Published in final edited form as:
*Annu Rev Vis Sci*. 2021 September 15; 7: 543–570. doi:10.1146/annurev-vision-093019-111701.

# Face Recognition by Humans and Machines: Three Fundamental Advances from Deep Learning

**Alice J. O'Toole**[1], **Carlos D. Castillo**[2]

[1]School of Behavioral and Brain Sciences, The University of Texas at Dallas, Richardson, Texas 75080, USA;

[2]Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA;

## Abstract

Deep learning models currently achieve human levels of performance on real-world face recognition tasks. We review scientific progress in understanding human face processing using computational approaches based on deep learning. This review is organized around three fundamental advances. First, deep networks trained for face identification generate a representation that retains structured information about the face (e.g., identity, demographics, appearance, social traits, expression) and the input image (e.g., viewpoint, illumination). This forces us to rethink the universe of possible solutions to the problem of inverse optics in vision. Second, deep learning models indicate that high-level visual representations of faces cannot be understood in terms of interpretable features. This has implications for understanding neural tuning and population coding in the high-level visual cortex. Third, learning in deep networks is a multistep process that forces theoretical consideration of diverse categories of learning that can overlap, accumulate over time, and interact. Diverse learning types are needed to model the development of human face processing skills, cross-race effects, and familiarity with individual faces.

### Keywords

face recognition; deep convolutional networks; human learning; machine learning; facial features; cross-race effects; face space

## 1. INTRODUCTION

The fields of vision science, computer vision, and neuroscience are at an unlikely point of convergence. Deep convolutional neural networks (DCNNs) now define the state of the art in computer-based face recognition and have achieved human levels of performance on real-world face recognition tasks (Jacquet & Champod 2020, Phillips et al. 2018, Taigman et al. 2014). This behavioral parity allows for meaningful comparisons of representations in two

otoole@utdallas.edu .

DISCLOSURE STATEMENT
C.D.C. is an equity holder in Mukh Technologies, which may potentially benefit from research results.

successful systems. DCNNs also emulate computational aspects of the ventral visual system (Fukushima 1988, Krizhevsky et al. 2012, LeCun et al. 2015) and support surprisingly direct, layer-to-layer comparisons with primate visual areas (Yamins et al. 2014). Nonlinear, local convolutions, executed in cascaded layers of neuron-like units, form the computational engine of both biological and artificial neural networks for human and machine-based face recognition. Enormous numbers of parameters, diverse learning mechanisms, and high-capacity storage in deep networks enable a wide variety of experiments at multiple levels of analysis, from reductionist to abstract. This makes it possible to investigate how systems and subsystems of computations support face processing tasks.

Our goal is to review scientific progress in understanding human face processing with computational approaches based on deep learning. As we proceed, we bear in mind wise words written decades ago in a paper on science and statistics: “All models are wrong, but some are useful” (Box 1979, p. 202) (see the sidebar titled Perspective: Theories and Models of Face Processing and the sidebar titled Caveat: Iteration Between Theory and Practice). Since all models are wrong, in this review, we focus on what is useful. For present purposes, computational models are useful when they give us insight into the human visual and perceptual system. This review is organized around three fundamental advances in understanding human face perception, using knowledge generated from deep learning models. The main elements of these advances are as follows.

1. Deep networks force us to rethink the universe of possible solutions to the problem of inverse optics in vision. The face representations that emerge from deep networks trained for identification operate invariantly across changes in image and appearance, but they are not themselves invariant.

2. Computational theory and simulation studies of deep learning indicate a reconsideration of a long-standing axiom in vision science that face or object representations can be understood in terms of interpretable features. Instead, in deep learning models, the concept of a nameable deep feature, localized in an output unit of the network or in the latent variables of the space, should be reevaluated.

3. Natural environments provide highly variable training data that can structure the development of face processing systems using a variety of learning mechanisms that overlap, accumulate over time, and interact. It is no longer possible to invoke learning as a generic theoretical account of a behavioral or neural phenomenon.

We focus on deep learning findings that are relevant for understanding human face processing—broadly construed. The human face provides us with diverse information, including identity, gender, race or ethnicity, age, and emotional state. We use the face to make inferences about a person’s social traits (Oosterhof & Todorov 2008). As we discuss below, deep networks trained for identification retain much of this diverse facial information (e.g., Colón et al. 2021, Dhar et al. 2020, Hill et al. 2019, Parde et al. 2017, Terhörst et al. 2020). The use of face recognition algorithms in applied settings (e.g., law enforcement) has spurred detailed performance comparisons between DCNNs and humans (e.g., Phillips et al. 2018). For analogous reasons, the problem of human-like race bias in DCNNs has also been studied (e.g., Cavazos et al. 2020; El Khiyari & Wechsler 2016; Grother et al. 2019;

Krishnapriya et al. 2019, 2020). Developmental data on infants' exposure to faces in the first year(s) of life offer insight into how to structure the training of deep networks (Smith & Slone 2017). These topics are within the scope of this review. Although we consider general points of comparison between DCNNs and neural responses in face-selective areas of the primate inferotemporal (IT) cortex, a detailed discussion of this topic is beyond the scope of this review. (For a review of primate face-selective areas that considers computational perspectives, see Hesse & Tsao 2020). In this review, we focus on the computational and representational principles of neural coding from a deep learning perspective.

The review is organized as follows. We begin with a brief review of where machine performance on face identification stands relative to humans in quantitative terms. Qualitative performance comparisons on identification and other face processing tasks (e.g., expression classification, social perception, development) are integrated into Sections 2–4. These sections consider advances in understanding human face processing from deep learning approaches. We close with a discussion of where the next steps might lead.

## 1.1. Where We Are Now: Human Versus Machine Face Recognition

Deep learning models of face identification map widely variable images of a face onto a representation that supports identification accuracy comparable to that of humans. The steady progress of machines over the past 15 years can be summarized in terms of the increasingly challenging face images that they can recognize (Figure 1). By 2007, the best algorithms surpassed humans on a task of identity matching for unfamiliar faces in frontal images taken indoors (O'Toole et al. 2007). By 2012, well-established algorithms exceeded human performance on frontal images with moderate changes in illumination and appearance (Kumar et al. 2009, Phillips & O'Toole 2014). Machine ability to match identity for in-the-wild images appeared with the advent of DCNNs in 2013–2014. Human face recognition was marginally more accurate than DeepFace (Taigman et al. 2014), an early DCNN, on the Labeled Faces in the Wild (LFW) data set (Huang et al. 2008). LFW contains in-the-wild images taken mostly from the front. DCNNs now fare well on in-the-wild images with significant pose variation (e.g., Maze et al. 2018, data set). Sengupta et al. (2016) found parity between humans and machines on frontal-to-frontal identity matching but human superiority on frontal-to-profile matching.

**Identity matching:**

process of determining if two or more images show the same identity or different identities; this is the most common task performed by machines

**Human face recognition:**

the ability to determine whether a face is known

## 1.2. Expert Humans and State-of-the-Art Machines Work Together

DCNNs can sometimes even surpass normal human performance. Phillips et al. (2018) compared humans and machines matching the identity of faces in high-quality frontal images. Although this is generally considered an easy task, the images tested were chosen to be highly challenging based on previous human and machine studies. Four DCNNs developed between 2015 and 2017 were compared to human participants from five groups: professional forensic face examiners, professional forensic face reviewers, superrecognizers (Noyes et al. 2017, Russell et al. 2009), professional fingerprint examiners, and students. Face examiners, reviewers, and superrecognizers performed more accurately than fingerprint examiners, and fingerprint examiners performed more accurately than students. Machine performance, from 2015 to 2017, tracked human skill levels. The 2015 algorithm (Parkhi et al. 2015) performed at the level of the students; the 2016 algorithm (Chen et al. 2016) performed at the level of the fingerprint examiners (Ranjan et al. 2017c); and the two 2017 algorithms (Ranjan et al. 2017,c) performed at the level of professional face reviewers and examiners, respectively. Notably, combining the judgments of individual professional face examiners with those of the best algorithm (Ranjan et al. 2017) yielded perfect performance. This suggests a degree of strategic diversity for the face examiners and the DCNN and demonstrates the potential for effective human–machine collaboration (Phillips et al. 2018).

Combined, the data indicate that machine performance has improved from a level comparable to that of a person recognizing unfamiliar faces to one comparable to that of a person recognizing more familiar faces (Burton et al. 1999, Hancock et al. 2000, Jenkins et al. 2011) (see Section 4.1).

## 2. RETHINKING INVERSE OPTICS AND FACE REPRESENTATIONS

Deep networks force us to rethink the universe of possible solutions to the problem of inverse optics in vision. These networks operate with a degree of invariance to image and appearance that was unimaginable by researchers less than a decade ago. Invariance refers to the model's ability to consistently identify a face when image conditions (e.g., viewpoint, illumination) and appearance (e.g., glasses, facial hair) vary. The nature of the representation that accomplishes this is not well understood. The inscrutability of DCNN codes is due to the enormous number of computations involved in generating a face representation from an image and the uncontrolled training data. To create a face representation, millions of nonlinear, local convolutions are executed over tens (to hundreds) of layers of units. Researchers exert little or no control over the training data, but instead source face images from the web with the goal of finding as much labeled training data as possible. The number of images per identity and the types of images (e.g., viewpoint, expression, illumination, appearance, quality) are left (mostly) to what is found through web scraping. Nevertheless, DCNNs produce a surprisingly structured and rich face representation that we are beginning to understand.

### 2.1. Mining the Face Identity Code in Deep Networks

The face representation generated by DCNNs for the purpose of identifying a face also retains detailed information about the characteristics of the input image (e.g., viewpoint, illumination) and the person pictured (e.g., gender, age). As shown below, this unified representation can solve multiple face processing tasks in addition to identification.

#### 2.1.1. Image characteristics.—

Face representations generated by deep networks both are and are not invariant to image variation. These codes can identify faces invariantly over image change, but they are not themselves invariant. Instead, face representations of a single identity vary systematically as a function of the characteristics of the input image. The representations generated by DCNNs are, in fact, representations of face images.

Work to dissect face identity codes draws on the metaphor of a face space (Valentine 1991) adapted to representations generated by a DCNN. Visualization and simulation analyses demonstrate that identity codes for face images retain ordered information about the input image (Dhar et al. 2020, Hill et al. 2019, Parde et al. 2017). Viewpoint (yaw and pitch) can be predicted accurately from the identity code, as can media source (still image or video frame) (Parde et al. 2017). Image quality (blur, usability, occlusion) is also available as the identity code norm (vector length).[1] Poor-quality images produce face representations centered in the face space, creating a DCNN garbage dump. This organizational structure was replicated in two DCNNs with different architectures, one developed by Chen et al. (2016) with seven convolutional layers and three fully connected layers and another developed by Sankaranarayanan et al. (2016) with 11 convolutional layers and one fully connected layer. Image quality estimates can also be optimized directly in a DCNN using human ratings (Best-Rowden & Jain 2018).

**Face space:**

representation of the similarity of faces in a multidimensional space

For a closer look at the structure of DCNN face representations, Hill et al. (2019) examined the representations of highly controlled face images in a face space generated by a deep network trained with in-the-wild images. The network processed images of three-dimensional laser scans of human heads rendered from five viewpoints under two illumination conditions (ambient, harsh spotlight). Visualization of these representations in the resulting face space showed a highly ordered pattern (see Figure 2). Consistent with the network's high accuracy at face identification, images clustered by identity. Identity clusters separated into regions of male and female faces (see Section 2.1.2). Within each identity cluster, the images separated by illumination condition—visible in the face space as chains of images. Within each illumination chain, the image representations were arranged in the space by viewpoint, which varied systematically along the image chain. To further probe the coding of identity, Hill et al. (2019) processed images of caricatures of the 3D heads (see also Blanz & Vetter 1999). Caricature representations were centered in each identity cluster, indicating that the network perceived a caricature as a good likeness of the identity.

