PNAS RESEARCH ARTICLE | PSYCHOLOGICAL AND COGNITIVE SCIENCES

Behavioral signatures of face perception emerge in deep neural networks optimized for face recognition

Katharina Dobs[a,b,c,d,1], Joanne Yuan[c], Julio Martinez[c,e], and Nancy Kanwisher[c,d,1]

Contributed by Nancy Kanwisher; received December 6, 2022; accepted June 8, 2023; reviewed by Sripati P. Arun and Chris I. Baker

Human face recognition is highly accurate and exhibits a number of distinctive and well-documented behavioral "signatures" such as the use of a characteristic representational space, the disproportionate performance cost when stimuli are presented upside down, and the drop in accuracy for faces from races the participant is less familiar with. These and other phenomena have long been taken as evidence that face recognition is "special". But why does human face perception exhibit these properties in the first place? Here, we use deep convolutional neural networks (CNNs) to test the hypothesis that all of these signatures of human face perception result from optimization for the task of face recognition. Indeed, as predicted by this hypothesis, these phenomena are all found in CNNs trained on face recognition, but not in CNNs trained on object recognition, even when additionally trained to detect faces while matching the amount of face experience. To test whether these signatures are in principle specific to faces, we optimized a CNN on car discrimination and tested it on upright and inverted car images. As we found for face perception, the car-trained network showed a drop in performance for inverted vs. upright cars. Similarly, CNNs trained on inverted faces produced an inverted face inversion effect. These findings show that the behavioral signatures of human face perception reflect and are well explained as the result of optimization for the task of face recognition, and that the nature of the computations underlying this task may not be so special after all.

face perception | deep neural networks | task optimization | 'why' questions

Significance

For decades, cognitive scientists have collected behavioral signatures of face recognition. Here, we move beyond the mere curation of behavioral phenomena to ask why the human face system works the way it does. We find that many classic signatures of human face perception emerge spontaneously in convolutional neural networks (CNNs) trained on face discrimination, but not in CNNs trained on object classification (or on both object classification and face detection), suggesting that these long-documented properties of the human face perception system reflect optimizations for face recognition, not by-products of a generic visual categorization system. This work further illustrates how CNN models can be synergistically linked to classic behavioral findings in vision research, thereby providing psychological insights into human perception.

For over 50 y, cognitive psychologists have documented the many ways that face recognition is "special" (1, 2). Face recognition performance drops disproportionately for inverted faces (i.e., face inversion effect) (3), is higher for faces of familiar than unfamiliar races (i.e., other-race effect) (4), and makes use of a characteristic "face space" (5). These and other behavioral signatures of the face system have been collected and curated as evidence that qualitatively distinct mechanisms are engaged in the recognition of faces compared to other objects. But largely missing from this long-standing literature is the question of *why* the human face recognition system might have these particular properties. Ideal observer methods have long been used to test whether specific behaviorally observed phenomena reflect optimized solutions to simple perceptual tasks, but this method is not well suited for complex real-world tasks (6) like face recognition. Recently, however, task-optimized deep neural networks are providing new traction on this classic question (7). In particular, if a specific human behavioral phenomenon is the expected result of optimization for a given task (whether through evolution or individual experience), then we should observe a similar phenomenon in a deep neural network optimized for that same task. Here, we use this logic to test the hypothesis that the classic behavioral signatures of face perception result specifically from optimization for the task of discriminating one face from another, by testing the prediction that these signatures will be found in convolutional neural networks (CNNs) trained on face recognition, but not in CNNs trained on object categorization or face detection, even when their overall face experience is matched.

Reason to suspect that training on faces may be necessary for CNNs to capture human face perception behavior comes from the ample evidence that face-trained networks perform well on face recognition tasks (8–10). But even if face experience is necessary, it could still be that training on face detection alone (without fine-grained face recognition) is sufficient for specific phenomena in human face perception to emerge. In contrast, reason to suspect that face experience may not be necessary to capture human behavior comes from previous findings that the features learned by CNNs optimized for object recognition are broadly useful for many tasks beyond visual object categorization (11–13) and highly predictive of human perceptual similarity (14). Object-trained networks are even sufficient to predict human behavior on fine-grained letter perception, outperforming networks that are specifically trained on letters (15). Further, object-trained networks are

Author contributions: K.D. and N.K. designed research; K.D. and J.Y. performed research; J.M. contributed new reagents/analytic tools; K.D. and J.Y. analyzed data; and K.D., J.Y., J.M., and N.K. wrote the paper.

Reviewers: S.P.A., Indian Institute of Science Bangalore; and C.I.B., NIH.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: katharina.dobs@psychol.uni-giessen.de or ngk@mit.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2220642120/-/DCSupplemental.

Published July 31, 2023.

currently the best model of face-specific neural responses in the primate brain (16, 17) and even appear to contain units selectively responsive to faces (18, 19). A third possibility is that none of the above training regimes might be able to capture all classic signatures of human face perception, and something else might be required, such as a face-specific inductive bias (20, 21) or a higher-level semantic processing of faces (22), to capture human behavioral signatures of face processing. Finally, these hypotheses are not mutually exclusive, and it is possible that different signatures of human face processing may result from optimization for different tasks.

Here, we tested humans in five different experiments on tasks that measure performance on real-world face recognition, face space, and two of the classic signatures of human face perception: the face inversion effect and other-race effect (see *SI Appendix*, Tables S1 and S2 for details). These behavioral face perception signatures were then directly compared to multiple CNNs based on the same architecture but optimized for different tasks (see *SI Appendix*, Tables S3 and S4 for details): One network was optimized on fine-grained face recognition, one was trained on object categorization only (without face categories in the output layer), one was trained on object categorization and face detection (assigning all faces to one output category), and one was not trained at all (i.e. the same CNN with random weights). A recent study found that face-trained but not object-trained CNNs approached human face recognition accuracy (23). Here, we begin by replicating this phenomenon in a large-scale cohort. We then compared not only overall accuracy but also the representational similarity space between the networks and humans. Next, we asked whether human face signatures emerge only from an optimization for face recognition, or whether an optimization for face detection would suffice. Last, we tested whether a classic face signature—the face inversion effect—is specific for faces per se or whether it can, in principle, emerge for other categories in networks optimized for fine-grained discrimination of those categories (24). Critically, although all CNNs necessarily start with a particular architecture and learning rule, none of the networks had any built-in inductive biases—except those introduced by the different objective functions—to produce these specific behavioral signatures.

Results

Does Humanlike Face Recognition Performance Reflect Optimization for Face Recognition in Particular? One of the most basic properties of human face recognition is simply that we are very good at it. Could our excellent accuracy at face recognition result from generic object categorization abilities, or does it reflect optimization for face recognition in particular? In experiment 1, we measured face recognition performance in people and CNNs using a difficult target-matching task of choosing which of two face images belong to the same identity as a third target image (Fig. 1 *B*, *Left*). Target and nontarget faces were all white females between 20 and 35 y of age, so discrimination of age, gender, or race alone would not suffice for high accuracy on this task. The correctly matching face image differed from the target face image in many low-level features and often in viewpoint, lighting, and facial expression, requiring participants to abstract across these differences to match the identity. Humans (n = 1,532) were tested in a large-scale online experiment using Amazon Mechanical Turk.

Four CNNs were tested on the same task. All four were based on the VGG16 architecture (Fig. 1*A*): one trained to discriminate face identities (Face-ID CNN in red), one trained on object categorization, excluding all animal categories (Obj-Cat CNN in yellow), one trained on object and face categorization (including all

