Seeking Safety in Knowledge

Jennifer Nagel
UNIVERSITY OF TORONTO

Sosa Prize lecture delivered at the one hundred twentieth Central Division meeting of the American Philosophical Association on February 24, 2023.

There is a difference between knowing something and just happening to be right about it. The nature of this contrast is hard to explain, but its sheer existence is somehow easy to recognize. In Plato's *Meno*, when Socrates sketches a theory of knowledge as a state of mind chained to its cause, he downplays that theory as mere conjecture. "And yet," Socrates states emphatically, "that knowledge differs from true opinion is no matter of conjecture with me. There are not many things which I profess to know, but this is most certainly one of them."[1]

Socrates's certainty on this matter is striking, but it does not take his philosophical vision to spot a difference between knowledge and mere true opinion. This distinction is also drawn by ordinary people without philosophical training, across genders and cultures;[2] indeed, there is evidence that nonhuman primates also distinguish knowledge from coincidentally correct judgment.[3]

What is not so clear is why we are naturally registering this contrast, or how. For practical purposes, as Plato observes, true opinion might seem to fit our needs exactly. When you come to a fork in the road on your way to Larissa, it could seem no better to have the right path pointed out by a knowledgeable expert than to have the same direction endorsed by a superficially similar guesser.[4] If a veridical hallucination leads a desert traveler to water in the valley ahead, his thirst ends up being quenched every bit as well as if he had known that there was water there.[5] It is easy to see why we care about the difference between truth and falsity, and easy to see how the world supplies us with feedback about the truth value of judgments. It is harder to explain exactly what we are gaining from tracking the stronger condition of knowledge, or how we are doing so.

Contemporary epistemology has zeroed in on a possible clue to this puzzle: whatever else it may involve, the difference between knowledge and mere true opinion seems to have a modal dimension. A landmark paper of Ernest Sosa's argued that the modal core of knowledge is a safety condition, according to which, when you know, "you would so believe only if your belief were true."[6] Unlike the person who knows the way, the guesser could easily have been wrong about which path to take; unlike the person who actually sees water, the hallucinator could easily have been wrong about where to find a drink. This general idea of knowledge demanding safe judgment, or the avoidance of error in sufficiently similar cases, has been embraced by contemporary epistemologists of various persuasions.[7] However, ongoing debates over safety raise notoriously difficult questions about what exactly it means to "so believe," or to have a sufficiently similar case to the case at hand. Meanwhile, the necessity of safety for knowledge has been challenged,[8] and indeed Sosa himself turned against the safety condition for a time,[9] only to return to it later in a more elaborate form.[10] Naive early theories about safety seem vulnerable to simple counterexamples,[11] and as theories have grown more sophisticated, they have been tested by increasingly exotic scenarios. One much-discussed case involves a grandfather clock, intermittently controlled by an invisible demon who wants to ensure that a subject forms a particular belief about the time.[12] A more recent article introduces a variant of an earlier case involving a costume party at a house that is hard to find, hosted by someone with an aversion to one potential guest; in the latest version, guests who almost decide to get dressed up as each other need to pass by a crossroads with an automatic device that may or may not direct them to their destination, depending on their apparent identity.[13] Alert to the threats posed by all these counterexamples, sophisticated recent theories have started to invoke powerful terms such as "competence,"[14] whose epistemic character is rich enough to generate worries that one mystery is being explained in terms of another, and possibly in a manner that incorporates some problematic circularity. Meanwhile, inventive counterexamples have arisen to challenge the extensional adequacy of these newer theories as well.[15]

It is puzzling that a contrast so readily grasped intuitively puts up so much resistance to being captured reflectively, in readily stated principles. As explicit theories of safety have grown more complex, it has become increasingly mystifying how we could naturally calculate the presence of knowledge in real time, if knowledge does indeed demand safety. It has also become increasingly mysterious what we would ever gain from doing so: Why on earth should we care so much about what would have happened if things had been slightly different? At the same time, we may find ourselves with methodological worries about our epistemological speculations, worries that may be heightened when strangely baroque scenarios either fail to generate clear intuitions in us, or generate intuitive responses whose legitimacy we are hard-pressed to explain.

If philosophers have not been gaining ground through a direct search for high-level explicit principles about safety, strategically navigating between stock examples and counterexamples, perhaps it is time to try a fresh approach. In what follows, I approach the problem of safety from below, focusing on a concrete cognitive capacity of ours, and looking at what its actual workings can tell us, first about the nature of epistemic safety, and then about the functional value of securing and tracking it. My chosen capacity is face recognition, in part because this is a fairly uncontroversial source of knowledge—we don't ordinarily doubt that a typical person can know at a glance that a good friend of theirs has entered the room—and in part because it still affords a useful variety of problem cases, for example, cases involving secret twins, plastic surgery, or one person mistaken for another at a distance. Having noted the existence of these problem cases, and having perhaps experienced a brief moment of fear that they will conspire to make true safety impossible, we will start with a focus on how face recognition works when all goes well. A clear sense of the good case will be useful in situating epistemic safety in the ordinary workings of this capacity, before turning to tackle problem cases and safety failures.