**DCNN face representation:**

output vector produced for a face image processed through a deep network trained for faces

[1] This is the case in networks trained with the Softmax objective function.

All results from Hill et al. (2019) were replicated using two networks with starkly different architectures. The first, developed by Ranjan et al. (2019), was based on a ResNet-101 with 101 layers and skip connections; the second, developed by Chen et al. (2016), had 15 convolution and pooling layers, a dropout layer, and one fully connected top layer. As measured using the brain-similarity metrics developed in Brain-Score (Schrimpf et al. 2018), one of these architectures (ResNet-101) was the third most brain-like of the 25 networks tested. The ResNet-101 network scored well on both neural (V4 and IT cortex) and behavioral predictability for object recognition. Hill et al.'s (2019) replication of this face space using a shallower network (Chen et al. 2016), however, suggests that network architecture may be less important than computational capacity in understanding high-level visual codes for faces (see Section 3.2).

**Brain-Score:**

neural and behavioral benchmarks that score an artificial neural network on its similarity to brain mechanisms for object recognition

Returning to the issue of human-like view invariance in a DCNN, Abudarham & Yovel (2020) compared the similarity of face representations computed within and across identities and viewpoints. Consistent with view-invariant performance, same-identity, different-view face pairs were more similar than different-identity, same-view face pairs. Consistent with a noninvariant face representation, correlations between similarity scores across head view decreased monotonically with increasing view disparity. These results support the characterization of DCNN codes as being functionally view invariant but with a view-specific code. Notably, earlier layers in the network showed view specificity, whereas higher layers showed view invariance.

It is worth digressing briefly to consider invariance in the context of neural approaches to face processing. An underlying assumption of neural approaches is that "a major purpose of the face patches is thus to construct a representation of individual identity invariant to view direction" (Hesse & Tsao 2020, pp. 703). Ideas about how this is accomplished have evolved. Freiwald & Tsao (2010) posited the progressive computation of invariance via the pooling of neurons across face patches, as follows. In early patches, a neuron responds to a specific identity from specific views; in middle face patches, greater invariance is achieved by pooling the responses of mirror-symmetric views of an identity; in later face patches, each neuron pools inputs representing all views of the same individual to create a fully view-invariant representation. More recently, Chang & Tsao (2017) proposed that the brain computes a view-invariant face code using shape and appearance parameters analogous to those used in a computer graphics model of face synthesis (Cootes et al. 1995) (see the sidebar titled Neurons, Neural Tuning, Population Codes, Features, and Perceptual Constancy). This code retains information about the face, but not about the particular image viewed.

Deep networks suggest an alternative that is largely consistent with neurophysiological data but interprets the data in a different light. Neurocomputational theory posits that the ventral visual system untangles face identity information from image parameters (DiCarlo &

Cox 2007). The idea is that visual processing starts in the image domain, where identity and viewpoint information are entangled. With successive levels of neural processing, manifolds corresponding to individual identities are untangled from image variation. This creates a representational space where identities can be separated with hyperplanes. Image information is not lost, but rather, is rearranged (for object recognition results, see Hong et al. 2016). The retention of image and identity information in DCNN face representations is consistent with this theory. It is also consistent with basic neuroscience findings indicating the emergence of a representation dominated by identity that retains sensitivity to image features (See Section 2.2).

**2.1.2. Appearance and demographics.—**Faces can be described using what computer vision researchers have called attributes or soft biometrics (hairstyle, hair color, facial hair, and accessories such as makeup and glasses). The definition of attributes in the computational literature is vague and can include demographics (e.g., gender, age, race) and even facial expression. Identity codes from deep networks retain a wide variety of face attributes. For example, Terhörst et al. (2020) built a massive attribute classifier (MAC) to test whether 113 attributes could be predicted from the face representations produced by deep networks [ArcFace (Deng et al. 2019) or FaceNet (Schroff et al. 2015)] for images from in-the-wild data sets (Huang et al. 2008, Liu et al. 2015). The MAC learned to map from DCNN-generated face representations to attribute labels. Cross-validated results showed that 39 of the attributes were easily predictable, and 74 of the 113 were predictable at reliable levels. Hairstyle, hair color, beard, and accessories were predicted easily. Attributes such as face geometry (e.g., round), periocular characteristics (e.g., arched eyebrows), and nose were moderately predictable. Skin and mouth attributes were not well predicted.

The continuous shuffling of identity, attribute, and image information across layers of the network was demonstrated by Dhar et al. (2020). They tracked the expressivity of attributes (identity, sex, age, pose) across layers of a deep network. Expressivity was defined as the degree to which a feature vector, from any given layer of a network, specified an attribute. Dhar et al. (2020) computed expressivity using a second neural network that estimated the mutual information between attributes and DCNN features. Expressivity order in the final fully connected layer of both networks (Resnet-101 and Inception Resnet v2; Ranjan et al. 2019) indicated that identity was most expressed, followed by age, sex, and yaw. Identity expressivity increased dramatically from the final pooling layer to the last fully connected layer. This echos the progressive increase in the detectability of view-invariant face identity representations seen across face patches in the macaque (Freiwald & Tsao 2010). It also raises the computational possibility of undetected viewpoint sensitivity in these neurons (see Section 3.1).

**Mutual information:**

a statistical term from information theory that quantifies the codependence of information between two random variables

#### 2.1.3. Social traits.—

People make consistent (albeit invalid) inferences about a person's social traits based on their face (Todorov 2017). These judgments have profound consequences. For example, competence judgments about faces predict election success at levels far above chance (Todorov et al. 2005). The physical structure of the face supports these trait inferences (Oosterhof & Todorov 2008, Walker & Vetter 2009), and thus it is not surprising that deep networks retain this information. Using face representations produced by a network trained for face identification (Sankaranarayanan et al. 2016), 11 traits (e.g., shy, warm, impulsive, artistic, lazy), rated by human participants, were predicted at levels well above chance (Parde et al. 2019). Song et al. (2017) found that more than half of 40 attributes were predicted accurately by a network trained for object recognition (VGG-16; Simonyan & Zisserman 2014). Human and machine trait ratings were highly correlated.

Other studies show that deep networks can be optimized to predict traits from images. Lewenberg et al. (2016) crowd-sourced large numbers of objective (e.g., hair color) and subjective (e.g., attractiveness) attribute ratings from faces. DCNNs were trained to classify images for the presence or absence of each attribute. They found highly accurate classification for the objective attributes and somewhat less accurate classification for the subjective attributes. McCurrie et al. (2017) trained a DCNN to classify faces according to trustworthiness, dominance, and IQ. They found significant accord with human ratings, with higher agreement for trustworthiness and dominance than for IQ.

#### 2.1.4. Facial expressions.—

Facial expressions are also detectable in face representations produced by identity-trained deep networks. Colón et al. (2021) found that expression classification was well above chance for face representations of images from the Karolinska data set (Lundqvist et al. 1998), which includes seven facial expressions (happy, sad, angry, surprised, fearful, disgusted, neutral) seen from five viewpoints (frontal and 90- and 45-degree left and right profiles). Consistent with human data, happiness was classified most accurately, followed by surprise, disgust, anger, neutral, sadness, and fear. Notably, accuracy did not vary across viewpoint. Visualization of the identities in the emergent face space showed a structured ordering of similarity in which viewpoint dominated over expression.

### 2.2. Functional Invariance, Useful Variability

The emergent code from identity-trained DCNNs can be used to recognize faces robustly, but it also retains extraneous information that is of limited, or no, value for identification. Although demographic and trait information offers weak hints to identity, image characteristics and facial expression are not useful for identification. Attributes such as glasses, hairstyle, and facial hair are, at best, weak identity cues and, at worst, misleading cues that will not remain constant over extended time periods. In purely computational terms, the variability of face representations for different images of an identity can lead to errors. Although this is problematic in security applications, coincidental features and attributes can be diagnostic enough to support acceptably accurate identification performance in day-to-day face recognition (Yovel & O'Toole 2016). (For related arguments based on adversarial images for object recognition, see Ilyas et al. 2019, Xie et al. 2020, Yuan et al. 2020.) A less-than-perfect identification system in computational terms, however, can be a surprisingly efficient, multipurpose face processing system that supports identification and the detection of visually derived semantic information [called attributes by Bruce & Young (1986)].

What do we learn from these studies that can be useful in understanding human visual processing of faces? First, we learn that it is computationally feasible to accommodate diverse information about faces (identity, demographics, visually derived semantic information), images (viewpoint, illumination, quality), and emotions (expression) in a unified representation. Furthermore, this diverse information can be accessed selectively from the representation. Thus, identity, image parameters, and attributes are all untangled when learning prioritizes the difficult within-category discrimination problem of face identification.

Second, we learn that to understand high-level visual representations for faces, we need to think in terms of categorical codes unbound from a spatial frame of reference. Although remnants of retinotopy and image characteristics remain in high-level visual areas (e.g., Grill-Spector et al. 1999, Kay et al. 2015, Kietzmann et al. 2012, Natu et al. 2010, Yue et al. 2010), the expressivity of spatial layout weakens dramatically from early visual areas to categorically structured areas in the IT cortex. Categorical face representations should capture what cognitive and perceptual psychologists call facial features (e.g., face shape, eye color). Indeed, altering these types of features in a face affects identity perception similarly for humans and deep networks (Abudarham et al. 2019). However, neurocomputational theory suggests that finding these features in the neural code will likely require rethinking the interpretation of neural tuning and population coding (see Section 3.2).

Third, if the ventral stream untangles information across layers of computations, then we should expect traces of identity, image data, and attributes at many, if not all, neural network layers. These may variously dominate the strength of the neural signal at different layers (see Section 3.1). Thus, various layers in the network will likely succeed in predicting several types of information about the face and/or image, though with differing accuracy. For now, we should not ascribe too much importance to findings about which specific layer(s) of a particular network predict specific attributes. Instead, we should pay attention to the pattern of prediction accuracy across layers. We would expect the following pattern. Clearly, for the optimized attribute (identity), the output offers the clearest access. For subject-related attributes (e.g., demographics), this may also be the case. For image-related attributes, we would expect every layer in the network to retain some degree of prediction ability. Exactly how, where, and whether the neural system makes use of these attributes for specific tasks remain open questions.

## 3. RETHINKING VISUAL FEATURES: IMPLICATIONS FOR NEURAL CODES

Deep learning models force us to rethink the definition and interpretation of facial features in high-level representations. Theoretical ideas about the brain's solution to complex real-world tasks such as face recognition must be reconciled at the level of neural units and representational spaces. Deep learning models can be used to test hypotheses about how faces are stored in the high-dimensional representational space defined by the pattern of responses of large numbers of neurons.

### 3.1. Units Confound Information that Separates in the Representation Space

Insight into interpreting facial features comes from deep network simulations aimed at understanding the relationship between unit responses and the information retained in the face representation. Parde et al. (2021) compared identification, gender classification, and viewpoint estimation in subspaces of a DCNN face space. Using an identity-trained network capable of all three tasks, they tested performance on the tasks using randomly sampled subsets of output units. Beginning at full dimensionality (512-units) and progressively decreasing sample size, they found no notable decline in identification accuracy for more than 3,000 in-the-wild-faces until the sample size reached 16 randomly chosen units (3% of full dimensionality). Correlations between unit responses across representations were near zero, indicating that individual units captured nonredundant identity cues. Statistical power for identification (i.e., separating identities) was uniformly high for all output units, demonstrating that units used their entire response range to separate identities. A unit firing at its maximum provided no more, and no less, information than any other response value. This distinction may seem trivial, but it is not. The data suggest that every output unit acts to separate identities to the maximum degree possible. As such, all units participate in coding all identities. In information theory terms, this is an ideal use of neural resources.