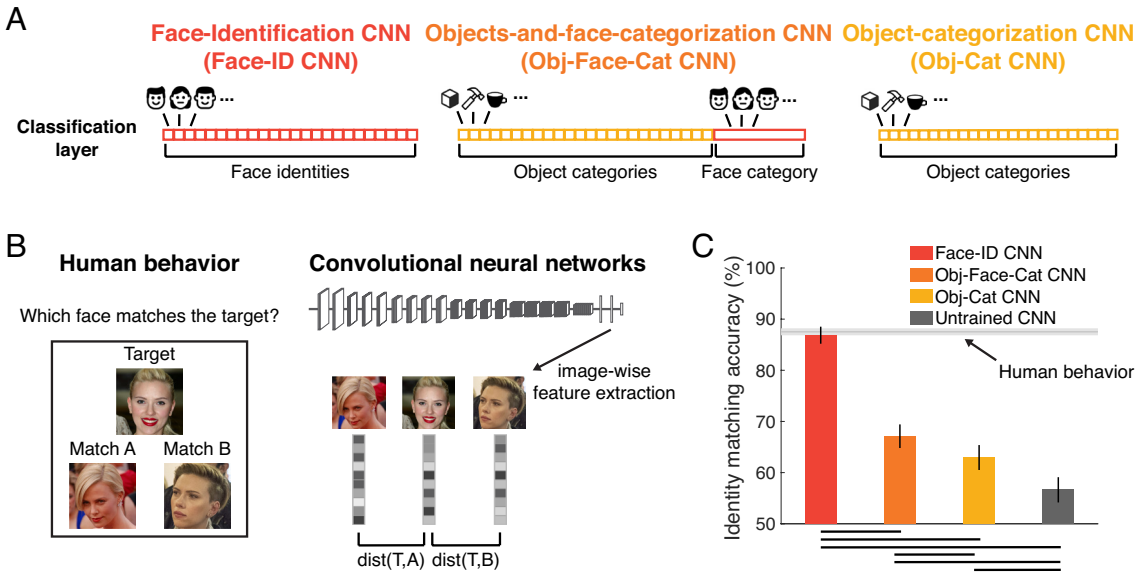

Fig. 1. Experiment 1: Only CNNs trained on face recognition achieve human-level accuracy. (*A*) We compared four CNNs with the VGG16 architecture to human behavior: one trained on face identity recognition (Face-ID CNN, red), one on object and face categorization, with all faces used for training the Face-ID CNN assigned to one category (Obj-Face-Cat CNN, orange), one trained on object categorization only (Obj-Cat CNN, yellow), and another untrained, randomly initialized CNN (Untrained CNN, dark gray). (*B*) Human face recognition performance (n = 1,532) was measured in a target-matching task on 40 female identities (5 images each) on Mechanical Turk. To measure performance in CNNs on the same task, we extracted the activation to each of the images in a trial and computed the correlation distance (1 – Pearson's *r*) between the target (T) and the two matching images (*A* and *B*). The network's choice was modeled as the minimal distance between the target and each of the matching images [e.g., dist(T,A)]. Image credit: Wikimedia Commons/Sgt. Bryson K. Jones, Sgt. Michael Connors, and Chairman of the Joint Chiefs of Staff. (*C*) Human performance was 87.5% correct (light gray horizontal line; the chance level was 50%). Only the face-trained CNN (red) achieved face recognition performance close to humans. Networks trained on object categorization and face detection (orange), or object categorization only (yellow) performed better than the untrained CNN (gray), but did not reach human-level recognition performance. Error bars denote bootstrapped 95% CIs. Black lines in the bottom indicate pairwise significant differences ($P < 0.05$, bootstrap tests, fdr-corrected). Images shown are not examples of the original stimulus set but in public domain and available at https://commons.wikimedia.org.

the faces and object training images, but assigning all face images to a single category; Obj-Face-Cat CNN in orange), and one untrained, randomly initialized CNN (Untrained CNN in gray). We chose the VGG16 architecture (25) because it provides a good fit to neural visual processing (26), it has been successfully trained for face recognition (27) and is widely used in cognitive neuroscience [see *SI Appendix, Supplementary Note 2* for similar results from two other commonly used architectures: Alexnet (28) and ResNet (29, 30)]. In CNNs, we extracted activation patterns from the penultimate fully connected layer (i.e., the decoding stage in a CNN; see *SI Appendix, Supplementary Note 1* for results from other layers) to the same images and computed the correlation distance (1 – Pearson's *r*) between the activation patterns of each pair of images (Fig. 1 *B*, *Right*). The network's choice was determined by which of the two matching images had an activation pattern that was closest to the target image. Importantly, none of the networks was trained on the face identities used as test stimuli.

Human participants were able to correctly match the target face in 87.5% of all trials (1C, light-gray horizontal line; the chance level was 50%). Although we intentionally chose less well-known faces for this test, it remained possible that our participants recognized some of the individuals, possibly inflating performance. To find out, we asked each participant whether one or more of the identities were familiar to them. Indeed, 71% of participants indicated that they were familiar with at least one of the identities, and 10% indicated that they were "not sure". When we ran the same analysis separately on those participants who indicated that they were familiar with at least one identity or not sure (n = 1,233) vs. those that were not familiar with any of the identities (n = 299), we found a significant but small drop in performance from 88.3% to 84.5% (P = 0, bootstrap test). This finding suggests that any contribution of familiarity with particular faces to the observed performance in this task was small (cf. Fig. 3*A* for very similar performance on completely unfamiliar faces).

Might this characteristically high human accuracy on a difficult, high-variance face recognition task, result from a system optimized only for generic object categorization, or would specific optimization for faces be required? We found that CNNs trained on object categorization performed far worse (Obj-Cat CNN: 63.0%; Fig. 1*C*, yellow) than humans, whereas the face-identity-trained CNN achieved human-level performance at 86.9% correct (Face-ID CNN; Fig. 1*C*, red; P = 0.44, bootstrap test), consistent with prior studies (8–10, 23, 31). Does the human-level accuracy of the face identity-trained CNN result from an optimization specifically for face identity discrimination, or would a CNN with the same amount of face experience but optimized for coarse face detection also achieve it? We found that the CNN trained to categorize objects and faces (assigning all faces to a single output category) performed significantly better than the CNN trained on object categorization only (Obj-Face-Cat CNN: 67%; Fig. 1*C*, orange; P < 0.01, bootstrap test), but it performed significantly worse than human performance (P = 0, bootstrap test). The untrained CNN achieved a performance of 56.7%, which was significantly above chance (Fig. 1*C*, gray; P = 0, bootstrap test)

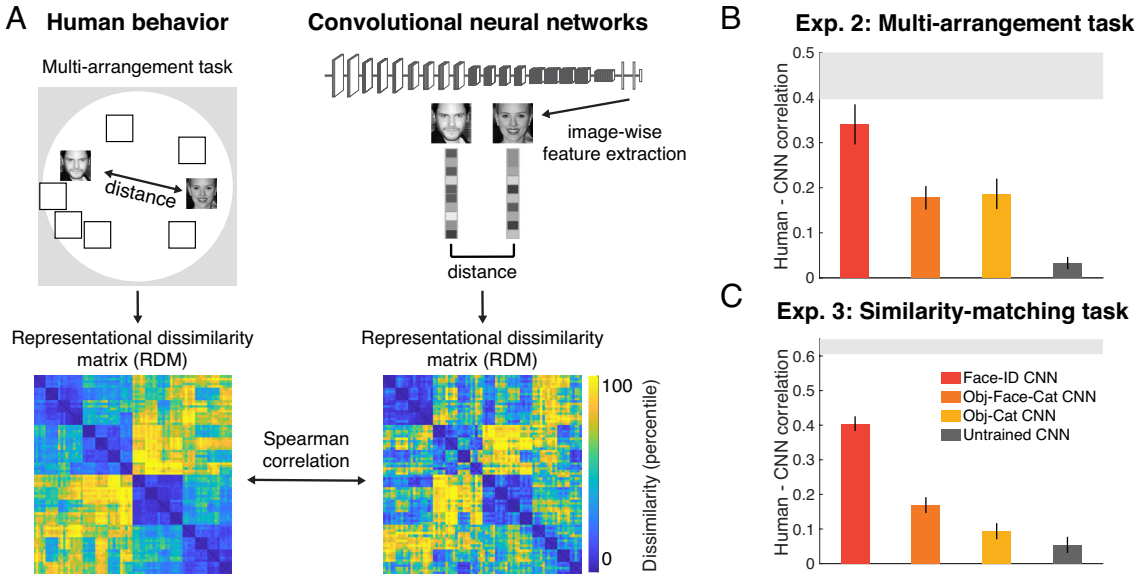

Fig. 2. Experiments 2 and 3: Face-trained but not object-trained CNNs match human face behavior. (*A*) To measure human representational similarities of faces in experiment 2, participants (n = 14) performed a multiarrangement task on 16 face identities (5 images each) resulting in a RDM for each participant (the average RDM shown in the *Bottom Left*). The correlation distance (1 – Pearson's *r*) between activations to each pair of the corresponding images in the penultimate fully connected layer was used to compute the networks' RDM (sample RDM shown for Face-ID CNN in the *Bottom Right*, see *SI Appendix, Supplementary Note 5* for other networks). Spearman rank correlation was used to measure the similarity between human behavioral and networks' RDMs. The 16 face identities were half female, half male and half older, half younger. The low dissimilarity clusters (in blue) along the diagonal of the RDMs correspond to "old female", "young female", "old male", and "young male" identities (from *Top Left* to *Bottom Right*), respectively. Image credit: Wikimedia Commons/ Sgt. Bryson K. Jones and usbotschaftberlin. (*B*) The face-identity-trained CNN (Face-ID CNN, red) matched human behavioral representational similarity best (close to noise ceiling; light gray bar). Neither the untrained CNN (dark gray) nor the object-categorization trained CNN (Obj-Cat CNN, yellow) or the CNN trained to categorize objects and to detect faces (Obj-Face-Cat CNN, orange) matched human representational similarities well. Error bars represent bootstrapped standard error of the mean (SEM) across participants. The gray area represents the noise ceiling. (*C*) The results of (*B*) were replicated in experiment 3 using a similarity-matching task (see Fig. 1*B* for the same stimulus presentation methods but different task: identity matching in experiment 1 but similarity matching here in experiment 3) on Mechanical Turk (n = 668) using a distinct dataset of 60 unfamiliar male identities (one image each). The Face-ID CNN again matched human behavioral representational similarity best, far outperforming the untrained CNN, the Obj-Cat CNN and the Obj-Face-Cat CNN. The corresponding RDMs are shown in *SI Appendix, Supplementary Note 5*. The gray area represents the split-half reliability (mean ± 2*SD across 50 random splits). Error bars denote bootstrapped 95% CIs across dissimilarity values. Images shown are not examples of the original stimulus set but in public domain and available at https://commons.wikimedia.org.