An interesting feature of face recognition is that we lack introspective access into its operations; although we can tell at a glance that the person immediately facing us is some particular close friend, we are hard-pressed to formulate any explicit description that uniquely selects that individual, or to state any high-level theoretical principles that generally structure our facial identifications. After reviewing how face recognition works, and explaining how it incorporates epistemic safety, I will argue that there is an important parallel between the recognition of a familiar face, and the recognition of knowledge itself. Indeed, the reason why we are unable to introspect our competence in face recognition turns out to be essentially the same as the reason why we are unable to provide a reductive analysis of knowledge. The concluding section of this paper tackles the higher-level question of human knowledge of knowledge, but to approach this problem from below, we will start on the ground floor, reviewing ordinary human knowledge of faces.

1. AN OVERVIEW OF FACE RECOGNITION

It has been estimated that the average adult can recognize the faces of roughly five thousand people, encompassing both personal acquaintances and celebrities.[16] Knowing a face does not require knowing a name: researchers count a face as known when the experimental subject can produce some appropriate identifying description ("the President of China"; "my favorite barista"), or, in some studies, when the subject simply responds to various different photos of the individual, interspersed amid distractors, by consistently judging that this person is familiar. A photo of a familiar person sparks an immediate and distinctive neural response, starting just 140ms after presentation,[17] and it generally takes less than a second to identify the person seen, assuming their face is known.[18]

It is remarkable that an average human can recognize so many people in this way.[19] Human faces are broadly similar in their basic configuration, so distinguishing individuals from one another requires subtle sensitivity to small differences, where combinations of anatomically realistic perceivable differences define a space of trillions of theoretically distinguishable faces.[20] To complicate matters further, photographs or percepts of a single individual exhibit large differences due to changing facial expressions, hairstyles, aging, blemishes, makeup, facial hair, illumination, angle of view, distance, and partial occlusion. A double challenge therefore confronts us: we need to explain "not only how we tell people apart, but also how we tell people together."[21]

The difficulty of telling people apart is well known, but the difficulty of "telling people together" may need to be spelled out. When a person is unfamiliar, it can be hard to judge that different pictures of them in fact depict the same person. This problem was often skimmed over in the early literature on face recognition, not least because many early studies started by minimizing the variation in their stimuli, for example by taking studio portraits of a series of individuals with neutral expressions, under uniform lighting, viewpoint, and pose. More naturalistic approaches to real-life face recognition need to use images taken from various angles, and in various conditions. In one study aiming to capture this kind of variability, researchers scraped twenty "ambient" images of each of two blonde Dutch celebrities from the internet, and printed them in greyscale onto laminated cards. British experimental participants were given the shuffled deck of forty images and instructed to sort them by identity into any number of groups, where the photos in each group were all to be of the same person. Not one of the twenty participants solved the puzzle correctly. It was rare for participants to lump photos of the two different women into the same pile (on average less than one such mistake per participant, mode 0, range 0–3). The main problem was in the other direction: the median number of perceived identities was between seven and eight, and the overall range was between three and sixteen. The researchers concluded: "This pattern indicates that the problem is primarily one of integrating dissimilar images. It is difficult to find commonalities among photos of the same face that justify grouping them together. At the same time, it is easy to find differences that justify grouping them separately."[22]

One might worry that these photos were atypically diverse in appearance (one of the celebrities was an actress, the other a model and TV star in multiple roles), or that the greyscale photos were simply too poor in quality to support accurate face identification. Twenty Dutch participants were recruited to perform the same task with those forty cards, and almost all of them did so perfectly (median 2, mode 2, range 2–5).[23] It is familiarity with a face that enables the recognition of its identity across variation. This is a robust result, confirmed across a variety of experimental procedures. When, and only when, people are familiar with an individual, they can easily recognize that person on a brief, low-quality, black-and-white security camera video feed, picking their photo out of an array presented just afterwards.[24] For unfamiliar faces, it can be hard even to match a target face to a simultaneously presented lineup of ten photos, with no advantage when the unfamiliar target is not just another photograph but present as a live actor currently in front of the subject.[25] Even passport officers show surprisingly poor performance identifying whether or not an unfamiliar live person facing them is the same as the person whose photograph they are currently viewing on screen.[26]