For gender classification and viewpoint estimation, performance declined at a much faster rate than for identification as units were deleted (Parde et al. 2021). Statistical power for predicting gender and viewpoint was strong in the distributed code but weak at the level of the unit. Prediction power for these attributes was again roughly equivalent for all units. Thus, individual units contributed to coding all three attributes, but identity modulated individual unit responses far more strongly than did gender or viewpoint. Notably, a principal component (PC) analysis of representations in the full-dimensional space revealed subspaces aligned with identity, gender, and viewpoint (Figure 3). Consistent with the strength of the categorical identity code in the representation, identity information dominated PCs explaining large amounts of variance, gender dominated the middle range of PCs, and viewpoint dominated PCs explaining small amounts of variation.

The emergence and effectiveness of these codes in DCNNs suggest that caution is needed in ascribing significance only to stimuli that drive a neuron to high rates of response. Small-scale modulations of neural responses can also be meaningful. Let us consider a concrete example. A neurophysiologist probing the network used by Parde et al. (2021) would find some neurons that respond strongly to a few identities. Interpreting this as identity tuning, however, would be an incorrect characterization of a code in which all units participate in coding all identities. Concomitantly, few units in the network would appear responsive to viewpoint or gender variations because unit firing rates would modulate only slightly with changes in viewpoint or gender. Thus, the distributed coding of view and gender across units would likely be missed. The finding that neurons in macaque face patch AM respond selectively (i.e., with high response rates) to identity over variable views (Freiwald & Tsao 2010) is consistent with DCNN face representations. It is possible, however, that these units also encode other face and image attributes, but with differential degrees of expressivity. This would be computationally consistent with the untangling theory and with DCNN codes.

**Macaque face patches:**

regions of the macaque cortex that respond selectively to faces, including the posterior lateral (PL), middle lateral (ML), middle fundus (MF), anterior lateral (AL), anterior fundus (AF), and anterior medial (AM)

Another example comes from the use of generative adversarial networks and related techniques to characterize the response properties of single (or multiple) neuron(s) in the primate visual cortex (Bashivan et al. 2019, Ponce et al. 2019, Yuan et al. 2020). These techniques have examined neurons in areas V4 (Bashivan et al. 2019) and IT (Ponce et al. 2019, Yuan et al. 2020). The goal is to progressively evolve images that drive neurons to their maximum response or that selectively (in)activate subsets of neurons. Evolved images show complex mosaics of textures, shapes, and colors. They sometimes show animals or people and sometimes reveal spatial patterns that are not semantically interpretable. However, these techniques rely on two strong assumptions. First, they assume that a neuron's response can be characterized completely in terms of the stimuli that activate it maximally, thereby discounting other response rates as noninformative. The computational utility of a unit's full response range in DCNNs suggests that reconsideration of this assumption is necessary. Second, these techniques assume that a neuron's response properties can be visualized accurately as a two-dimensional image. Given the categorical, nonretinotopic nature of representations in high-level visual areas, this seems problematic. If the representation under consideration is not in the image or pixel domain, then image-based visualization may offer limited, and possibly misleading, insight into the underlying nature of the code.

### 3.2. Direct-Fit Models and Deep Learning

In rethinking visual features at a theoretical level, direct-fit models of neural coding appear to best explain deep learning findings in multiple domains (e.g., face recognition, language) (Hasson et al. 2020). These models posit that neural computation fits densely sampled data from the environment. Implementation is accomplished using "overparameterized optimization algorithms that increase predictive (generalization) power, without explicitly modeling the underlying generative structure of the world" (Hasson et al. 2020, p. 418). Hasson et al. (2020) begins with an ideal model in a small-parameter space (Figure 4). When the underlying structure of the world is simple, a small-parameter model will find the underlying generative function, thereby supporting generalization via interpolation and extrapolation. Despite decades of effort, small-parameter functions have not solved real-world face recognition with performance anywhere near that of humans.

When the underlying structure of the world is complex and multivariate, direct-fit models offer an alternative to models based on small-parameter functions. With densely sampled real-world training data, each new observation can be placed in the context of past experience. More formally, direct-fit models solve the problem of generalization to new exemplars by experience-scaffolded interpolation (Hasson et al. 2020). This produces face recognition performance in the range of that of humans. A fundamental element of the success of deep networks is that they model the environment with big data, which can be structured in overparameterized spaces. The scale of the parameterization and the requirement to operate on real-world data are pivotal. Once the network is sufficiently parameterized to fit the data, the exact details of its architecture are not important. This may explain why starkly different network architectures arrive at similarly structured representations (Hill et al. 2019, Parde et al. 2017, Storrs et al. 2020).

Returning to the issue of features, in neurocomputational terms, the strength of connectivity between neurons at synapses is the primary locus of information, just as weights between units in a deep network comprise information. We expect features, whatever they are, to be housed in the combination of connection strengths among units, not in the units themselves. In a high-dimensional multivariate encoding space, they are hyperplane directions through the space. Thus, features are represented across many computing elements, and each computing element participates in encoding many features (Hasson et al. 2020, Parde et al. 2021). If features are directions in a high-dimensional coding space (Goodfellow et al. 2014), then units act as an arbitrary projection surface from which this information can be accessed—albeit in a nontransparent form.

A downside of direct-fit models is that they cannot generalize via extrapolation. The other-race effect is an example of how face recognition may fail due to limited experience (Malpass & Kravitz 1969) (see Section 4.3.2). The extrapolation limit may be countered, however, by the capacity of direct-fit models to acquire expertise within the confines of experience. For example, in human perception, category experience selectively structures representations as new exemplars are learned. Collins & Behrmann (2020) show that this occurs in a way that reflects the greater experience that humans have with faces and computer-generated objects from novel made-up categories of objects, which the authors call YUFOs. They tracked the perceived similarity of pairs of other-race faces and YUFOs as people learned novel exemplars of each. Experience changed perceived similarities more selectively for faces than for YUFOs, enabling more nuanced discrimination of exemplars from the experienced category of faces.

In summary, direct-fit models offer a framework for thinking about high-level visual codes for faces in a way that unifies disparate data on single units and high-dimensional coding spaces. These models are fueled by the rich experience that we (models) gain from learning (training on) real-world data. They solve complex visual tasks with interpolated solutions that elude transparent semantic interpretation.

## 4. RETHINKING LEARNING IN HUMANS AND DEEP NETWORKS

Deep network models of human face processing force us to consider learning as a complex and diverse set of mechanisms that can overlap, accumulate over time, and interact. Learning in both humans and artificial neural networks can refer to qualitatively different phenomena. In both cases, learning involves multiple steps. For DCNNs, these steps are fundamental to a network’s ability to recognize faces across image and appearance variation. Human visual learning is likewise diverse and unfolds across the developmental lifespan in a process governed by genetics and environmental input (Goodman & Shatz 1993). The stepwise implementation of learning is one way that DCNNs differ from previous face recognition networks. Considered as manipulable modeling tools, the learning steps in DCNNs force us to think in concrete and nuanced ways about how humans learn faces.

In this section, we outline the learning layers in human face processing (Section 4.1), introduce the layers of learning used in training machines (Section 4.2), and consider the relationship between the two in the context of human behavior (Section 4.3.1). The human learning layers support a complex, biologically realized face processing system. The machine learning layers can be thought of as building blocks that can be combined in a variety of ways to model human behavioral phenomena. At the outset, we note that machine learning is designed to maximize performance—not to model the development of the human face processing system (Smith & Slone 2017). Concomitantly, the sequential presentation of training data in DCNNs differs from the pattern of exposure that infants and young children have with faces and objects (Jayaraman et al. 2015). The machine learning steps, however, can be modified to model human learning more closely. In practical terms, fully trained DCNNs, available on the web, are used (almost exclusively) to model human neural systems (see the sidebar titled Caveat: Iteration Between Theory and Practice). It is important, therefore, to understand how (and why) these models are configured as they are and to understand the types of learning tools available for modeling human face processing. These steps may provide computational grounding for basic learning mechanisms hypothesized in humans.

### 4.1. Human Learning for Face Processing

To model human face processing, researchers need to consider the following types of learning. The most specific form of learning is familiar face recognition. People learn the faces of specific familiar individuals (e.g., friends, family, celebrities). Familiar faces are recognized robustly over challenging changes in appearance and image characteristics. The second-most specific is local population tuning. People recognize own-race faces more accurately than other-race faces, a phenomenon referred to as the other-race effect (e.g., Malpass & Kravitz 1969). This likely results from tuning to the statistical properties of the faces that we see most frequently—typically faces of our own race. The third-most specific is nfamiliar face recognition. People can differentiate unfamiliar faces perceptually. Unfamiliar refers to faces that a person has not encountered previously or has encountered infrequently. Unfamiliar face recognition is less robust to image and appearance change than is familiar face recognition. The least specific form of learning is object recognition. At a fundamental level of analysis, faces are objects, and both share early visual processing wetware.

### 4.2. How Deep Convolutional Neural Networks Learn Face Identification

Training DCNNs for face recognition involves a sequence of learning stages, each with a concrete objective. Unlike human learning, machine learning stages are executed in strict sequence. The goal across all stages of training is to build an effective method for converting images of faces into points in a high-dimensional space. The resulting high-dimensional space allows for easy comparison among faces, search, and clustering. In this section, we sketch out the engineering approach to learning, working forward from the most general to the most specific form of learning. This follows the implementation order used by engineers.

**4.2.1. Object classification (between-category learning): Stage 1.—**Deep networks for face identification are commonly built on top of DCNNs that have been pretrained for object classification. Pretraining is carried out using large data sets of objects, such as those available in ImageNet (Russakovsky et al. 2015), which contains more than 14 million images of over 1,000 classes of objects (e.g., volcanoes, cups, chihuahuas). The object categorization training procedure involves adjusting the weights on all layers of the network. For training to converge, a large training set is required. The loss function optimized in this procedure typically uses the well-understood cross-entropy loss + Softmax combination. Most practitioners do not execute this step because it has been performed already in a pretrained model downloaded from a public repository in a format compatible with DCNN software libraries [e.g., PyTorch (Paszke et al. 2019), TensorFlow (Abadi et al. 2016)]. Networks trained for object recognition have proven better for face identification than networks that start with a random configuration (Liu et al. 2015, Yi et al. 2014).

**4.2.2. Face recognition (within-category learning): Stage 2.—**Face recognition training is implemented in a second stage of training. In this stage, the last fully connected layer that connects to object-category nodes (e.g., volcanoes, cups) is removed from the results of the Stage 1 training. Next, a fully connected layer that maps to the number of face identities available for face training is connected. Depending on the size of the face training set, the weights of either all layers or all but a few layers at the beginning of the network are updated. The former is common when very large numbers of face identities are available for training. In academic laboratories, data sets include 5–10 million face images of 40,000–100,000 identities. In industry, far larger data sets are often used (Schroff et al. 2015). A technical difficulty encountered in retraining an object classification network to a face recognition network is the large increase in the number of categories involved (approximately 1,000 objects versus 50,000+ faces). Special loss functions can address this issue [e.g., L2-Softmax/crystal loss (Ranjan et al. 2017), NormFace (Wang et al. 2017), angular Softmax (Li et al. 2018), additive Softmax (Wang et al. 2018), additive angular margins (Deng et al. 2019)].

When the Stage 2 face training is complete, the last fully connected layer that connects to the 50,000+ face identity nodes is removed, leaving below it a relatively low-dimensional (128- to 5,000-unit) layer of output units. This can be thought of as the face representation. This output represents a face image, not a face identity. At this point in training, any arbitrary face image from any identity (known or unknown to the network) can be processed by the DCNN to produce a compact face image descriptor across the units of this layer. If the network functions perfectly, then it will produce identical codes for all images of the same person. This would amount to perfect image and appearance generalization. This is not usually achieved, even when the network is highly accurate (see Section 2).