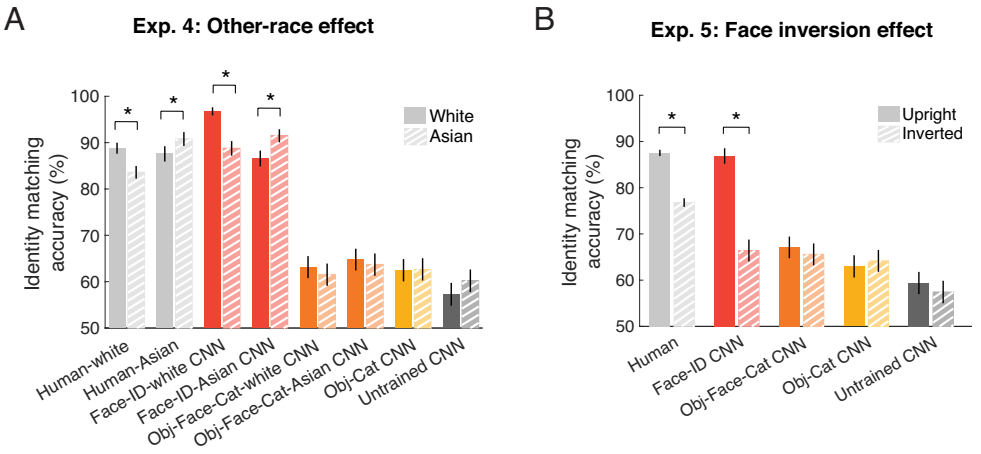

Fig. 3. Face-identity-trained CNNs show humanlike other-race effects (Experiment 4) and face inversion effects (Experiment 5). Face recognition performance was measured in a target-matching task (Fig. 1*B*) in human participants (light gray), multiple face-trained CNNs (red), an object-and-face-trained CNN (Obj-Face-Cat CNN, orange), an object-trained CNN (Obj-Cat CNN, yellow) and an untrained CNN (Untrained CNN, dark gray). In CNNs, activations were extracted from the penultimate fully connected layer to different stimuli sets and compared to human behavior (Fig. 1*B*). (*A*) To measure the other-race effect, performance on unfamiliar, young female white identities (darker bars) was compared to performance on unfamiliar, young Asian female identities (lighter, striped bars). White participants (n = 269) and the face-identity-trained CNN (Face-ID-white CNN; with Asian identities removed from the training set), but not networks trained on object categorization and face detection of only white (Obj-Face-Cat-white CNN) or only Asian faces (Obj-Face-Cat-Asian CNN) or the object-trained or untrained CNN, showed significantly lower performance on Asian than white faces. In contrast, Asian participants (n = 102) and the CNN trained on Asian identities (Face-ID-Asian CNN) showed significantly lower performance on white than on Asian faces. Asterisks indicate significant differences between conditions (*P = 0, bootstrap test). (*B*) Identity matching accuracy for upright (n = 1,532; darker bars) and inverted (n = 1,219; lighter, striped bars) on white female identities from human behavior and CNNs. Only the face-identity-trained CNN (Face-ID CNN) showed the face inversion effect, i.e., lower performance for inverted than upright faces, mirroring human behavior. Error bars denote bootstrapped 95% CIs. Asterisks above bars indicate significant differences between conditions (*P = 0, bootstrap test).

but much lower than human performance (Fig. 1*C*, light-gray horizontal line) or performance of all other trained networks (all $P = 0$, bootstrap tests).

Does the face-identity-trained CNN not just match the overall performance of humans, but also use similar strategies to solve the identity matching task (32, 33)? To address this question, we performed an analysis of the errors being made by humans and CNNs and computed the trial-by-trial predictivity of the human behavioral choices by CNNs (see *SI Appendix, Supplementary Note 3* for details). We indeed found that humans performed significantly better on triplets in which the Face-ID CNN was correct (human performance 88.2%) than on triplets for which it was incorrect (human performance: 83.3%; $P = 0$, bootstrap test). Moreover, the Face-ID CNN predicted the behavioral choices on a trial-by-trial level well (Pearson's r: 0.74), outperforming all other CNNs. These findings suggest that the face-identity-trained CNN not only achieves a similar recognition accuracy to humans, but uses a similar strategy to do so.

Thus, we found that CNNs optimized for face recognition were able to achieve face recognition performance comparable to humans [consistent with prior studies (9, 10, 23)], but this was not the case for CNNs trained on object categorization [despite their high usefulness for other tasks (13, 14)] or untrained CNNs. Further, the CNN trained on both object classification and face detection, which had "experienced" as many faces as the CNN optimized for face recognition but was trained to assign all faces to a single face category, performed far worse than the CNN trained on face recognition. Taken together, these findings suggest that humans' high accuracy at face recognition is not the result of a system optimized for generic object categorization, even with large numbers of faces in the training data, but more likely reflects optimization (through evolution or individual experience) for face recognition in particular.

Do CNNs Represent Faces in a Similar Fashion to Humans? The analyses so far show that CNNs trained on face recognition achieve accuracy levels similar to humans when tested on the same task. But do they achieve this high performance in the same way? To address this question, we assessed the perceived similarity of face images in humans and compared them to CNNs using representational similarity analysis (RSA). Specifically, in experiment 2, we asked whether the similarity between face representations in CNNs resemble those in humans. Human participants (n = 14) performed a multiarrangement task (34) on images belonging to 16 different identities (half female, half male and half older, half younger; 5 images each) for a total of 80 face images (Fig. 2*A*). In this task, participants were asked to place each image in a 2D space that captures similarities in the appearance of faces. Using RSA, we compared the resulting behavioral representational dissimilarity matrices (RDMs) to the RDMs of all four CNNs, obtained by computing the correlation distance between the activation patterns from the penultimate fully connected layer for the same stimuli (see *SI Appendix, Supplementary Note 6* for results from other layers and *SI Appendix, Supplementary Note 7* for similar results from two other architectures: Alexnet and ResNet).

For the face-identity-trained CNN (Exp. 2; Fig. 2*B*; Face-ID CNN in red), correlations between the network's face representations and human behavior were high in the penultimate layer (Spearman's r: 0.34), almost reaching the noise ceiling (i.e., the maximum correlation possible given the consistency across participants; light gray vertical bar). In contrast, the CNN trained object categorization only (Fig. 2*B*; Obj-Cat CNN in yellow) and the CNN trained on object and face categorization (Fig. 2*B*; Obj-Face-Cat CNN in orange) represented faces significantly less similarly to humans (Obj-Cat CNN: Spearman's r: 0.19; Obj-Face-Cat CNN: Spearman's r: 0.21; both $P = 0$, bootstrap test). The representational dissimilarities of the untrained CNN (Fig. 2*B*; Untrained CNN in dark gray) showed a significant but low correlation with human behavior (Spearman's r: 0.03; $P = 0.02$, bootstrap test). Thus, the decoding stage of processing in face-identity-trained CNNs, but not CNNs trained on object categorization, face detection or untrained CNNs, match human behavior well, indicating that faces are similarly represented in human behavior and CNNs optimized for face recognition.

The previous dataset contained multiple images of the same identity, thus human participants might have been biased to simply place images of the same identity together, without taking into account fine-grained details within or between identities. Do these results generalize to other tasks and datasets that rely less on identity recognition? To find out, in experiment 3, we measured representational dissimilarities in humans (n = 668) and CNNs on a completely different dataset using 60 images of distinct, nonfamous young (approximate age between 20 and 30 y) male identities in a similarity-matching task (cf. Fig. 1*B* for the same task but on identity matching instead of similarity matching). We found the same pattern of results (Exp. 3; Fig. 2*C*; see *SI Appendix*, *Supplementary Note 5* for behavioral and CNN RDMs). Specifically, the face-identity-trained CNN was again more similar to human behavioral similarities (Fig. 2*C*; Spearman's *r*: 0.40) than the other three networks (all $P = 0$, bootstrap test; Obj-Cat CNN *r*: 0.09, Obj-Face-Cat CNN r: 0.17, untrained CNN *r*: 0.05, all correlated with behavior above chance, all $P <= 0.02$).

Taken together, face-identity-trained networks, but not networks that were untrained or did not have training on face identification, represented faces similarly to humans [consistent with recent studies (31, 35)], suggesting that optimization for face identification was necessary to match the human face representations tapped in behavioral judgments.

Do CNNs Show Classic Signatures of Human Face Processing? So far, we have found that CNNs optimized for face recognition achieve human-level face recognition performance and represent faces in a similar way, but CNNs optimized for object classification (even when trained extensively on face detection) do not. These findings suggest that human face recognition performance is unlikely to reflect a system optimized (through evolution, individual experience, or both) generically for object categorization and/or face detection alone. Optimization for face recognition in particular seems to be required to capture human face recognition performance. But what about the classic behavioral signatures of human face processing, like the other-race effect and the face inversion effect? Why might human face recognition exhibit these phenomena? Might they also result from optimization for face recognition in particular? If so, we should expect to find that a CNN trained on face recognition, but not CNNs trained on object recognition (even if faces are included as an object category), would exhibit these same phenomena. In both cases, we predict the *presence* of the two signatures in the face-trained network, but not its magnitude, because the networks do not exactly match human experience.

In the other-race effect (36), humans show lower performance recognizing subgroups of faces they have had less exposure to during development. Previous work has suggested that CNNs also show such experience-dependent deficits when specific demographics are underrepresented during training (37–39). But are these effects comparable to the other-race effect in humans, and could they result directly from optimization for recognition of faces of predominantly one race? Or is training of face detection or even passive exposure to faces of one predominant race sufficient? To find out, in experiment 4, we tested white and Asian participants on a set of unfamiliar white and Asian female identities and compared them to CNNs using the same target-matching task (Fig. 1*B*). To test the other-race effect in CNNs, we trained a CNN on face recognition on a dataset of only Asian identities (Face-ID-Asian CNN), and another CNN on a predominantly white dataset with all Asian identities removed (Face-ID-white CNN). Further, we trained two networks on object categorization and face detection using only Asian identities (Obj-Face-Cat-Asian CNN) or only white identities (Obj-Face-Cat-white CNN).

