. F_{inf}, x_{half}, and α were fitted from the observed data (f and x); and thus, $k_{Logistic}$ is 3.

We fitted a *step function* of $f(x) = a$ when $x \geq c$, and $f(x) = b$ when $x < c$, where f is the firing rate and x is the number of identity exposure. We fitted the parameters using multidimensional unconstrained nonlinear minimization (Nelder-Mead) method to minimize the least squares. a, b, and c were fitted from the observed data (f and x); and thus, k_{Step} is 3.

With the model comparison, we found that the response of 169 neurons (48.56%) could be best explained by a linear model, the response of 58 neurons (16.67%) could be best explained by a logistic model, and the response of 119 neurons (34.20%) could be best explained by a step-like threshold model (separate analysis for familiar faces and unfamiliar faces found similar results).

Computational modeling of face learning

We simulated the face learning process by fine-tuning a DNN model step-by-step (Figures 7A and 7B). We used the pre-trained VGG-Face model as the starting model (Figure 7A), and fine-tuned this model using the same CelebA stimuli as in our main experiment. During fine-tuning, we unlocked all layers of the model, and used a small learning rate for feature extractors (i.e., convolutional layers; 10^{-4}) but a large learning rate for classifiers (i.e., fully connected layers; 10^{-2}). This decision was made with the intention of preserving the original pre-trained model as much as possible for feature extraction, while facilitating the convergence of the model. This approach aimed to strike a balance between extracting informative features from the limited data available and enabling efficient classification. For each iteration, we utilized three images from each of the 50 identities, resulting in a pooled training dataset of 150 images. We conducted a total of three iterations. We used the 10th image of each identity in the dataset to test model recognition performance and calculate representational distance between DNN units. To enhance the accuracy of model performance estimation, we incorporated an additional 5 novel images per identity from the CelebA database during the assessment of model performance. We calculated the Euclidean distance between DNN units for each pair of identities (in total 1225 pairs for 50 identities), as we calculated the representational distance with neurons.

Statistics

We used t-tests to compare conditions/groups (e.g., familiar vs. unfamiliar, different levels of visual similarity) and linear regression to analyze changes in firing rate as a function of face learning. For testing the correspondence between visual similarity matrices, we used Spearman's rank correlation and a non-parametric permutation test to confirm the correlation results. Our statistical threshold was set at $p < 0.05$, and we corrected for multiple comparisons using Bonferroni correction except for the results presented in Figures 5I and 5J. We also used linear mixed effect model to control for nested factors such as session and patient.

We primarily selected two groups of neurons. The first group consisted of neurons whose response differentiated familiar vs. unfamiliar faces using a two-tailed two-sample t test. The second group consisted of neurons whose response varied linearly as a function of identity exposure using linear regression. We also selected face-responsive neurons using a two-tailed paired t test and identity-selective neurons using a one-way ANOVA (see above). For each selection, we used a statistical threshold of $p < 0.05$ for each neuron. Consistent with the vast neurophysiology literature,[82] we did not correct for multiple comparisons at the individual neuron level but used a binomial test to determine whether the number of selected neurons was significantly above chance (5% for $p < 0.05$).

As in our previous studies (e.g.,[70,80]), we used two-tailed two-sample Kolmogorov-Smirnov (KS) tests to compare cumulative distribution functions (Figure S2A) and bootstrap tests to determine significance in decoding analysis (Figure S2I). Due to the large number of temporally correlated data points, we used false discovery rate (FDR)[29] to correct for multiple comparisons for PSTH (Figures 2A–2D), cumulative firing rate (Figure S2D), and decoding time course (Figure S2I).