In this state, the network is commonly employed to recognize faces not seen in training (unfamiliar faces). Stage 2 training supports a surprising degree of generalization (e.g., pose, expression, illumination, and appearance) for images of unfamiliar faces. This general face learning gives the system special knowledge of faces and enables it to perform within-category face discrimination for unfamiliar faces (O’Toole et al. 2018). With or without Stage 3 training, the network is now capable of converting images of faces into points in a high-dimensional space, which, as noted above, is the primary goal of training. In practice, however, Stages 3 and 4 can provide a critical bridge to modeling behavioral characteristics of the human face processing system.

**4.2.3. Adapting to local statistics of people and visual environments: Stage 3.—**The objective of Stage 3 training is to finalize the modification of the DCNN weights to better adapt to the application domain. The term application domain can refer to faces from a particular race or ethnicity or, as it is commonly used in industry, to the type of images to be processed (e.g., in-the-wild faces, passport photographs). This training is a crucial step in many applications because there will be no further transformation of the weights. Special care is needed in this training to avoid collapsing the representation into a form that is too specific. Training at this stage can improve performance for some faces and decrease it for others.

Whereas Stages 1 and 2 are used in the vast majority of published computational work, in Stage 3, researchers diverge. Although there is no standard implementation for this training, fine-tuning and learning a triplet loss embedding (van der Maaten & Weinberger 2012) are common methods. These methods are conceptually similar but differ in implementation. In both methods, (*a*) new layers are added to the network, (*b*) specific subsets of layers are frozen or unfrozen, and (*c*) optimization continues with an appropriate loss function using a new data set with the desired domain characteristics. Fine-tuning starts from an already-viable network state and updates a nonempty subset of weights, or possibly all weights. It is typically implemented with smaller learning rates and can use smaller training sets than those needed for full training. Triplet loss is implemented by freezing all layers and adding a new, fully connected layer. Minimization is done with the triplet loss, again on a new (smaller) data set with the desired domain characteristics.

A natural question is why Stage 2 (general face training) is not considered fine-tuning. The answer, in practice, comes down to viability and volume. When the training for Stage 2 starts, the network is not in a viable state to perform face recognition. Therefore, it requires a voluminous, diverse data set to function. Stage 3 begins with a functional network and can be tuned effectively with a small targeted data set.

This face knowledge history provides a tool for adapting to local face statistics (e.g., race) (O’Toole et al. 2018).

**4.2.4. Learning individual people: Stage 4.—**In psychological terms, learning individual familiar faces involves seeing multiple, diverse images of the individuals to whom the faces belong. As we see more images of a person, we become more familiar with their face and can recognize it from increasingly variable images (Dowsett et al. 2016, Murphy et al. 2015, Ritchie & Burton 2017). In computational terms, this translates into the question of how a network can learn to recognize a random set of special (familiar) faces with greater accuracy and robustness than other nonspecial (unfamiliar) faces—assuming, of course, the availability of multiple, variable images of the special faces. This stage of learning is defined, in nearly all cases, outside of the DCNN, with no change to weights within the DCNN.

The problem is as follows. The network starts with multiple images of each familiar identity and can produce a representation for each of the images–but what then? There is no standard familiarization protocol, but several approaches exist. We categorize these approaches first and link them to theoretical accounts of face familiarity in Section 4.3.3.

The first approach is averaging identity codes, or 1-class learning. It is common in machine learning to use an average (or weighted average) of the DCNN-generated face image representations as an identity code (see also Crosswhite et al. 2018, Su et al. 2015). Averaging creates a person-identity prototype (Noyes et al. 2021) for each familiar face.

The second is individual face contrast, or 2-class learning. This technique employs direct learning of individual identities by contrasting them with all other identities. There are two classes because the model learns what makes each identity (positive class) different than all other identities (negative class). The distinctiveness of each familiar face is enhanced relative to all other known faces (e.g., Noyes et al. 2021).

The third is multiple face contrast, or K-class learning. This refers to the use of identification training for a random set of (familiar) faces with a simple network (often a one-layer network). The network learns to map DCNN-generated face representations of the available images onto identity nodes.

The fourth approach is fine-tuning individual face representations. Fine-tuning has also been used for learning familiar identities (Blauch et al. 2020a). It is an unusual method because it alters weights within the DCNN itself. This can improve performance for the familiarized faces but can limit the network's ability to represent other faces.

These methods create a personal face learning history that supports more accurate and robust face processing for familiar people (O'Toole et al. 2018).

## 4.3. Mapping Learning Between Humans and Machines

Deep networks rely on multiple types of learning that can be useful in formulating and testing complex, nuanced hypotheses about human face learning. Manipulable variables include order of learning, training data, and network plasticity at different learning stages. We consider a sample of topics in human face processing that can be investigated by manipulating learning in deep networks. Because these investigations are just beginning, we provide an overview of the work in progress and discuss possible next steps in modeling.

**4.3.1. Development of face processing.—**Early infants' experience with faces is critical for the development of face processing skills (Maurer et al. 2002). The timing of this experience has become increasingly clear with the availability of data sets gathered using head-mounted cameras in infants (1–15 months of age) (e.g., Jayaraman et al. 2015, Yoshida & Smith 2008). In seeing the world from the perspective of the infant, it becomes clear that the development of sensorimotor abilities drives visual experience. Infants' experience transitions from seeing only what is made available to them (often faces in the near range), to seeing the world from the perspective of a crawler (objects and environments), to seeing hands and the objects that they manipulate (Fausey et al. 2016, Jayaraman et al. 2015, Smith & Slone 2017, Sugden & Moulson 2017). Between 1 and 3 months of age, faces are frequent, temporally persistent, and viewed frontally at close range. This early experience with faces is limited to a few individuals. Faces become less frequent as the child's first year progresses and attention shifts to the environment, to objects, and later to hands (Jayaraman & Smith 2019).

The prevalence of a few important faces in the infants' visual world suggests that early face learning may have an out-sized influence on structuring visual recognition systems. Infants' visual experience of objects, faces, and environments can provide a curriculum for teaching machines (Smith et al. 2018). DCNNs can be used to test hypotheses about the emergence of competence on different face processing tasks. Some basic computational challenges, however, need to be addressed. Training with very large numbers of objects (or faces) is required for deep network learning to converge (see Section 4.2.1). Starting small and building competence on multiple domains (faces, objects, environments) might require basic changes to deep network training. Alternatively, the small number of special faces in an infant's life might be considered familiar faces. Perception and memory of these faces may be better modeled using tools that operate outside the deep network on representations that develop within the network (Stage 4 learning; Section 4.2.4). In this case, the quality of the representation produced at different points in a network's development of more general visual knowledge varies (Stages 1 and 2 of training; Sections 4.2.1 and 4.2.2). The learning of these special faces early in development might interact with the learning of objects and scenes at the categorical level (Rosch et al. 1976, Yovel et al. 2012). A promising approach would involve pausing training in Stages 1 and 2 to test face representation quality at various points along the way to convergence.

**4.3.2. Race bias in the performance of humans and deep networks.—**People recognize own-race faces more accurately than other-race faces. For humans, this other-race effect begins in infancy (Kelly et al. 2005, 2007) and is manifest in children (Pezdek et al. 2003). Although it is possible to reverse these effects in childhood (Sangrigoli et al. 2005), training adults to recognize other-race faces yields only modest gains (e.g., Cavazos et al. 2019, Hayward et al. 2017, Laurence et al. 2016, Matthews & Mondloch 2018, Tanaka & Pierce 2009). Concomitantly, evidence for the experience-based contact hypothesis is weak when it is evaluated in adulthood (Levin 2000). Clearly, the timing of experience is critical in the other-race effect. Developmental learning, which results in perceptual narrowing during a critical childhood period, may provide a partial account of the other-race effect (Kelly et al. 2007, Sangrigoli et al. 2005, Scott & Monesson 2010).

**Perceptual narrowing:**

sculpting of neural and perceptual processing via experience during a critical period in child development

Face recognition algorithms from the 1990s and present-day DCNNs differ in accuracy for faces of different races (for a review, see Cavazos et al. 2020; for a comprehensive test of race bias in DCNNs, see Grother et al. 2019). Although training with faces of different races is often cited as a cause of race effects, it is unclear which training stage(s) contribute to the bias. It is likely that biased learning affects all learning stages. From the human perspective, for many people, experience favors own-race faces across the lifespan, potentially impacting performance through multiple learning mechanisms (developmental, unfamiliar, and familiar face learning). DCNN training may also use race-biased data at all stages. For humans, understanding the role of different types of learning in the other-race effect is challenging because experience with faces cannot be controlled. DCNNs can serve as a tool for studying critical periods and perceptual narrowing. It is possible to compare the face representations that emerge from training regimes that vary in the time course of exposure to faces of different races. The ability to manipulate training stage order, network plasticity, and training set diversity in deep networks offers an opportunity to test hypotheses about how bias emerges. The major challenge for DCNNs is the limited availability of face databases that represent the diversity of humans.

**4.3.3. Familiar versus unfamiliar face recognition.—**Face familiarity in a deep network can be modeled in more ways than we can count. The approaches presented in Section 4.2.4 are just a beginning. Researchers should focus first on the big questions. How do familiar and unfamiliar face representations differ—beyond simple accuracy and robustness? This has been much debated recently, and many questions remain (Blauch et al. 2020a,b; Young & Burton 2020; Yovel & Abudarham 2020). One approach is to ask where in the learning process representations for familiar and unfamiliar faces diverge. The methods outlined in Section 4.2.4 make some predictions.

In the individual and multiple face contrast methods, familiar and unfamiliar face representations are not differentiated within the deep network. Instead, familiar face representations generated by the DCNN are enhanced in another, simpler network populated with known faces. A familiar face’s representation is affected, therefore, by the other faces that we know well. Contrast techniques have preliminary empirical support. In the work of Noyes et al. (2021), familiarization using individual-face contrast improved identification for both evasion and impersonation disguise. It also produced a pattern of accuracy similar to that seen for people familiar with the disguised individuals (Noyes & Jenkins 2019). For humans who were unfamiliar with the disguised faces, the pattern of accuracy resembled that seen after general face training inside of the DCNN. There is also support for multiple-face contrast familiarization. Perceptual expertise findings that emphasize the selective effects of the exemplars experienced during highly skilled learning are consistent with this approach (Collins & Behrmann 2020) (see Section 3.2).

Familiarization by averaging and fine-tuning both improve performance, but at a cost. For example, averaging the DCNN representations increased performance for evasion disguise by increasing tolerance for appearance variation (Noyes et al. 2021). It decreased performance, however, for imposter disguise by allowing too much tolerance for appearance variation. Averaging methods highlight the need to balance the perception of identity across variable images with an ability to tell similar faces apart.

Familiarization via fine-tuning was explored by Blauch et al. (2020a), who varied the number of layers tuned (all layers, fully connected layers, only the fully connected layer mapping the perceptual layer to identity nodes). Fine-tuning applied at lower layers alters the weights within the deep network to produce a perceptual representation potentially affected by familiar faces. Fine-tuning in the mapping layer is equivalent to multiclass face contrast learning (Blauch et al. 2020b). Blauch et al. (2020b) show that fine-tuning the perceptual representation, which they consider analogous to perceptual learning, is not necessary for producing a familiarity effect (Blauch et al. 2020a).

These approaches are not (necessarily) mutually exclusive and therefore can be combined to exploit useful features of each.

**4.3.4. Objects, faces, both.—**The organization of face-, body-, and object-selective areas in the ventral temporal cortex has been studied intensively (cf. Grill-Spector & Weiner 2014). Neuroimaging studies in childhood reveal the developmental time course of face selectivity and other high-level visual tasks (e.g., Natu et al. 2016; Nordt et al. 2019, 2020). How these systems interact during development in the context of constantly changing input from the environment is an open question. DCNNs can be used to test functional hypotheses about the development of object and face learning (see also Grill-Spector et al. 2018).