Face recognition performance of the white participants (n = 269) was lower for this Asian female test set (82.6%) set than for the white female test set (86.3%), replicating the other-race effect (Fig. 3*A*; light-gray bars; $P = 0$, bootstrap test across participants). Importantly, Asian participants (n = 102) showed the opposite, performing significantly better on the Asian test set (90.1%) than on the white test set (87.4%; $P = 0$, bootstrap test). We found a significant interaction between the stimulus set and participants' race ($P = 0$, bootstrap test across participants), indicating that this effect cannot be simply due to differences in difficulty between stimuli sets. Our key question was whether this effect can be explained as a direct, perhaps inevitable, result of optimization for face recognition based on the demographically biased samples typical of human experience. Indeed, the recognition performance of the CNN trained on faces with Asian identities removed from the training set (Fig. 3*A*; Face-ID-white CNN) was significantly lower on Asian faces (88.8%) than on white faces (96.8%; $P = 0$, bootstrap test), while the performance of the CNN trained on Asian faces (Face-ID-Asian CNN) was lower for white (86.6%) than for Asian face stimuli (91.5%; $P = 0$, bootstrap test). Despite being matched in face experience, we did not find a significant difference in performance for the CNNs trained on object categorization and face detection of white identities (Obj-Face-Cat-white CNN: 63.1% white vs. 61.5% Asian; $P = 0.51$, bootstrap test) or Asian identities (Obj-Face-Cat-Asian CNN: 64.8% white vs. 63.7% Asian; $P = 0.63$, bootstrap test). Moreover, neither the CNN trained on object categorization nor the untrained CNN showed a significant drop in performance for the Asian set compared to the dataset of white female identities (Obj-Cat CNN: 62.5% white vs. 62.6% Asian; $P = 0.93$, bootstrap test; Untrained CNN: 57.2% white vs. 60.2% Asian; $P = 0.19$, bootstrap test). Overall, these results suggest a reason why humans show an other-race effect: It is a natural consequence of training to discriminate faces from a specific race.

Might optimization for recognition of (upright) faces similarly explain why humans show a face inversion effect? To test this prediction, in experiment 5, we used the target-matching task with the white female identities used before (Fig. 1*B*), but we presented them upside-down to both humans (n = 1,219) and networks [Fig. 3*B* and see Fig. 4 for similar results using SVM (support vector machine) decoding]. Replicating multiple prior studies (3), human participants showed lower performance for inverted (76.8%) than for upright faces (87.5%; $P = 0$; bootstrap test). This significant drop was also found as a significant within-participant difference using the subset of participants (n = 364) who performed both the upright and inverted tasks (accuracy inverted: 75.9% vs. upright: 87.5%; $P = 0$, bootstrap test across participants). We found that the face-identity-trained CNN (Fig. 3*B*; red) was the only network whose performance was lower for inverted than upright images (86.9% upright vs. 66.4% inverted; $P = 0$, bootstrap test). Neither the object-categorization trained (Fig. 3*B*; yellow) nor the object-and-face-categorization trained (Fig. 3*B*; orange) or the untrained CNN (Fig. 3*B*; dark gray) showed a significant difference in performance between upright and inverted faces (Obj-Cat CNN: 63% upright vs. 64.2% inverted, $P = 0.30$; Obj-Face-Cat CNN: 67.1% upright vs. 65.6% inverted, $P = 0.24$ Untrained CNN: 59.4% upright vs. 57.4% inverted; $P = 0.08$, bootstrap test). Moreover, the face inversion effect was significantly larger for the face-identity-trained than for all other CNNs (all $P = 0$, bootstrap test). Thus, even though the object-trained network and the object-and-face-categorization trained network were exposed to faces, and much more to upright than inverted faces, neither shows the robust face inversion effect seen in the face-identity-trained network. These findings show that the face inversion effect does not automatically arise from

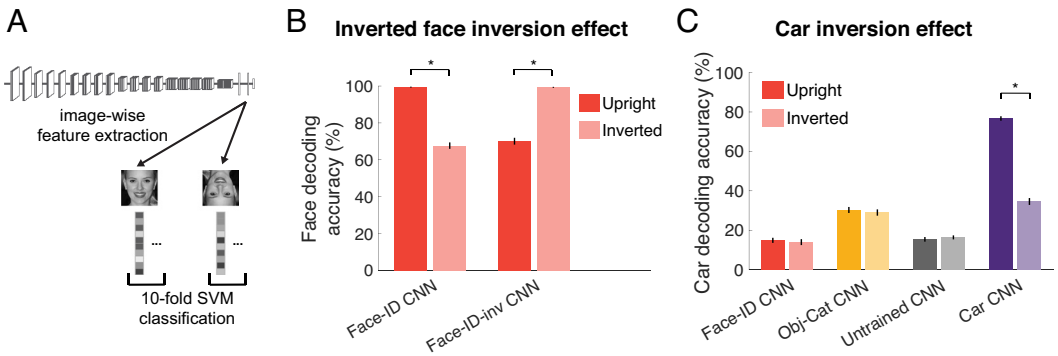

Fig. 4. Inversion effects are not specific to upright faces per se. (*A*) Fine-grained category decoding accuracy was measured in CNNs trained on multiple tasks. Activations were extracted from the penultimate fully connected layer to different stimuli sets shown upright (darker bars) and inverted (lighter bars) and used to train and test a SVM. Image credit: Wikimedia Commons/Sgt. Bryson K. Jones. (*B*) To measure an "inverted face inversion" effect, 100 face identities (10 images each) were decoded from two face-identity-trained CNNs: one trained on upright faces (Face-ID CNN, *Left*) and one trained on inverted faces (Face-ID-inv CNN, *Right*). While the CNN trained on upright faces showed higher accuracy for upright than inverted faces, the CNN trained on inverted faces showed the opposite. (*C*) We decoded 100 car model/make categories (10 images each) from the face-identity-trained CNN (Face-ID CNN, red), the CNN trained on object categorization (Obj-Cat CNN, yellow), the untrained CNN (Untrained CNN, gray), and a CNN trained to categorize car models/makes (Car CNN, purple). Only the car-trained CNN showed an inversion effect for cars, i.e., lower performance for inverted than upright cars. Error bars denote SEM across classification folds. Asterisks indicate significant differences across conditions (*P < 1e-5, two-sided paired t test).

extensive exposure or even training to detect upright but not inverted faces, but it does result from optimization on (upright) face identification, providing a likely explanation of why humans show this effect.

How Special Is the Face Inversion Effect? If indeed optimization for the recognition of upright faces is sufficient to produce a face inversion effect, hence explaining why humans show this phenomenon, does this reflect something special about face stimuli per se? Or, might any stimulus category produce inversion effects given sufficient training at fine-grained discrimination of exemplars within a category (2)? This question has long been pursued in the psychology literature (40), but the evidence that face-like inversion effects can result from perceptual expertise is mixed (41, 42). On the other hand, few, if any, humans have as much perceptual expertise on another stimulus category as they do for faces, and it remains unclear whether face-sized behavioral inversion effects might arise for nonface stimuli if they did. But with machine learning methods, we can give a network as much training on nonfaces as on faces. Here, we trained CNNs on inverted faces and on another fine-grained discrimination task (i.e., cars) to measure inversion effects in these networks and compared them to the previously used face-trained, object-trained, and untrained CNNs. To evaluate the performance of these networks, we used a classification approach by training and testing a linear SVM on activations for upright and inverted stimulus sets extracted from the penultimate fully connected layer of these CNNs (Fig. 4*A*). The stimulus sets used here are larger (1,000 images) than those we used before because we were no longer constrained by the limitations of human experiments.

Indeed, we can produce an "inverted face inversion effect" for faces by training the network only on inverted faces (Fig. 4*B*; Face-ID-inv CNN). That is, face identity decoding accuracy of a CNN trained on inverted faces was significantly larger for inverted faces (99.3%) than for upright faces (70.1%; P < 1e-5; two-sided paired t test). In contrast, the decoding accuracy of the Face-ID CNN trained on upright faces showed the regular face inversion effect (Face-ID CNN: 99.4% upright vs. 67.7% inverted; P < 1e-5). Note that the performance on the trained conditions (e.g., performance of upright faces for the Face-ID CNN) was larger than what we found before (i.e., in Fig. 1*C*). This difference could be due to differences in the stimulus sets or the different analysis method (i.e., SVM decoding approach) used here (see *SI Appendix*, Table S3 for an overview), which allows for reweighting the features. Furthermore, we even found that a network trained on car discrimination showed a car inversion effect (Fig. 4*C*; Car CNN: 76.7% upright vs. 34.6% inverted; P < 1e-5), but neither the object-trained (without cars included in the training set; Obj-Cat CNN: 30.3% upright vs. 29% inverted; P = 0.66) nor the face-identity-trained model (Face-ID CNN: 14.9% upright vs. 14% inverted; P = 0.66) or the untrained model (Untrained CNN: 15.4% upright vs. 16.4% inverted; P = 0.66) did. These findings indicate that inversion effects are not specific to faces per se but can in principle arise naturally for other stimulus classes from training on only upright stimuli.

Discussion

Why does human face recognition show the particular behavioral signatures it does? Here, we show that the characteristic human high accuracy, face space, other-race effects, and face inversion effects, are all found in CNNs optimized for face identification (on upright demographically biased training stimuli characteristic of human experience), but none of these effects are found in networks optimized for generic object categorization. Further, none of these effects arise in networks that are trained on face detection only, despite having the same amount of face experience as the face-identity-trained networks. This finding shows that face experience alone is not sufficient to produce these effects. Rather, it is optimization for the specific task of discriminating individual faces from each other that produces these effects. These findings enable us to go beyond the mere documentation of the special signatures of the human face system, to provide an answer to the question of *why* human face recognition exhibits these phenomena. What our findings suggest is that these classic behavioral signatures of face recognition in humans result from optimization for face recognition in particular. Thus we might expect any system optimized on this task to show the same phenomena.