In the case of machine learning, face recognition networks are more accurate when pretrained to categorize objects (Liu et al. 2015, Yi et al. 2014), and networks trained with only faces are more accurate for face recognition than networks trained with only objects (Abudarham & Yovel 2020, Blauch et al. 2020a). Human-like viewpoint invariance was found in a DCNN trained for face recognition but not in one trained for object recognition (Abudarham & Yovel 2020). In machine learning, networks are trained first with objects, and then with faces. Moreover, networks can simultaneously learn object and face recognition (Dobs et al. 2020), which incurs minimal duplication of neural resources.

### 4.4. New Tools, New Questions, New Data, and a New Look at Old Data

Psychologists have long posited diverse and complex learning mechanisms for faces. Deep networks provide new tools that can be used to model human face learning with greater precision than was possible previously. This is useful because it encourages theoreticians to articulate hypotheses in ways specific enough to model. It may no longer be sufficient to explain a phenomenon in terms of generic learning or contact. Concepts such as perceptual narrowing should include ideas about where and how in the learning process this narrowing occurs. A major challenge ahead is the sheer number of knobs to be set in deep networks. Plasticity, for example, can be dialed up or down, and it can be applied to selected network layers or specific face diets administered across multiple learning stages (in sequence or simultaneously). The list goes on. In all of the topics discussed, and others not discussed, theoretical ideas should specify the manipulations thought to be most critical. We should follow the counsel of Box (1976) to avoid worrying selectivity and instead focus on what is most important. New tools succeed when they facilitate the discovery of things that we did not know or had not hypothesized. Testing these hypotheses will require new data and may suggest a reevaluation of existing data.

## 5. THE PATH FORWARD

In this review, we highlight fundamental advances in thinking brought about by deep learning approaches. These networks solve the inverse optics problem for face identification by untangling image, appearance, and identity over layers of neural-like processing. This demonstrates that robust face identification can be achieved with a representation that includes specific information about the face image(s) actually experienced. These representations retain information about appearance, perceived traits, expressions, and identity.

Direct-fit models posit that deep networks operate by placing new observations into the context of past experience. These models depend on overparameterized networks that create a high-dimensional space from real-world training data. Face representations housed within this space project onto units, thereby confounding stimulus features that (may) separate in the high-dimensional space. This raises questions about the transparency and interpretability of information gained by examining the response properties of network units. Deep networks can be studied at the both micro- and macroscale simultaneously and can be used to formulate hypotheses about the underlying neural code for faces. A key to understanding face representations is to reconcile the responses of neurons to the structure of the code in the high-dimensional space. This is a challenging problem best approached by combining psychological, neural, and computational methods.

The process of training a deep network is complex and layered. It draws on learning mechanisms aimed at objects and faces, visual categories of faces (e.g., race), and special familiar faces. Psychological and neural theory considers the many ways in which people and brains learn faces from real-world visual experience. DCNNs offer the potential to implement and test sophisticated hypotheses about how humans learn faces across the lifespan.

We should not lose sight of the fact that a compelling reason to study deep networks is that they actually work, i.e., they perform nearly as well as humans, on face recognition tasks that have stymied computational modelers for decades. This might qualify as a property of deep networks that is importantly right (Box 1976). There is a difference, of course, between working and working like humans. Determining whether a deep network can work like humans, or could be made to do so by manipulating other properties of the network (e.g., architectures, training data, learning rules), is work that is just beginning.

## ACKNOWLEDGMENTS

The authors are supported by funding provided by National Eye Institute grant R01EY029692-03 to A.J.O. and C.D.C.

## LITERATURE CITED

Abadi M, Barham P, Chen J, Chen Z, Davis A, et al. 2016. Tensorflow: a system for large-scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), pp. 265–83. Berkeley, CA: USENIX

Abudarham N, Shkiller L, Yovel G. 2019. Critical features for face recognition. Cognition 182:73–83 [PubMed: 30218914]

Abudarham N, Yovel G. 2020. Face recognition depends on specialized mechanisms tuned to view-invariant facial features: insights from deep neural networks optimized for face or object recognition. bioRxiv 2020.01.01.890277. 10.1101/2020.01.01.890277

Azevedo FA, Carvalho LR, Grinberg LT, Farfel JM, Ferretti RE, et al. 2009. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. J. Comp. Neurol 513(5):532–41 [PubMed: 19226510]

Barlow HB. 1972. Single units and sensation: a neuron doctrine for perceptual psychology? Perception 1(4):371–94 [PubMed: 4377168]

Bashivan P, Kar K, DiCarlo JJ. 2019. Neural population control via deep image synthesis. Science 364(6439):eaav9436

Best-Rowden L, Jain AK. 2018. Learning face image quality from human assessments. IEEE Trans. Inform. Forensics Secur 13(12):3064–77

Blanz V, Vetter T. 1999. A morphable model for the synthesis of 3d faces. In Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, pp. 187–94. New York: ACM

Blauch NM, Behrmann M, Plaut DC. 2020a. Computational insights into human perceptual expertise for familiar and unfamiliar face recognition. Cognition 208:104341 [PubMed: 32586632]

Blauch NM, Behrmann M, Plaut DC. 2020b. Deep learning of shared perceptual representations for familiar and unfamiliar faces: reply to commentaries. Cognition 208:104484 [PubMed: 33504433]

Box GE. 1976. Science and statistics. J. Am. Stat. Assoc 71(356):791–99

Box GEP. 1979. Robustness in the strategy of scientific model building. In Robustness in Statistics, ed. Launer RL, Wilkinson GN, pp. 201–36. Cambridge, MA: Academic Press

Bruce V, Young A. 1986. Understanding face recognition. Br. J. Psychol 77(3):305–27 [PubMed: 3756376]

Burton AM, Bruce V, Hancock PJ. 1999. From pixels to people: a model of familiar face recognition. Cogn. Sci 23(1):1–31

Cavazos JG, Noyes E, O'Toole AJ. 2019. Learning context and the other-race effect: strategies for improving face recognition. Vis. Res 157:169–83 [PubMed: 29604301]

Cavazos JG, Phillips PJ, Castillo CD, O'Toole AJ. 2020. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? IEEE Trans. Biom. Behav. Identity Sci 3(1):101–11 [PubMed: 33585821]

Chang L, Tsao DY. 2017. The code for facial identity in the primate brain. Cell 169(6):1013–28 [PubMed: 28575666]

Chen JC, Patel VM, Chellappa R. 2016. Unconstrained face verification using deep CNN features. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–9. Piscataway, NJ: IEEE

Cichy RM, Kaiser D. 2019. Deep neural networks as scientific models. Trends Cogn. Sci 23(4):305–17 [PubMed: 30795896]

Collins E, Behrmann M. 2020. Exemplar learning reveals the representational origins of expert category perception. PNAS 117(20):11167–77 [PubMed: 32366664]

Colón YI, Castillo CD, O'Toole AJ. 2021. Facial expression is retained in deep networks trained for face identification. J. Vis 21(4):4

Cootes TF, Taylor CJ, Cooper DH, Graham J. 1995. Active shape models-their training and application. Comput. Vis. Image Underst 61(1):38–59

Crosswhite N, Byrne J, Stauffer C, Parkhi O, Cao Q, Zisserman A. 2018. Template adaptation for face verification and identification. Image Vis. Comput 79:35–48

Deng J, Guo J, Xue N, Zafeiriou S. 2019. Arcface: additive angular margin loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4690–99. Piscataway, NJ: IEEE

Dhar P, Bansal A, Castillo CD, Gleason J, Phillips P, Chellappa R. 2020. How are attributes expressed in face DCNNs? In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020), pp. 61–68. Piscataway, NJ: IEEE

DiCarlo JJ, Cox DD. 2007. Untangling invariant object recognition. Trends Cogn. Sci 11(8):333–41 [PubMed: 17631409]

Dobs K, Kell AJ, Martinez J, Cohen M, Kanwisher N. 2020. Using task-optimized neural networks to understand why brains have specialized processing for faces. J. Vis 20(11):660

Dowsett A, Sandford A, Burton AM. 2016. Face learning with multiple images leads to fast acquisition of familiarity for specific individuals. Q. J. Exp. Psychol 69(1):1–10

El Khiyari H, Wechsler H. 2016. Face verification subject to varying (age, ethnicity, and gender) demographics using deep learning. J. Biom. Biostat 7:323

Fausey CM, Jayaraman S, Smith LB. 2016. From faces to hands: changing visual input in the first two years. Cognition 152:101–7 [PubMed: 27043744]

Freiwald WA, Tsao DY. 2010. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. Science 330(6005):845–51 [PubMed: 21051642]

Freiwald WA, Tsao DY, Livingstone MS. 2009. A face feature space in the macaque temporal lobe. Nat. Neurosci 12(9):1187–96 [PubMed: 19668199]

Fukushima K 1988. Neocognitron: a hierarchical neural network capable of visual pattern recognition. Neural Netw 1(2):119–30

Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, et al. 2014. Generative adversarial nets. In NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems, pp. 2672–80. New York: ACM

Goodman CS, Shatz CJ. 1993. Developmental mechanisms that generate precise patterns of neuronal connectivity. Cell 72:77–98 [PubMed: 8428376]

Grill-Spector K, Kushnir T, Edelman S, Avidan G, Itzchak Y, Malach R. 1999. Differential processing of objects under various viewing conditions in the human lateral occipital complex. Neuron 24(1):187–203 [PubMed: 10677037]

Grill-Spector K, Weiner KS. 2014. The functional architecture of the ventral temporal cortex and its role in categorization. Nat. Rev. Neurosci 15(8):536–48 [PubMed: 24962370]

Grill-Spector K, Weiner KS, Gomez J, Stigliani A, Natu VS. 2018. The functional neuroanatomy of face perception: from brain measurements to deep neural networks. Interface Focus 8(4):20180013 [PubMed: 29951193]

Gross CG. 2002. Genealogy of the "grandmother cell". Neuroscientist 8(5):512–18 [PubMed: 12374433]

Grother P, Ngan M, Hanaoka K. 2019. Face recognition vendor test (FRVT) part 3: demographic effects. Rep., Natl. Inst. Stand. Technol., US Dept. Commerce, Gaithersburg, MD

Hancock PJ, Bruce V, Burton AM. 2000. Recognition of unfamiliar faces. Trends Cogn. Sci 4(9):330–37 [PubMed: 10962614]

Hasson U, Nastase SA, Goldstein A. 2020. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. Neuron 105(3):416–34 [PubMed: 32027833]

Hayward WG, Favelle SK, Oxner M, Chu MH, Lam SM. 2017. The other-race effect in face learning: using naturalistic images to investigate face ethnicity effects in a learning paradigm. Q. J. Exp. Psychol 70(5):890–96

Hesse JK, Tsao DY. 2020. The macaque face patch system: a turtle's underbelly for the brain. Nat. Rev. Neurosci 21(12):695–716 [PubMed: 33144718]

Hill MQ, Parde CJ, Castillo CD, Colon YI, Ranjan R, et al. 2019. Deep convolutional neural networks in the face of caricature. Nat. Mach. Intel 1(11):522–29

Hong H, Yamins DL, Majaj NJ, DiCarlo JJ. 2016. Explicit information for category-orthogonal object properties increases along the ventral stream. Nat. Neurosci 19(4):613–22 [PubMed: 26900926]

Hornik K, Stinchcombe M, White H. 1989. Multilayer feedforward networks are universal approximators. Neural Netw 2(5):359–66

Huang GB, Lee H, Learned-Miller E. 2012. Learning hierarchical representations for face verification with convolutional deep belief networks. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2518–25. Piscataway, NJ: IEEE

Huang GB, Mattar M, Berg T, Learned-Miller E. 2008. Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Paper presented at the Workshop on Faces in "Real-Life" Images: Detection, Alignment, and Recognition, Marseille, France

Ilyas A, Santurkar S, Tsipras D, Engstrom L, Tran B, Madry A. 2019. Adversarial examples are not bugs, they are features. arXiv:1905.02175 [stat.ML]

Issa EB, DiCarlo JJ. 2012. Precedence of the eye region in neural processing of faces. J. Neurosci 32(47):16666–82 [PubMed: 23175821]

Jacquet M, Champod C. 2020. Automated face recognition in forensic science: review and perspectives. Forensic Sci. Int 307:110124 [PubMed: 31927397]

Jayaraman S, Fausey CM, Smith LB. 2015. The faces in infant-perspective scenes change over the first year of life. PLOS ONE 10(5):e0123780 [PubMed: 26016988]

Jayaraman S, Smith LB. 2019. Faces in early visual environments are persistent not just frequent. Vis. Res 157:213–21 [PubMed: 29852210]

Jenkins R, White D, Van Montfort X, Burton AM. 2011. Variability in photos of the same face. Cognition 121(3):313–23 [PubMed: 21890124]

Kandel ER, Schwartz JH, Jessell TM, Siegelbaum S, Hudspeth AJ, Mack S, eds. 2000. Principles of Neural Science, Vol. 4. New York: McGraw-Hill

Kay KN, Weiner KS, Grill-Spector K. 2015. Attention reduces spatial uncertainty in human ventral temporal cortex. Curr. Biol 25(5):595–600 [PubMed: 25702580]

Kelly DJ, Quinn PC, Slater AM, Lee K, Ge L, Pascalis O. 2007. The other-race effect develops during infancy: evidence of perceptual narrowing. Psychol. Sci 18(12):1084–89 [PubMed: 18031416]

Kelly DJ, Quinn PC, Slater AM, Lee K, Gibson A, et al. 2005. Three-month-olds, but not newborns, prefer own-race faces. Dev. Sci 8(6):F31–36 [PubMed: 16246233]

Kietzmann TC, Swisher JD, König P, Tong F. 2012. Prevalence of selectivity for mirror-symmetric views of faces in the ventral and dorsal visual pathways. J. Neurosci 32(34):11763–72 [PubMed: 22915118]

Krishnapriya KS, Albiero V, Vangara K, King MC, Bowyer KW. 2020. Issues related to face recognition accuracy varying based on race and skin tone. IEEE Trans. Technol. Soc 1(1):8–20

Krishnapriya K, Vangara K, King MC, Albiero V, Bowyer K. 2019. Characterizing the variability in face recognition accuracy relative to race. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Vol. 1, pp. 2278–85. Piscataway, NJ: IEEE

Krizhevsky A, Sutskever I, Hinton GE. 2012. Imagenet classification with deep convolutional neural networks. In NIPS'12: Proceedings of the 25th International Conference on Neural Information Processing Systems, pp. 1097–105. New York: ACM

Kumar N, Berg AC, Belhumeur PN, Nayar SK. 2009. Attribute and simile classifiers for face verification. In Proceedings of the 2009 IEEE International Conference on Computer Vision, pp. 365–72. Piscataway, NJ: IEEE

Laurence S, Zhou X, Mondloch CJ. 2016. The flip side of the other-race coin: They all look different to me. Br. J. Psychol 107(2):374–88 [PubMed: 26366460]

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. Nature 521(7553):436–44 [PubMed: 26017442]

Levin DT. 2000. Race as a visual feature: using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. J. Exp. Psychol. Gen 129(4):559–74 [PubMed: 11142869]

Lewenberg Y, Bachrach Y, Shankar S, Criminisi A. 2016. Predicting personal traits from facial images using convolutional neural networks augmented with facial landmark information. arXiv:1605.09062 [cs.CV]

Li Y, Gao F, Ou Z, Sun J. 2018. Angular softmax loss for end-to-end speaker verification. In Proceedings of the 11th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 190–94. Baixas, France: ISCA

Liu Z, Luo P, Wang X, Tang X. 2015. Deep learning face attributes in the wild. In Proceedings of the 2015 IEEE International Conference on Computer Vision, pp. 3730–38. Piscataway, NJ: IEEE

Lundqvist D, Flykt A, Ohman A. 1998. Karolinska directed emotional faces. Database of standardized facial images, Psychol. Sect., Dept. Clin. Neurosci. Karolinska Hosp., Solna, Swed. https://www.kdef.se/#:~:text=The%20Karolinska%20Directed%20Emotional%20Faces,from%20the%20original%20KDEF%20images

Malpass RS, Kravitz J. 1969. Recognition for faces of own and other race. J. Personal. Soc. Psychol 13(4):330–34

Matthews CM, Mondloch CJ. 2018. Improving identity matching of newly encountered faces: effects of multi-image training. J. Appl. Res. Mem. Cogn 7(2):280–90

Maurer D, Le Grand R, Mondloch CJ. 2002. The many faces of configural processing. Trends Cogn. Sci 6(6):255–60 [PubMed: 12039607]

Maze B, Adams J, Duncan JA, Kalka N, Miller T, et al. 2018. IARPA Janus Benchmark—C: face dataset and protocol. In Proceedings of the 2018 International Conference on Biometrics (ICB), pp. 158–65. Piscataway, NJ: IEEE

McCurrie M, Beletti F, Parzianello L, Westendorp A, Anthony S, Scheirer WJ. 2017. Predicting first impressions with deep learning. In Proceedings of the 2017 IEEE International Conference on Automatic Face & Gesture Recognition *(FG* 2017), pp. 518–25. Piscataway, NJ: IEEE

Murphy J, Ipser A, Gaigg SB, Cook R. 2015. Exemplar variance supports robust learning of facial identity. J. Exp. Psychol. Hum. Percept. Perform 41(3):577–81 [PubMed: 25867504]

Natu VS, Barnett MA, Hartley J, Gomez J, Stigliani A, Grill-Spector K. 2016. Development of neural sensitivity to face identity correlates with perceptual discriminability. J. Neurosci 36(42):10893–907 [PubMed: 27798143]

Natu VS, Jiang F, Narvekar A, Keshvari S, Blanz V, O’Toole AJ. 2010. Dissociable neural patterns of facial identity across changes in viewpoint. J. Cogn. Neurosci 22(7):1570–82 [PubMed: 19642884]

Nordt M, Gomez J, Natu V, Jeska B, Barnett M, Grill-Spector K. 2019. Learning to read increases the informativeness of distributed ventral temporal responses. Cereb. Cortex 29(7):3124–39 [PubMed: 30169753]

Nordt M, Gomez J, Natu VS, Rezai AA, Finzi D, Grill-Spector K. 2020. Selectivity to limbs in ventral temporal cortex decreases during childhood as selectivity to faces and words increases. J. Vis 20(11):152

Noyes E, Jenkins R. 2019. Deliberate disguise in face identification. J. Exp. Psychol. Appl 25(2):280–90 [PubMed: 30730157]

Noyes E, Parde C, Colon Y, Hill M, Castillo C, et al. 2021. Seeing through disguise: getting to know you with a deep convolutional neural network. Cognition. In press

Noyes E, Phillips P, O’Toole A. 2017. What is a super-recogniser. In Face Processing: Systems, Disorders and Cultural Differences, ed. Bindemann M, pp. 173–201. Hauppage, NY: Nova Sci. Publ.

Oosterhof NN, Todorov A. 2008. The functional basis of face evaluation. PNAS 105(32):11087–92 [PubMed: 18685089]

O’Toole AJ, Castillo CD, Parde CJ, Hill MQ, Chellappa R. 2018. Face space representations in deep convolutional neural networks. Trends Cogn. Sci 22(9):794–809 [PubMed: 30097304]

O’Toole AJ, Phillips PJ, Jiang F, Ayyad J, Pénard N, Abdi H. 2007. Face recognition algorithms surpass humans matching faces over changes in illumination. IEEE Trans. Pattern Anal. Mach. Intel (9):1642–46

Parde CJ, Castillo C, Hill MQ, Colon YI, Sankaranarayanan S, et al. 2017. Face and image representation in deep CNN features. In Proceedings of the 2017 IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 673–80. Piscataway, NJ: IEEE

Parde CJ, Colón YI, Hill MQ, Castillo CD, Dhar P, O’Toole AJ. 2021. Face recognition by humans and machines: closing the gap between single-unit and neural population codes—insights from deep learning in face recognition. J. Vis In press

Parde CJ, Hu Y, Castillo C, Sankaranarayanan S, O’Toole AJ. 2019. Social trait information in deep convolutional neural networks trained for face identification. Cogn. Sci 43(6):e12729 [PubMed: 31204800]

Parkhi OM, Vedaldi A, Zisserman A. 2015. Deep face recognition. Rep., Vis. Geom. Group, Dept. Eng. Sci., Univ. Oxford, UK

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, et al. 2019. Pytorch: an imperative style, high-performance deep learning library. In NeurIPS 2019: Proceedings of the 32nd International Conference on Neural Information Processing Systems, pp. 8024–35. New York: ACM

Pezdek K, Blandon-Gitlin I, Moore C. 2003. Children’s face recognition memory: more evidence for the cross-race effect. J. Appl. Psychol 88(4):760–63 [PubMed: 12940414]

Phillips PJ, Beveridge JR, Draper BA, Givens G, O’Toole AJ, et al. 2011. An introduction to the good, the bad, & the ugly face recognition challenge problem. In Proceedings of the 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 346–53. Piscataway, NJ: IEEE

Phillips PJ, O’Toole AJ. 2014. Comparison of human and computer performance across face recognition experiments. Image Vis. Comput 32(1):74–85

Phillips PJ, Yates AN, Hu Y, Hahn CA, Noyes E, et al. 2018. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. PNAS 115(24):6171–76 [PubMed: 29844174]

Poggio T, Banburski A, Liao Q. 2020. Theoretical issues in deep networks. PNAS 117(48):30039–45 [PubMed: 32518109]

Ponce CR, Xiao W, Schade PF, Hartmann TS, Kreiman G, Livingstone MS. 2019. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. Cell 177(4):999–1009 [PubMed: 31051108]

Ranjan R, Bansal A, Zheng J, Xu H, Gleason J, et al. 2019. A fast and accurate system for face detection, identification, and verification. IEEE Trans. Biom. Behav. Identity Sci 1(2):82–96

Ranjan R, Castillo CD, Chellappa R. 2017. L2-constrained softmax loss for discriminative face verification. arXiv:1703.09507 [cs.CV]

Ranjan R, Sankaranarayanan S, Castillo CD, Chellappa R. 2017c. An all-in-one convolutional neural network for face analysis. In Proceedings of the 2017 IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), pp. 17–24. Piscataway, NJ: IEEE

Richards BA, Lillicrap TP, Beaudoin P, Bengio Y, Bogacz R, et al. 2019. A deep learning framework for neuroscience. Nat. Neurosci 22(11):1761–70 [PubMed: 31659335]

Ritchie KL, Burton AM. 2017. Learning faces from variability. Q. J. Exp. Psychol 70(5):897–905

Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P. 1976. Basic objects in natural categories. Cogn. Psychol 8(3):382–439

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, et al. 2015. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis 115(3):211–52

Russell R, Duchaine B, Nakayama K. 2009. Super-recognizers: people with extraordinary face recognition ability. Psychon. Bull. Rev 16(2):252–57 [PubMed: 19293090]

Sangrigoli S, Pallier C, Argenti AM, Ventureyra V, de Schonen S. 2005. Reversibility of the other-race effect in face recognition during childhood. Psychol. Sci 16(6):440–44 [PubMed: 15943669]

Sankaranarayanan S, Alavi A, Castillo C, Chellappa R. 2016. Triplet probabilistic embedding for face verification and clustering. arXiv:1604.05417 [cs.CV]

Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, et al. 2018. Brain-score: Which artificial neural network for object recognition is most brain-like? bioRxiv 407007. 10.1101/407007