We further find that the most classic signature of the face system, the face inversion effect, need not in principle be restricted to face stimuli. In CNNs trained on cars, we find a car inversion effect, and in CNNs trained only on inverted faces, we find an inverted inversion effect. Similarly, recent work has found an inversion effect for birds in a CNN trained on bird discrimination

(43). This kind of behavioral inversion effect for nonface objects of expertise has long been sought experimentally in humans, but the enterprise has remained inconclusive (40–42), probably because it is very difficult to find any stimulus class for which humans have as much expertise as they do for faces. With CNNs however, we can control exactly how much experience and what task each network is trained on. These methods have enabled us to show that we should not expect faces to be special in the kinds of representations we extract from them. Rather, faces are special in the human brain (44), and in networks jointly trained of face and object recognition (45) in that distinct neurons and network units are engaged in their processing.

Our results also give some hints about the possible origins of the other-race effect. Like our human participants, we found that the CNN trained on predominantly white faces showed a drop in performance for Asian faces. This finding mirrors recent reports of bias in AI face recognition systems (37–39, 46) and suggests a computational account of the other-race effect in humans (36). Thus, achieving high face recognition accuracy in machines (and possibly also humans) requires not only extensive face experience, but extensive experience within each of multiple face types. This finding, along with our finding that the face inversion effect arises spontaneously in CNNs trained to discriminate face identities but not in CNN trained on face detection and/or object classification, accords with other findings showing signatures of human face perception in face-identity-trained networks, such as face familiarity effects (23), the Thatcher illusion (47) and view-invariant identity representations (31, 48).

Of course, CNNs differ from brains in myriad ways, perhaps most strikingly in how they learn. We are not arguing that the human face system develops in the same way CNNs do, certainly not from extensive labeled examples trained with backpropagation (49). Rather, our point is that CNNs allow us to move beyond the mere documentation of behavioral signatures of face processing as curiosities to be collected, to the more interesting enterprise of asking which of these signatures may be explained as a consequence of the computational optimization for face recognition (6, 7).

Despite the consistency of our results with CNNs trained on face recognition, many puzzles remain. For example, given our finding that face-identity-trained networks better explain human face perception, why do object-trained CNNs perform similarly or even better at explaining face-specific neural responses (16, 17)? One possible explanation is that human face behavior might be read out from later stages of neural processing than have been investigated so far in studies examining the correspondence between CNNs and neural responses. This hypothesis is supported by several findings suggesting that face-specific regions in the superior temporal sulcus (50, 51), or areas beyond the core system of face perception (52, 53) may be involved in face identity recognition in humans or monkeys (54). Another explanation could be that the resolution of neuroimaging methods in humans is insufficient to read out identity information in face-specific areas (55). However, methodological limitations of functional magnetic resonance imaging (fMRI) cannot fully explain this discrepancy, because face-trained models also did not outperform object-trained models in predicting human intracranial data (16), which provides higher spatial resolution than fMRI. A third explanation could be that object-specific features can be repurposed for face perception by reweighting the features (as is typically done when building encoding models). However, we recently found that even when training a linear classifier on object-specific features, those features were much less useful for face identification than face-trained features (45). Last, neural face representations might be optimized for fine-grained face discrimination and face detection. While standard face-trained CNNs are trained to discriminate different faces from each other, the face-specific features they develop were not optimized to distinguish faces from objects. This hypothesis could be addressed by training a CNN on fine-grained face recognition and object categorization simultaneously. Indeed, our recent work suggests that networks optimized for both face and object recognition spontaneously segregate face from object processing in the network and are able to capture human behavioral representational space for faces and objects (45). It will be of interest to directly compare this network to neural responses (56). In the future, these questions might be answered by combining human behavior, neural data, and deep neural networks to find out which task optimization best explains neural face responses and where in the brain the face representations tapped in behavioral tasks reside.

Would more complex types of networks better match human face perception? While it is possible that recurrent neural networks (57, 58) or three-dimensional generative models of face perception (17, 35, 59, 60) could also explain our data, it is unlikely that they would consistently outperform face-identity-trained CNNs given the high correspondence (sometimes even reaching noise ceiling) we observed between CNNs and human face perception behavior for most of the signatures. This suggests that simple feedforward CNNs are sufficient for modeling these face signatures. However, while face-identity-trained CNNs matched all signatures qualitatively, not all signatures were quantitatively matched (e.g., the face inversion effect). Furthermore, feedforward CNNs have been shown to not perform as well on all face tasks [e.g., the Hollow-face effect (59)], and it will be critical to study these tasks further. Additionally, our task was designed to test face recognition under relatively high image variation conditions, but it remains possible that tests with even higher image variation would reveal a gap between humans and feed-forward networks.

Our work also provides some clues into the origins of the human face system, by showing that humanlike face recognition can in principle arise from face-specific experience alone, but only if networks are trained to discriminate individual faces from each other. Importantly, however, training on object classification alone, even with extensive experience on face detection appears not to be sufficient. This finding highlights the fact that the behavior of a network depends not only on the training diet, but also on the training task. It remains an open question whether the relevant face experience that shaped the human face recognition system occurred during evolution, or modern individual experience, or (as is usually the case) both.

In sum, our findings, that face-identity-trained but not object-trained models, even when trained on face detection with the same amount of face experience, match many of the classic signatures of human face recognition enable us to explain these signatures of human face processing as the expected result of optimization for this specific task. Our study joins several other recent investigations that use deep neural networks to explain human perceptual phenomena as the result of optimization for a given task (6, 7). Moreover, the existence of special neural populations selectively engaged in face recognition can also be explained as the result of joint optimization for face and object recognition (45). Each of these studies uses CNNs to move beyond the mere characterization of perceptual phenomena to address the more fundamental question of why our perceptual systems work the way they do. This strategy builds upon earlier work using ideal observer methods in perception but enables us to now tackle more complex real-world perceptual problems.

Materials and Methods

Comparing Human Face Recognition Performance in the Target-Matching Task to Task-Optimized CNNs (Experiments 1 and 5).

Participants. In experiment 1, a total of 1,540 individual workers from the online crowdsourcing platform Amazon Mechanical Turk participated in the target-matching tasks (Fig. 1*B*) on white upright stimuli. A total of 8 workers were excluded from the analysis due to overly fast responses (response time in more than five trials <500 ms or more than 10 trials <800 ms). All workers were located in the United States. We asked workers to voluntarily provide their sex, race, and age (i.e., in ranges "18 to 24", "25 to 34", "35 to 44", "45 to 54", "55 to 64", and "65 or older"). The average workers' age was between 25 and 34 y, 57% of workers were female, 42% were male, and 1% reported "other" or did not report their sex. The majority of the workers were white (70%), 15% were Black, 10% were Asian, and 5% reported other or did not report their race.

In addition to the set of workers participating in the target-matching task on upright stimuli (experiment 1), a total of 1,237 individual workers from Mechanical Turk performed the same target-matching task on inverted stimuli (experiment 5). A total of 18 workers were excluded from the analysis due to overly fast responses (response time in more than five trials <500 ms or more than 10 trials <800 ms). Of the remaining workers, 64 workers had also participated in the target-matching task on upright face images (experiment 1). In addition to these 64 workers, we were able to recruit 300 of the workers that participated in the target-matching task on inverted images to also perform the target-matching task on upright images. In total, we recruited 364 workers who performed both the upright and inverted versions of the target-matching task (providing a within-subject comparison).

To avoid familiarity effects with identities, each worker was only allowed to perform one set of 21 trials (using all 40 distinct identities) per task. Some workers were still able to perform more sets of trials due to technical restrictions on Mechanical Turk. In this case, only the first set of trials was included in the analysis. The number of workers was chosen such that each trial was sampled 20 times across workers. This number was chosen based on previous studies that sampled triplets on Mechanical Turk (61).

Stimuli and behavioral target-matching task. To measure human behavioral face recognition performance, participants performed a target-matching task on Mechanical Turk. To construct this task, we chose 5 images of each of 40 identities. To ensure that the task would not merely rely on external face dimensions (e.g., age, gender), we restricted the stimuli to white female identities of similar age (approximately 20 to 35 y). We further tried to choose individuals who were less famous to reduce familiarity effects. Moreover, we chose images of these identities that varied across lighting, hairstyle, pose, and other low- and mid-level image features (as validated by the low performance of early convolutional layers on this task, around 60% correct; see *SI Appendix, Supplementary Note 1*). We first selected identities which fulfilled these criteria from the Labeled Faces in the Wild dataset (62). Since the Faces in the Wild dataset did not contain sufficient identities to fulfill these criteria, we manually supplemented more identities by selecting them from the internet and by using identities from the VGGFace2 dataset. We then randomly chose images to build triplets in which each target identity (2 images) was paired with each other identity as distractor (1 image) for a total of 1,560 (40 × 39) triplets. Critically, none of the test identities were used for training the Face CNN. Also importantly, from the 156,000 possible triplets [40 identities × 5 images × 4 images (same identity) × 195 images (distractor identity)] available, we only used the 1,560 triplets that were presented to the human participants to compute the identity recognition performance of CNNs. On Amazon Mechanical Turk, we asked human participants to choose which of two face images (e.g., matches A or B) belonged to the same identity as the third target image. The position of the matching images (left or right) was pseudorandomized across trials. To reduce perceptual and behavioral noise as much as possible, participants had unlimited time to perform each trial. Each participant performed 20 trials in which 20 distinct identities were paired with the remaining half of the 20 identities such that each identity was shown only once during the set of trials. Each triplet was repeated 20 times across participants, and the average identity matching performance across all triplets and human participants served as measure for human face recognition performance.