Schroff F, Kalenichenko D, Philbin J. 2015. Facenet: a unified embedding for face recognition and clustering. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–23. Piscataway, NJ: IEEE

Scott LS, Monesson A. 2010. Experience-dependent neural specialization during infancy. Neuropsychologia 48(6):1857–61 [PubMed: 20153343]

Sengupta S, Chen JC, Castillo C, Patel VM, Chellappa R, Jacobs DW. 2016. Frontal to profile face verification in the wild. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–9. Piscataway, NJ: IEEE

Sim T, Baker S, Bsat M. 2002. The CMU pose, illumination, and expression (PIE) database. In Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition, pp. 53–58. Piscataway, NJ: IEEE

Simonyan K, Zisserman A. 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [cs.CV]

Smith LB, Jayaraman S, Clerkin E, Yu C. 2018. The developing infant creates a curriculum for statistical learning. Trends Cogn. Sci 22(4):325–36 [PubMed: 29519675]

Smith LB, Slone LK. 2017. A developmental approach to machine learning? Front. Psychol 8:2124 [PubMed: 29259573]

Song A, Linjie L, Atalla C, Gottrell G. 2017. Learning to see people like people: predicting social impressions of faces. Cogn. Sci 2017:1096–101

Storrs KR, Kietzmann TC, Walther A, Mehrer J, Kriegeskorte N. 2020. Diverse deep neural networks all predict human it well, after training and fitting. bioRxiv 2020.05.07.082743. 10.1101/2020.05.07.082743

Su H, Maji S, Kalogerakis E, Learned-Miller E. 2015. Multi-view convolutional neural networks for 3d shape recognition. In Proceedings of the 2015 IEEE International Conference on Computer Vision, pp. 945–53. Piscataway, NJ: IEEE

Sugden NA, Moulson MC. 2017. Hey baby, what's "up"? One-and 3-month-olds experience faces primarily upright but non-upright faces offer the best views. Q. J. Exp. Psychol 70(5):959–69

Taigman Y, Yang M, Ranzato M, Wolf L. 2014. Deepface: closing the gap to human-level performance in face verification. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–8. Piscataway, NJ: IEEE

Tanaka JW, Pierce LJ. 2009. The neural plasticity of other-race face recognition. Cogn. Affect. Behav. Neurosci 9(1):122–31 [PubMed: 19246333]

Terhörst P, Fährmann D, Damer N, Kirchbuchner F, Kuijper A. 2020. Beyond identity: What information is stored in biometric face templates? arXiv:2009.09918 [cs.CV]

Thorpe S, Fize D, Marlot C. 1996. Speed of processing in the human visual system. Nature 381(6582):520–22 [PubMed: 8632824]

Todorov A 2017. Face Value: The Irresistible Influence of First Impressions. Princeton, NJ: Princeton Univ. Press

Todorov A, Mandisodza AN, Goren A, Hall CC. 2005. Inferences of competence from faces predict election outcomes. Science 308(5728):1623–26 [PubMed: 15947187]

Valentine T 1991. A unified account of the effects of distinctiveness, inversion, and race in face recognition. Q. J. Exp. Psychol. A 43(2):161–204 [PubMed: 1866456]

van der Maaten L, Weinberger K. 2012. Stochastic triplet embedding. In Proceedings of the 2012 IEEE International Workshop on Machine Learning for Signal Processing, pp. 1–6. Piscataway, NJ: IEEE

Walker M, Vetter T. 2009. Portraits made to measure: manipulating social judgments about individuals with a statistical face model. J. Vis 9(11):12

Wang F, Liu W, Liu H, Cheng J. 2018. Additive margin softmax for face verification. IEEE Signal Process. Lett 25:926–30

Wang F, Xiang X, Cheng J, Yuille AL. 2017. Normface: $L_2$ hypersphere embedding for face verification. In MM '17: Proceedings of the 25th ACM International Conference on Multimedia, pp. 1041–49. New York: ACM

Xie C, Tan M, Gong B, Wang J, Yuille AL, Le QV. 2020. Adversarial examples improve image recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 819–28. Piscataway, NJ: IEEE

Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. PNAS 111(23):8619–24 [PubMed: 24812127]

Yi D, Lei Z, Liao S, Li SZ. 2014. Learning face representation from scratch. arXiv:1411.7923 [cs.CV]

Yoshida H, Smith LB. 2008. What's in view for toddlers? Using a head camera to study visual experience. Infancy 13(3):229–48 [PubMed: 20585411]

Young AW, Burton AM. 2020. Insights from computational models of face recognition: a reply to Blauch, Behrmann and Plaut. Cognition 208:104422 [PubMed: 32800311]

Yovel G, Abudarham N. 2020. From concepts to percepts in human and machine face recognition: a reply to Blauch, Behrmann & Plaut. Cognition 208:104424 [PubMed: 32819709]

Yovel G, Halsband K, Pelleg M, Farkash N, Gal B, Goshen-Gottstein Y. 2012. Can massive but passive exposure to faces contribute to face recognition abilities? J. Exp. Psychol. Hum. Percept. Perform 38(2):285–89 [PubMed: 22288697]

Yovel G, O'Toole AJ. 2016. Recognizing people in motion. Trends Cogn. Sci 20(5):383–95 [PubMed: 27016844]

Yuan L, Xiao W, Kreiman G, Tay FE, Feng J, Livingstone MS. 2020. Adversarial images for the primate brain. arXiv:2011.05623 [q-bio.NC]

Yue X, Cassidy BS, Devaney KJ, Holt DJ, Tootell RB. 2010. Lower-level stimulus features strongly influence responses in the fusiform face area. Cereb. Cortex 21(1):35–47 [PubMed: 20375074]

### PERSPECTIVE: THEORIES AND MODELS OF FACE PROCESSING

Box (1976) reminds us that scientific progress comes from motivated iteration between theory and practice. In understanding human face processing, theories should be used to generate the questions, and machines (as models) should be used to answer the questions. Three elemental concepts are required for scientific progress. The first is flexibility. Effective iteration between theory and practice requires feedback between what the theory predicts and what the model reveals. The second is parsimony. Because all models are wrong, excessive elaboration will not find the correct model. Instead, economical descriptions of a phenomenon should be preferred over complex descriptions that capture less fundamental elements of human perception. Third, Box (1976, p. 792) cautions us to avoid "worrying selectivity" in model evaluation. As he puts it, "since all models are wrong, the scientist must be alert to what is importantly wrong."

These principles represent a scientific ideal, rather than a reality in the field of face perception by humans and machines. Applying scientific principles to computational modeling of human face perception is challenging for diverse reasons (see the sidebar titled Caveat: Iteration Between Theory and Practice below). We argue, as Cichy & Kaiser (2019) have, that although the utility of scientific models is usually seen in terms of prediction and explanation, their function for exploration should not be underrated. As scientific models, DCNNs carry out high-level visual tasks in neurally inspired ways. They are at a level of development that is ripe for exploring computational and representational principles that actually work but are not understood. This is a classic problem in reverse engineering—yet the use of deep learning as a model introduces a dilemma. The goal of reverse engineering is to understand how a functional but highly complex system (e.g., the brain and human visual system) solves a problem (e.g., recognizes a face). To accomplish this, a well-understood model is used to test hypotheses about the underlying mechanisms of the complex system. A prerequisite of reverse engineering is that we understand how the model works. Failing that, we risk using one poorly understood system to test hypotheses about another poorly understood system. Although deep networks are not black boxes (every parameter is knowable) (Hasson et al. 2020), we do not fully understand how they recognize faces (Poggio et al. 2020). Therefore, the primary goal should be to understand deep networks for face recognition at a conceptual and representational level.

## CAVEAT: ITERATION BETWEEN THEORY AND PRACTICE

Box (1976) noted that scientific progress depends on motivated iteration between theory and practice. Unfortunately, a motivation to iterate between theory and practice is not a reasonable expectation for the field of computer-based face recognition. Automated face recognition is big business, and the best models were not developed to study human face processing. DCNNs provide a neurally inspired, but not copied, solution to face processing tasks. Computer scientists formulated DCNNs at an abstract level, based on neural networks from the 1980s (Fukushima 1988). Current DCNN-based models of human face processing are computationally refined, scaled-up versions of these older networks. Algorithm developers make design and training decisions for performance and computational efficiency. In using DCNNs to model human face perception, researchers must choose between smaller, controlled models and larger-scale, uncontrolled networks (see also Richards et al. 2019). Controlled models are easier to analyze but can be limited in computational power and training data diversity. Uncontrolled models better emulate real neural systems but may be intractable. The easy availability of cutting-edge pretrained face recognition models, with a variety of architectures, has been the deciding factor for many research labs with limited resources and expertise to develop networks. Given the widespread use of these models in vision science, brain-similarity metrics for artificial neural networks have been developed (Schrimpf et al. 2018). These produce a Brain-Score made up of a composite of neural and behavioral benchmarks. Some large-scale (uncontrolled) network architectures used in modeling human face processing (See Section 2.1) score well on these metrics.

A promising long-term strategy is to increase the neural accuracy of deep networks (Grill-Spector et al. 2018). The ventral visual stream and DCNNs both enable hierarchical and feedforward processing. This offers two computational benefits consistent with DCNNs as models of human face processing. First, the universal approximation theorem (Hornik et al. 1989) ensures that both types of networks can approximate any complex continuous function relating the input (visual image) to the output (face identity). Second, linear and nonlinear feedforward connections enable fast computation consistent with the speed of human facial recognition (Grill-Spector et al. 2018, Thorpe et al. 1996). Although current DCNNs lack other properties of the ventral visual system, these can be implemented as the field progresses.

### NEURONS, NEURAL TUNING, POPULATION CODES, FEATURES, AND PERCEPTUAL CONSTANCY

Barlow (1972, p. 371) wrote, "Results obtained by recording from single neurons in sensory pathways…obviously tell us something important about how we sense the world around us; but what exactly have we been told?" In answer, Barlow (1972, p. 371) proposed that "our perceptions are caused by the activity of a rather small number of neurons selected from a very large population of predominantly silent cells. The activity of each single cell is thus an important perceptual event and it is thought to be related quite simply to our subjective experience." Although this proposal is sometimes caricatured as the grandmother cell doctrine (see also Gross 2002), Barlow simply asserts that single-unit activity can be interpreted in perceptual terms, and that the responses of small numbers of units, in combination, underlie subjective perceptual experience. This proposal reflects ideas gleaned from studies of early visual areas that have been translated, at least in part, to studies of high-level vision.

Over the past decade, single neurons in face patches have been characterized as selective for facial features (e.g., aspect ratio, hair length, eyebrow height) (Freiwald et al. 2009), face viewpoint and identity (Freiwald & Tsao 2010), eyes (Issa & DiCarlo 2012), and shape or appearance parameters from an active appearance model of facial synthesis (Chang & Tsao 2017). Neurophysiological studies of face and object processing also employ techniques aimed at understanding neural population codes. Using the pattern of neural responses in a population of neurons (e.g., IT), linear classifiers are used often to predict subjective percepts (commonly defined as the image viewed). For example, Chang & Tsao (2017) showed that face images viewed by a macaque could be reconstructed using a linear combination of the activity of just 205 face cells in face patches ML–MF and AM. This classifier provides a real neural network model of the face-selective cortex that can be interpreted in simple terms.

Population code models generated from real neural data (a few hundred units), however, differ substantially in scale from the face- and object-selective cortical regions that they model ($1mm^3$ of the cerebral cortex contains approximately 50,000 neurons and 300 million adjustable parameters; Azevedo et al. 2009, Kandel et al. 2000, Hasson et al. 2020). This difference in scale is at the core of a tension between model interpretability and real-world task generalizability (Hasson et al. 2020). It also creates tension between the neural coding hypotheses suggested by deep learning and the limitations of current neuroscience techniques for testing these hypotheses. To model neural function, an electrode gives access to single neurons and (with multi-unit recordings) to relatively small numbers of neurons (a few hundred). Neurocomputational theory based on direct fit models posits that overparameterization (i.e., the extremely high number of parameters available for neural computation) is critical to the brain's solution to real-world problems (see Section 3.2). Bridging the gap between the computational and neural scale of these perspectives remains an ongoing challenge for the field.

## SUMMARY POINTS

1. Face representations generated by DCNN networks trained for identification retain information about the face (e.g., identity, demographics, attributes, traits, expression) and the image (e.g., viewpoint).
2. Deep learning face networks generate a surprisingly structured face representation from unstructured training with in-the-wild face images.
3. Individual output units from deep networks are unlikely to signal the presence of interpretable features.
4. Fundamental structural aspects of high-level visual codes for faces in deep networks replicate over a wide variety of network architectures.
5. Diverse learning mechanisms in DCNNs, applied simultaneously or in sequence, can be used to model human face perception across the lifespan.

### FUTURE ISSUES

1. Large-scale systematic manipulations of training data (race, ethnicity, image variability) are needed to give insight into the role of experience in structuring face representations.
2. Fundamental challenges remain in understanding how to combine deep networks for face, object, and scene recognition in ways analogous to the human visual system.
3. Deep networks model the ventral visual stream at a generic level, arguably up to the level of the IT cortex. Future work should examine how downstream systems, such as face patches, could be connected into this system.
4. In rethinking the goals of face processing, we argue in this review that some longstanding assumptions about visual representations should be reconsidered. Future work should consider novel experimental questions and employ methods that do not rely on these assumptions.

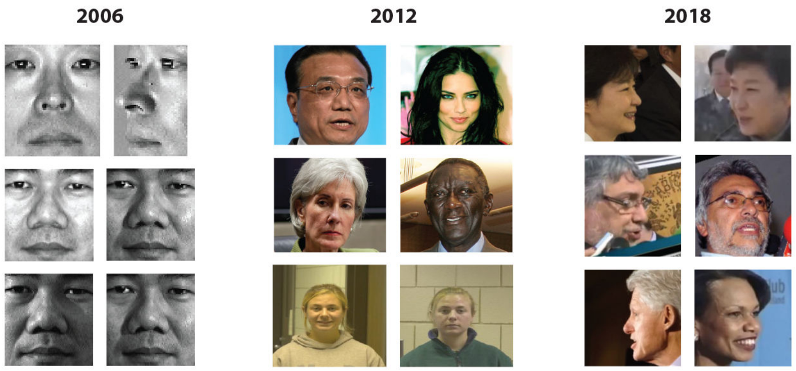

**Figure 1.**

The progress of computer-based face recognition systems can be tracked by their ability to recognize faces with increasing levels of image and appearance variability. In 2006, highly controlled, cropped face images with moderate variability, such as the images of the same person shown, were challenging (images adapted with permission from Sim et al. 2002). In 2012, algorithms could tackle moderate image and appearance variability (the top 4 images are extreme examples adapted with permission from Huang et al. 2012; the bottom two images adapted with permission from Phillips et al. 2011). By 2018, deep convolutional neural networks (DCNNs) began to tackle wide variation in image and appearance, (images adapted with permission from the database in Maze et al. 2018). In the 2012 and 2018 images, all side-by side images show the same person except the bottom pair of 2018 panels.

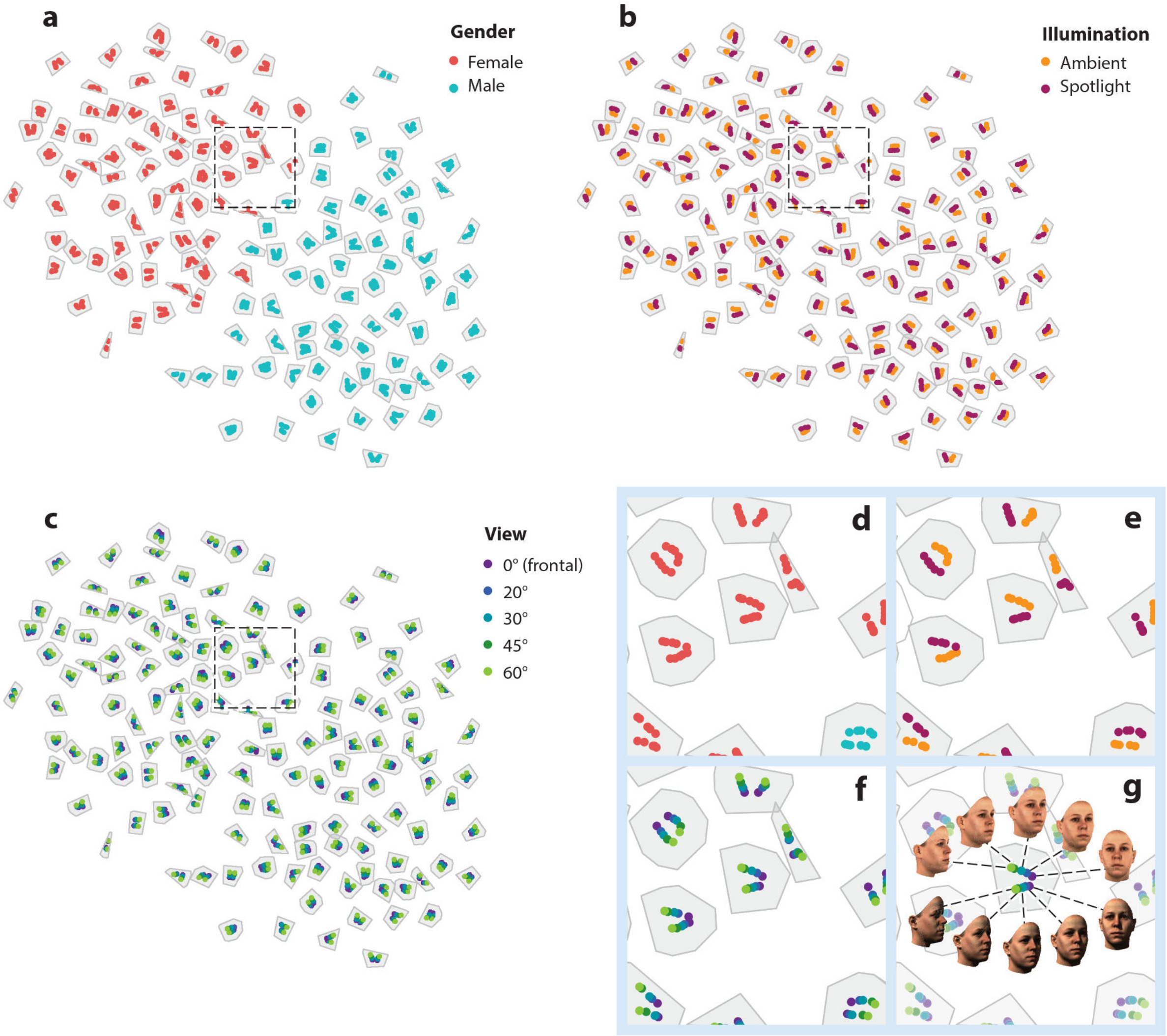

**Figure 2.**

Visualization of the top-level deep convolutional neural network (DCNN) similarity space for all images from Hill et al. (2019). (*a–f*) Points are colored according to different variables. Grey polygonal borders are for illustration purposes only and show the convex hull of all images of each identity. These convex hulls are expanded by a margin for visibility. The network separates identities accurately. In panels *a* and *d*, the space is divided into male and female sections. In panels *b* and *e*, illumination conditions subdivide within identity groupings. In panels *c* and *f*, the viewpoint varies sequentially within illumination clusters. Dotted-line boxes in panels *a–c* show areas enlarged in panels *d–g*. Figure adapted with permission from Hill et al. (2019).

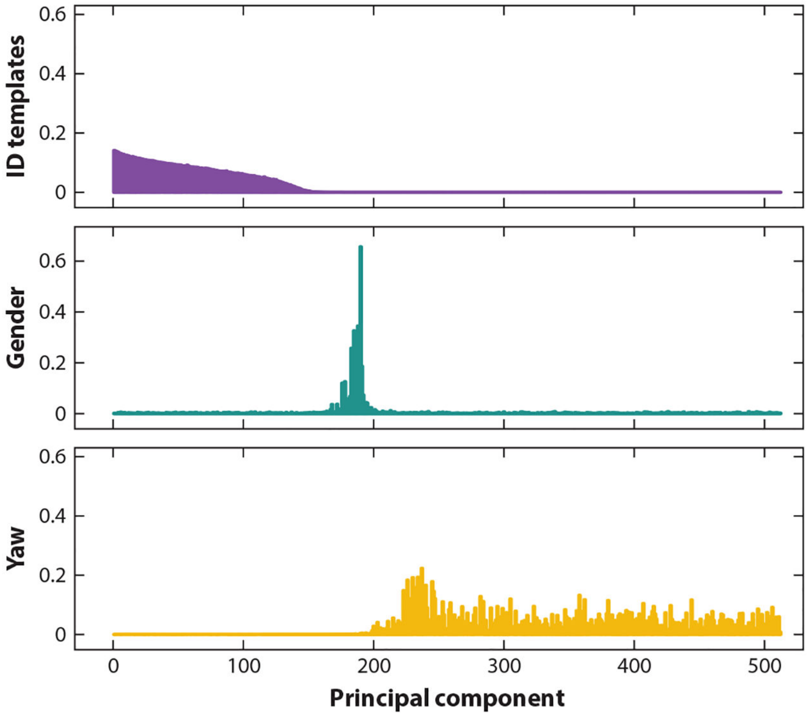

**Figure 3.**

Illustration of the separation of the task-relevant information into subspaces for an identity-trained deep convolutional neural network (DCNN). Each plot shows the similarity (cosine) between principal components (PCs) of the face space and directional vectors in the space that are diagnostic of identity (*top*), gender (*middle*), and viewpoint (*bottom*). Figure adapted with permission from Parde et al. (2021).

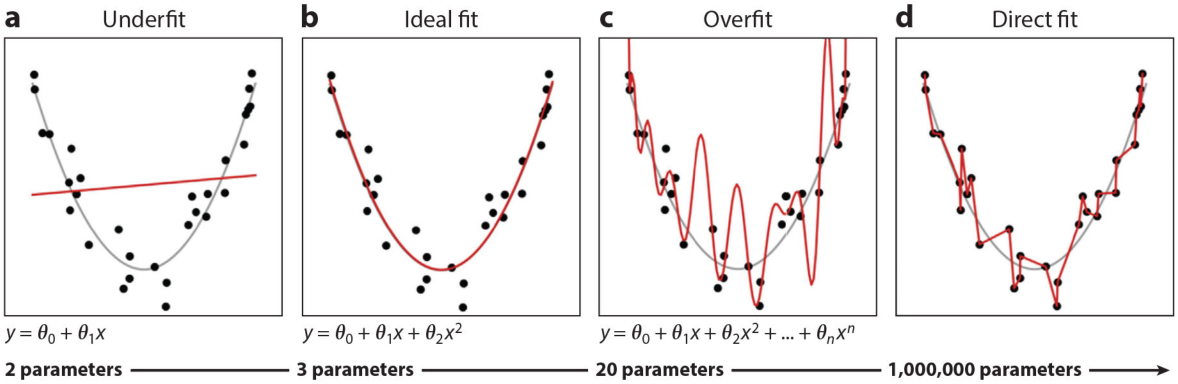

**Figure 4.**

(*a*) A model with too few parameters fails to fit the data. (*b*) The ideal-fit model fits with a small number of parameters and has generative power that supports interpolation and extrapolation. (*c*) An overfit function can model noise in the training data. (*d*) An overparameterized model generalizes well to new stimuli within the scope of the training samples. Figure adapted with permission from Hasson et al. (2020).