To measure human behavioral face recognition performance on inverted faces (experiment 5), participants performed the same target-matching task on Mechanical Turk as described above, using the identical stimuli we used for the target-matching task on upright faces, but now presenting each face upside down.

Untrained and trained CNNs. To explore the role of task optimization for face recognition, we used four CNNs trained on different tasks and with varying amounts of face experience. All CNNs were based on the VGG16 architecture (25) [see *SI Appendix, Supplementary Note 2* for parallel results based on Alexnet (28) and ResNet (29, 30) architectures]. First, to test the performance of an untrained network, we used a randomly initialized network (Untrained CNN). Second, to measure the performance on face recognition on a CNN trained on generic objects, we included a CNN trained on object categorization only (Obj-Cat CNN) using 423 manually sampled object categories of the ILSVRC-2012 database (63). Third, we included a CNN trained on face identity categorization only (Face-ID CNN) using 1,714 identities from the VGGFace2 database (8). Details about the training and test sets of the latter two networks have been described previously (45). Fourth, to test whether mere face experience (without discriminating individual faces) was sufficient to match human face processing, we included a CNN that was matched in the amount of face experience to the face-identity-trained CNN, but trained on coarse face detection only. Specifically, we trained a CNN on object and face categorization (Obj-Face-Cat CNN) on the exact same 423 object categories as the object-trained CNN with one additional category (i.e., 424 categories in total) that included all face images used to trained the face-identity-trained CNN. Note, that this network did not only include the same amount of face images as the face-identity-trained model but was also trained on a much larger training set.

Target-matching task in CNNs. To directly compare the face recognition performance between humans and CNNs, for the behavioral target-matching tasks on face images, we presented the same stimuli to the four different CNNs: the untrained CNN, the object-categorization-trained CNN, the object-and-face-categorization CNN, and the face-identification CNN. For each tasks and stimuli, we simulated the behavioral task in CNNs by extracting activation from the penultimate fully connected layer to each of the 200 images (see *SI Appendix, Supplementary Note 1* for additional layers). We measured the pairwise similarity for each pair of images using correlation distance (1 – Pearson's *r*) between the activation patterns. The network's choice was modeled as which of the two matching images was closest to the target image. Critically, we only used the 1,560 triplets that were presented to the human participants to compute the identity recognition performance. Note that when we compared CNNs to participants who indicated that they were unfamiliar with all identities, we only compared the performance on the triplets performed by those participants. By averaging the choice accuracy (0 or 1 for incorrect or correct, respectively) across all 1,560 triplets, we obtained a corresponding identity matching performance for each of the different CNNs.

To directly compare the face recognition performance between humans and CNNs for the behavioral target-matching tasks on upright and inverted face images, we presented the inverted stimuli to the four different CNNs and extracted activations from the penultimate fully connected layer (see *SI Appendix, Supplementary Note 1* for other layers). All other analyses were identical to the analyses described for the target-matching task on upright face images above. To test whether the face inversion effect was restricted to a particular architecture, we further investigated the face inversion effect in Alexnet and ResNet architectures (*SI Appendix, Supplementary Note 2*).

Using a Multiarrangement Task to Compare Human Perceptual Similarity of Faces to Task-Optimized CNNs (Experiment 2).

Participants. Behavioral data from 14 laboratory participants (7 female; mean age 25.9, SD = 4.33) from a previously published study (64) were used to perform the RSA using the multiarrangement task. As described previously, all participants provided informed, written consent prior to the experiment and were compensated financially for their time. The Massachusetts Institute of Technology (MIT) Committee on the Use of Humans as Experimental Subjects approved the experimental protocol (COUHES No 1606622600).

Stimuli and behavioral representational dissimilarities. To find out whether humans and CNNs represent faces similarly, we performed RSA in two different experiments. The experimental design and stimuli to obtain the behavioral data have been explained in detail previously (64), so here, we just briefly summarize the stimuli and task. Participants performed a multiarrangement task (34) using 80 face stimuli. Stimuli consisted of 5 images of each of 16 celebrities, which varied orthogonally in gender and age, such that half were female and half were male and half of them were young (below ~35 y) and half were old (above ~60 y). Participants performed the multiarrangement experiment online using their own computer. During the task, participants were instructed to arrange different

subsets of the images based on their perceived similarity ("similar images together, dissimilar images apart") by dragging and dropping them in a circle. After the completion of the experiment, the pairwise squared on-screen distances between the arranged images was computed and resulted in an RDM (see Fig. 2*A* bottom left for visualization of the mean behavioral RDM). For each participant, we extracted the lower off-diagonal data from the behavioral RDM to obtain a vector of pairwise dissimilarities used for computing the correlations.

We additionally computed the noise ceiling for the representational dissimilarities given the inconsistencies across participants using a method described previously (65). Briefly, we estimated the upper bound of the noise ceiling as the mean correlation of each participant's vector of perceived dissimilarities with the group mean (including the participant itself). In contrast, the lower bound was computed by taking the mean correlation of each participant with all other participants.

RSA between humans and CNNs. To obtain representational dissimilarities in CNNs, we presented the same stimuli as used for the human participants to the four CNNs. For each CNN, we extracted the activation patterns to each image separately from the fully connected penultimate layer (see *SI Appendix, Supplementary Note 6* for other layers) and computed the correlation distance (1 – Pearson's *r*) between each pair of activation patterns. This resulted in one RDM for each of the four CNNs (see *SI Appendix, Supplementary Note 5* for visualization of the RDMs).

To compute the similarity between the human RDMs and the RDMs obtained for the CNNs, we rank correlated each participant's behavioral dissimilarity vector with the corresponding CNN dissimilarity vectors. The average rank correlation across participants served as similarity measure between human participants and CNNs.

Statistical inference. To measure statistical significance, we used bootstrap tests. Specifically, we bootstrapped the participant-specific dissimilarity vectors 10,000 times and correlated them with the CNN dissimilarity vectors to obtain an empirical distribution of the correlations. The SD of these distributions defined the SEM for the correlation between humans and CNNs. To test for differences (or differences of differences) between correlations, we bootstrapped the participants 10,000 times and computed the mean difference between correlations resulting in an empirical distribution of correlation differences. All *P*-values were derived as explained below (see section *Statistical Inference*).

Using a Similarity-Matching Task to Compare Human Perceptual Similarity of Faces to Task-Optimized CNNs (Experiment 3). To test whether the results from the multiarrangement task from experiment 2 would generalize to a different dataset and task, we tested participants (n = 668) in a similarity-matching task on Amazon Mechanical Turk using a distinct set of 60 unfamiliar male identities (1 image each) from the Flickr-Faces-HQ database (66). We then converted the participants' choices in this task into pairwise dissimilarity values and used RSA to measure the similarity of human behavior with the four trained CNNs. We analyzed the results as in experiment 2 except that we bootstrapped across the behavioral dissimilarity values instead of across participants, since participants contributed to varying degrees to this task. More details about the participants, stimuli and analyses can be found in the *SI Appendix, Supplementary Note 4*.

Comparing the Other-Race Effect between Humans and Task-Optimized CNNs (Experiment 4).

Participants. To study the other-race effect (Fig. 3*A*), we collected data using a different image dataset. Specifically, we collected data on unfamiliar, young white and Asian female stimuli using the target-matching task. To sample white participants (n = 396), we used Amazon Mechanical Turk and only included workers who were located in the United States, who listed their race as white and who reported that during elementary school at least 50% of their peers were white and less than 50% were Asian. Of those workers, 127 had to be excluded due to overly fast responses, overly deterministic responses (more than 65% left or right clicks only, when chance level was 50%) or because their response differed in the catch trial. The average workers' age was between 25 and 44 y, 46% of workers were female, 53% were male and 1% reported other or did not report their sex. Amazon Mechanical Turk does not provide access to workers that are based in East-Asian countries. Therefore, to additionally collect Asian participants on the same task, we used Clickworker (www.clickworker.com, which provides access to workers from some countries in East Asia) to recruit participants and directed them to perform the experiment on the Meadows platform (www.meadows-research.com). We were able to recruit 132 participants who listed their race as Asian and who reported that during elementary school at least 50% of their peers were Asian. Of those participants, 30 had to be excluded using the same exclusion criteria as for the white participants. The average workers' age was between 35 and 44 y, 44% of workers were female, 54% were male, and 2% reported other or did not report their sex.

Stimuli and behavioral target-matching task on white and Asian faces. To test the other-race effect in humans and compare them to CNNs (see section below), we ran the same target-matching task (see experiment 1) on unfamiliar, young, female white and Asian faces. To exclude that familiarity with some of the identities would influence the results, we collected a novel set of 5 images of each of 80 identities (40 of each race) by using photos provided by colleagues, and by sampling photos of identities on Instagram (with less than 2,000 followers). All of the identities were female and between 20 to 35 y of age, and none of them was used as training for the CNNs. For each race, we built triplets in which each target identity (2 images) was paired with each other identity as distractor (1 image) for a total of 1,560 (40 × 39) triplets. During the experiment, each participant performed 20 trials of each race, randomly interleaved, for a total of 40 trials. For each participant, 20 distinct identities of a race were paired with the remaining half of the 20 identities of the same race, such that each identity was shown only once during the set of trials. To measure within-participant reliability, we included an additional trial in which one randomly chosen trial from the set of 40 trials was repeated.

Testing the other-race effect in CNNs. To test the other-race effect in CNNs, we trained four additional CNNs. We first trained a VGG16 architecture on a dataset of mainly white identities (Face-ID-white CNN). To obtain such a dataset, we manually removed all Asian identities from the identities we previously selected from the VGGFace2 dataset. Specifically, we removed 60 Asian identities from the set of 1,714 identities, for a total of 1,654 remaining mainly white identities. We then trained another VGG16 network on Asian identities (Face-ID-Asian CNN) using the Asian Face Dataset (67). We randomly chose 1,654 identities of this dataset to match the number of identities of the Face-ID-white CNN. These identities had a minimum of 105 images per identity (~174 k images). To avoid imbalanced classes, we therefore chose 105 images for each of the identities of both datasets using 100 for training and five images for validating. To further test whether training on face detection would be sufficient for the other-race effect to emerge in CNNs, we additionally trained two VGG16 networks on object categorization and face detection using the white identities only (Obj-Face-Cat-white) or the Asian identities only (Obj-Face-Cat-Asian).

We then presented the same unfamiliar white and Asian face stimuli as used during the behavioral tasks to the four trained CNNs as well as the Obj-Cat CNN and the untrained CNN and extracted activations from the penultimate fully connected layer (see *SI Appendix, Supplementary Note 1* for other layers). All other analyses were identical to the analyses to the target-matching task on upright face images as described above. To test whether the other-race effect was restricted to a particular architecture, we further investigated the other-race effect in Alexnet and ResNet architectures (*SI Appendix, Supplementary Note 2*).

Statistical inference. For all analyses, we used nonparametric statistical tests that do not rely on assumptions about the distributions of the data. For the target-matching tasks, we bootstrapped the combination of paired identities shown as a trial (i.e., 1,560 triplets) 10,000 times and averaged the responses to obtain a distribution of accuracies. The 2.5th and the 97.5th percentiles of this distribution were used as 95% CI for the behavioral and CNN performances. For statistical inference of the differences between performances, we bootstrapped the triplets 10,000 times and computed the mean difference between accuracies resulting in an empirical distribution of performance differences. To test for interaction effects, we performed the same analysis but bootstrapped the difference of the differences 10,000 times. The number of differences (or differences of differences) that were smaller or larger than zero divided by the number of bootstraps defined the *P*-value (i.e., two-sided testing). In the case of within-participant tests, we performed the same analysis but bootstrapped the participants (instead of triplets) 1,000 times to obtain a distribution of performance differences or difference of differences in case of interaction effects. All *P*-values were corrected for multiple comparisons using false discovery rate at a 0.05 level.

Human Experimental Protocol. All human participants collected for the experiments in this study (experiments 1, 3, 4 and 5) provided informed consent and were compensated financially for their time. The experimental protocol was approved by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES No 1806424985) and conducted following all ethical regulations for conducting behavioral experiments.

Extended Testing for Inversion Effects in Task-Optimized CNNs.

Training CNNs. To test whether a CNN trained only on inverted faces (Face-ID-inv CNN) would show an inverted face inversion effect, we additionally trained a CNN on inverted face images. We used the same architecture, training dataset and parameters as for the Face-ID CNN but showed all images inverted instead of upright during training.

Further, to test whether inversion effects could emerge in CNNs trained on a different domain of stimuli, we trained a VGG16 architecture on car model/make discrimination (Car CNN) using the CompCars dataset (68). To obtain enough images per class, we concatenated images of this dataset from the same model/make but of different years into one class. In this fashion, we ended up with 1,109 classes with 45 images for training and 5 images for validating per class (~50 k training and ~5.5 k validation images).

Decoding of visual categories in CNNs. To test whether CNNs trained on varying tasks differ in how much information about a visual category they contain, we decoded exemplars of independent sets of visual categories from activations extracted from those networks (Fig. 4*A*).

To test the inverted face inversion effect (Fig. 4*B*), we used 100 held-out face identities (50 female; 10 images per identity; 1,000 images in total) from the VGGFace2 dataset that were not included in the training set of the CNNs. We presented these images upright and inverted and extracted activations from both the CNN trained on upright faces (Face-ID CNN) and the CNN trained on inverted images (Face-ID-inv CNN).

To test whether the inversion effect was specific to faces, we further tested for an inversion effect for fine-grained car decoding (Fig. 4*C*). We selected 10 images of 100 model/make categories from the CompCars dataset (1,000 images in total) that were not including in the training of any network and extracted activations to those images from the Face-ID CNN, the Obj-Cat CNN (which had no vehicle-related categories in the training set), the untrained CNN and the Car CNN.

For both of these analyses, we extracted the activation in the penultimate fully connected layer of each network to the image sets. For each task and activations from each network, we trained and tested a 100-way linear SVM (with L2 regularization) on the corresponding activation patterns using a leave-one-image-out (i.e., 10-fold) cross-validation scheme. We computed the mean and SEM across classification folds and used two-sided paired t-tests across classification folds to test for differences between decoding accuracies.

Training Parameters for CNNs. For all trained networks (i.e., Face-ID, Face-ID-white, Face-ID-Asian, Face-ID-inv, Obj-Face-Cat, Obj-Face-Cat-white, Obj-Face-Cat-Asian, Obj-Cat, Car CNN), we used similar training parameters as suggested in ref. 25: stochastic gradient descent with momentum with an initial learning rate of 10^{-3}, a weight decay of 10^{-4} and momentum of 0.9. We trained each network for at least 50 epochs (i.e., full passes over the training set) and the learning rate was reduced twice when the training loss saturated to 10^{-4} and 10^{-5}, respectively. All CNNs were trained until the training loss reached saturation before being compared to human data. To update the weights during training, we computed the cross-entropy loss on random batches of 128 images and back-propagated the loss. Each image was scaled to a minimum side length (height or width) of 256 pixels, normalized to a mean and SD of 0.5. For data augmentation, we showed images randomly as gray-scale with a probability of 20% and chose random crops of the size of 224 × 224 pixels (out of the 256 × 256 pixel-sized images) for each image during training. The test images were scaled, normalized, and center-cropped.

Data, Materials, and Software Availability. The code to train computational models has been previously made available at https://github.com/martinezjulio/sdnn (69). All stimuli used for the online experiments and the source data underlying Figs. 1–4 and *SI Appendix*, Figs. S1–S10 are available at https://osf.io/dbks3/ (70). The stimuli used for the laboratory experiment have been previously made available at https://osf.io/gk6f5/ (71).

ACKNOWLEDGMENTS. We thank Martin Hebart for his help and advice in setting up the experiments on Mechanical Turk, Rebecca Saxe for valuable comments on the manuscript, Elizabeth Mieczkowski for help with online experiments and members of the Kanwisher lab for fruitful discussions and feedback. This work was supported a Feodor-Lynen postdoctoral fellowship of the Humboldt Foundation to K.D., the Deutsche Forschungsgemeinschaft (German Research Foundation; project number 222641018–SFB/TRR 135), "The Adaptive Mind", funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research and Art to K.D., NIH grant Grant DP1HD091947 to N.K. and NSF Science and Technology Center for Brains, Minds, and Machines.

Author affiliations: [a]Department of Psychology, Justus Liebig University Giessen, Giessen 35394, Germany; [b]Center for Mind, Brain and Behavior (CMBB), University of Marburg and Justus Liebig University Giessen, Marburg 35302, Germany; [c]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; [d]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139; and [e]Department of Psychology, Stanford University, Stanford, CA 94305

1. M. J. Farah, Is face recognition 'special'? Evidence from neuropsychology. *Behav. Brain Res.* **76**, 181–189 (1996).
2. R. Diamond, S. Carey, Why faces are and are not special: An effect of expertise. *J. Exp. Psychol. Gen.* **115**, 107–117 (1986).
3. R. K. Yin, Looking at upside-down faces. *J. Exp. Psychol.* **81**, 1–5 (1969).
4. R. K. Bothwell, J. C. Brigham, R. S. Malpass, Cross-racial identification. *Pers. Soc. Psychol. Bull.* **15**, 19–25 (1989).
5. T. Valentine, "Face-space models of face recognition" in *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges*, M. J. Wenger, J. T. Townsend, Eds. (Lawrence Erlbaum Associates Publishers, 2001), pp. 83–113.
6. A. J. Kell, J. H. McDermott, Deep neural network models of sensory systems: Windows onto the role of task constraints. *Curr. Opin. Neurobiol.* **55**, 121–132 (2019).
7. N. Kanwisher, M. Khosla, K. Dobs, Using artificial neural networks to ask 'why' questions of minds and brains. *Trends Neurosci.* **46**, 240–254 (2023).
8. Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age" in *IEEE International Conference on Automatic Face & Gesture Recognition* (IEEE Computer Society, 2018), pp. 67–74.
9. Y. Taigman, M. Yang, M. A. Ranzato, L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 1701–1708.
10. P. J. Phillips *et al.*, Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 6171–6176 (2018).
11. R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2014), pp. 580–587.
12. S. Kornblith, J. Shlens, Q. V. Le, "Do better imagenet models transfer better?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 2661–2671.
13. M. Huh, P. Agrawal, A. A. Efros, "What makes ImageNet good for transfer learning?" in *NIPS Workshop on Large Scale Computer Vision Systems* (2016), pp. 1–10.
14. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2018), pp. 586–595.
15. D. Janini, C. Hamblin, A. Deza, T. Konkle, General object-based features account for letter perception. *PLoS Comput. Biol.* **18**, e1010522 (2022).
16. S. Grossman *et al.*, Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nat. Commun.* **10**, 4934 (2019).
17. L. Chang, B. Egger, T. Vetter, D. Y. Tsao, Explaining face representation in the primate brain using different computational models. *Curr. Biol.* **31**, 2785–2795.e4 (2021), 10.1016/j.cub.2021.04.014.
18. J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, H. Lipson, Understanding neural networks through deep visualization. arXiv [Preprint] (2015). https://doi.org/10.48550/arxiv.1506.06579 (Accessed 13 July 2023).
19. S. Xu, Y. Zhang, Z. Zhen, J. Liu, The face module emerged in a deep convolutional neural network selectively deprived of face experience. *Front. Comput. Neurosci.* **15**, 626259 (2021).
20. S. Sutherland, B. Egger, J. Tenenbaum, "Building 3D Morphable models from a single scan" in *1st Workshop on Traditional Computer Vision in the Age of Deep Learning (TradiCV)* (2021).
21. L. J. Powell, H. L. Kosakowski, R. Saxe, Social origins of cortical face areas. *Trends Cogn. Sci.* **22**, 752–763 (2018).
22. A. Shoham, I. Grosbard, O. Patashnik, D. Cohen-Or, G. Yovel, Deep learning algorithms reveal a new visual-semantic representation of familiar faces in human perception and memory. bioRxiv [Preprint] (2022). https://doi.org/10.1101/2022.10.16.512398 (Accessed 13 June 2023).
23. N. M. Blauch, M. Behrmann, D. C. Plaut, Computational insights into human perceptual expertise for familiar and unfamiliar face recognition. *Cognition* **208**, 104341 (2020).
24. C. Rezlescu, A. Chapman, T. Susilo, A. Caramazza, Large inversion effects are not specific to faces and do not vary with object expertise. PsyArXiv [Preprint] (2016). https://doi.org/10.31234/osf.io/xzbe5 (Accessed 13 July 2023).
25. K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition" in *International Conference on Learning Representations* (2015), pp. 1–14.
26. M. Schrimpf *et al.*, Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* **108**, 413–423 (2020).
27. O. M. Parkhi, A. Vedaldi, A. Zisserman, "Deep face recognition" in *Proceedings of the British Machine Vision Conference (BMVC)* (2015), pp. 41.1–41.12.
28. A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet classification with deep convolutional neural networks" in *Advanced NIPS Neural Information Processing System* (2012), pp. 1097–1105.

29. D. Han, J. Kim, J. Kim, *Deep Pyramidal Residual Networks in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)), Conference on Computer Vision and Pattern Recognition (CVPR)*. (2017). pp. 6307–6315.
30. K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.
31. N. Abudarham, I. Grosbard, G. Yovel, Face recognition depends on specialized mechanisms tuned to view-invariant facial features: Insights from deep neural networks optimized for face or object recognition. *Cogn. Sci.* **45**, e13031 (2021).
32. H. Katti, S. P. Arun, Are you from North or South India? A hard face-classification task reveals systematic representational differences between humans and machines *J. Vis.* **19**, 1 (2019).
33. C. M. Funke *et al.*, Five points to check when comparing visual perception in humans and machines. *J. Vis.* **21**, 16–16 (2021).
34. N. Kriegeskorte, M. Mur, Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Front. Psychol.* **3**, 245 (2012).
35. K. M. Jozwik *et al.*, Face dissimilarity judgments are predicted by representational distance in morphable and image-computable models. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2115047119 (2022).
36. C. A. Meissner, J. C. Brigham, Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychol. Public Policy Law* **7**, 3–35 (2001).
37. I. D. Raji *et al.*, "Saving face: Investigating the ethical concerns of facial recognition auditing" in *AAAI/ACM Conference on AI, Ethics, and Society* (2020), pp. 145–151.
38. J. G. Cavazos, P. J. Phillips, C. D. Castillo, A. J. O'Toole, Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Trans. Biom. Behav. Identity Sci.* **3**, 101–111 (2021).
39. J. Tian, H. Xie, S. Hu, J. Liu, Multidimensional face representation in a deep convolutional neural network reveals the mechanism underlying AI racism. *Front. Comput. Neurosci.* **15**, 620281 (2021).
40. I. Gauthier, C. A. Nelson, The development of face expertise. *Curr. Opin. Neurobiol.* **11**, 219–224 (2001).
41. R. Robbins, E. McKone, No face-like processing for objects-of-expertise in three behavioural tasks. *Cognition* **103**, 34–79 (2007).
42. A. Campbell, J. W. Tanaka, Inversion impairs expert budgerigar identity recognition: A face-like effect for a nonface object of expertise. *Perception* **47**, 647–659 (2018).
43. G. Yovel, I. Grosbard, N. Abudarham, Deep learning models challenge the prevailing assumption that face-like effects for objects of expertise support domain-general mechanisms. *Proc. Biol. Sci.* **290**, 20230093 (2023).
44. G. Schalk *et al.*, Facephenes and rainbows: Causal evidence for functional and anatomical specificity of face and color processing in the human brain. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 12285–12290 (2017).
45. K. Dobs, J. Martinez, A. J. E. Kell, N. Kanwisher, Brain-like functional specialization emerges spontaneously in deep neural networks. *Sci. Adv.* **8**, eabl8913 (2022).
46. J. Buolamwini, T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification" in *Proceedings of Machine Learning Research* (2018), pp. 77–91.
47. G. Jacob, R. T. Pramod, H. Katti, S. P. Arun, Qualitative similarities and differences in visual object representations between brains and deep networks. *Nat. Commun.* **12**, 1872 (2021).
48. A. Farzmahdi, K. Rajaei, M. Ghodrati, R. Ebrahimpour, S.-M. Khaligh-Razavi, A specialized face-processing model inspired by the organization of monkey face patches explains several face-specific phenomena observed in humans. *Sci. Rep.* **6**, 25025 (2016).
49. C. Zhuang *et al.*, Unsupervised neural network models of the ventral visual stream. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2014196118 (2021).
50. S. Anzellotti, A. Caramazza, Multimodal representations of person identity individuated with fMRI. *Cortex* **89**, 85–97 (2017).
51. K. Dobs, J. Schultz, I. Bülthoff, J. L. Gardner, Task-dependent enhancement of facial expression and identity representations in human cortex. *NeuroImage* **172**, 689–702 (2018).
52. J. S. Guntupalli, K. G. Wheeler, M. I. Gobbini, Disentangling the representation of identity from head view along the human face processing pathway. *Cereb. Cortex* **27**, 46–53 (2016).
53. N. Kriegeskorte, E. Formisano, B. Sorger, R. Goebel, Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 20600–20605 (2007).
54. S. M. Landi, W. A. Freiwald, Two areas for familiar face recognition in the primate brain. *Science* **357**, 591–595 (2017).
55. J. Dubois, A. O. de Berker, D. Y. Tsao, Single-unit recordings in the Macaque face patch system reveal limitations of fMRI MVPA. *J. Neurosci.* **35**, 2791–2802 (2015).
56. K. Kar, N. Kanwisher, K. Dobs, "Deep neural networks optimized for both face detection and face discrimination most accurately predict face-selective neurons in macaque inferior temporal cortex" in *Conference on Cognitive Computational Neuroscience, Conference on Cognitive Computational Neuroscience* (2023).
57. T. C. Kietzmann *et al.*, Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 21854–21863 (2019).
58. K. Kar, J. Kubilius, K. Schmidt, E. B. Issa, J. J. DiCarlo, Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* **22**, 974–983 (2019).
59. I. Yildirim, M. Belledonne, W. Freiwald, J. Tenenbaum, Efficient inverse graphics in biological face processing. *Sci. Adv.* **6**, eaax5979 (2020).
60. C. Daube *et al.*, Grounding deep neural network predictions of human categorization behavior in understandable functional features: The case of face identity. *Patterns* **2**, 100348 (2021).
61. M. N. Hebart, C. Y. Zheng, F. Pereira, C. I. Baker, Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nat. Hum. Behav.* **4**, 1173–1185 (2020).
62. G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments in workshop on faces" in *"Real-Life" (Detection, Alignment, and Recognition, Images* (2008), pp. 1–11.
63. J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2009), pp. 248–255.
64. K. Dobs, L. Isik, D. Pantazis, N. Kanwisher, How face perception unfolds over time. *Nat. Commun.* **10**, 1258 (2019).
65. H. Nili *et al.*, A toolbox for representational similarity analysis. *PLoS Comp. Biol.* **10**, e1003553-11 (2014).
66. T. Karras, S. Laine, T. Aila, "A style-based generator architecture for generative adversarial networks" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 4401–4410.
67. Z. Xiong *et al.*, "An Asian face dataset and how race influences face recognition" in *Pacific Rim Conference on Multimedia* (2018) pp. 372–383.
68. L. Yang, P. Luo, C. C. Loy, X. Tang, "A large-scale car dataset for fine-grained categorization and verification" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 3973–3981.
69. K. Dobs, J. Martinez, A. Kell, N. Kanwisher, Brain-like functional specialization emerges spontaneously in deep neural networks. Github. https://github.com/martinezjulio/sdnn. Deposited 15 December 2021.
70. K. Dobs, J. Yuan, J. Martinez, N. Kanwisher, Data from "Behavioral signatures of face perception emerge in deep neural networks optimized for face recognition". Available at http://doi.org/10.17605/OSF.IO/DBKS3. Deposited 30 April 2021.
71. K. Dobs, L. Isik, D. Pantazis, N. Kanwisher, Data from "MEG decoding of face dimensions". OSF. Available at https://doi.org/10.17605/OSF.IO/GK6F5. Deposited 4 October 2018.