The capacity to recognize an individual face is a many-to-one mapping from a domain of possible percepts (approximated by photos) to a target identity. It is important for the capacity to cover *possible* percepts: face recognition is not simply a matter of memorizing a stock of actual photos or percepts and then retrieving the correlated identity when we encounter some exact member of that stock. In order to select the correct underlying identity when we encounter someone in the wild, or see a new photo of them, we need to be equipped to deal with novel combinations of expression, viewpoint, aging, lighting, fluctuation in weight, hairstyle, and so forth. Somehow, past experience informs a model in us that is applicable to new data, and the telling-people-together problem suggests that the formation of this model is subtly individualized for each identity. There is no single recipe for converting a frontal view to a profile, or for predicting the dynamic changes incurred by shifts between various neutral and emotional expressions. In the words of one review, "the ways in which one person's face varies are different from the ways in which someone else's face will vary. To recognise Angela Merkel from any image of her, then, our brains need to have learned how to take into account this idiosyncratic, Merkel-specific variability."[27]

If the exact variations in the full set of a single person's possible images are unique to her, this is not to say that there are no general patterns to be detected: all variations happen within bounds set by the genetic determination of human face shapes, the physiological course of aging, and so forth. For any given face, experience with similar faces is helpful. Across all variations, face recognition is enhanced for the types of face with which an individual has most contact: people are generally more accurate in recognizing the faces of people of their own ethnicity,[28] with improved performance in recognizing faces of other ethnicities where there is greater cross-ethnic exposure.[29] Three-year-old children are better at distinguishing the faces of people who are the age of their caregivers, but above the age of five, children and adults of all ages are more accurate at recognizing the faces of their own age cohort.[30] Enhanced sensitivity to what individuates certain types of faces is driven not just by the quantity of contact, but also by social conditions motivating attention and depth of processing; performance with members of a group is improved by a need to track individual identities, as opposed to tracking mere group membership.[31]

The various differences in physiognomy that are used to distinguish and reidentify faces can be taken to structure a multidimensional Euclidean "face space" with a zone for each face, defined by its value for each of the dimensions, where this space is ultimately anchored either on some point of origin representing the average of all dimensions, or on some set of known face exemplars.[32] The dominant exemplar model is now expressed in terms of a Voronoi diagram (see Figure 1), in which each face defines a recognition zone or cell anchored on a prototype, divided from other cells by bisecting the distance along each dimension between this prototype and its nearest neighbours.[33] Any point within the recognition zone for a known face is closer to the prototype for that face than to the prototype for any other face. We can recognize the same person through changes such as aging, as long as these transformations do not render them more similar to another person than to the original. The overall structure of the "face space" that each person uses in recognizing others is biased by their own experience of faces, enabling finer-grained distinctions along dimensions that mark differences among more of the individuals one has encountered,

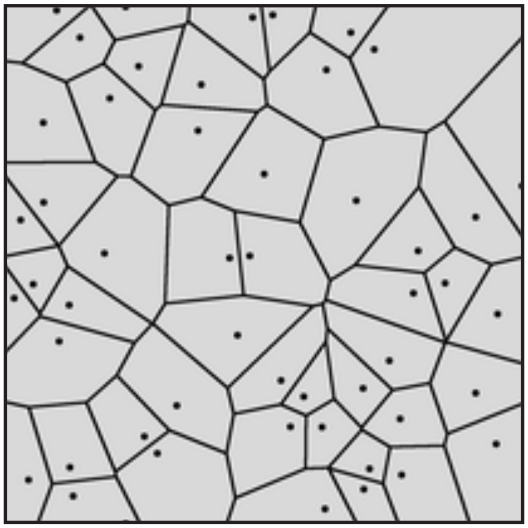

Figure 1. A two-dimensional Voronoi Diagram. Voronoi diagrams can be composed of polygons of any dimensionality, with all points inside each polygon closer to their own anchor point than to the anchor point of any other polygon.

where the selection of these most discriminating dimensions can vary by ethnicity and age.

The puzzle of identifying the relevant dimensions of facial variation is a difficult one. To see the depth of the problem, it will help to run through a quick history of efforts to solve it in artificial intelligence, as applied to all three of the standard applications of face recognition: verification (deciding whether two photos are of the same individual), clustering (sorting images by identity), and identification (matching a new image to a specific identity).

2. FACE RECOGNITION IN ARTIFICIAL INTELLIGENCE

Early approaches aimed to extract a relatively compact set of parameters that would together demarcate the individuality of faces, structuring "face space." Bruce and Young's Principal Components Analysis model, for example, sought a simplified code of the significant dimensions of facial variation. In this model, "a familiar face is represented by an interconnected set of descriptions—some describing the configuration of the whole face, and some describing the details of particular features."[34] The "face recognition unit" for each known face was taken to include both "view-centered descriptions and more abstract, expression-independent descriptions of the features and relations between them";[35] incoming percepts or photographs were taken to have their relevant structural features encoded to align with this stored structural unit. Matching over some threshold would activate the "person identification node" for the relevant individual, which could in turn trigger name generation, retrieval of biographical details, and so forth. (This person identification node could also be activated in other ways, for example by voice or gait recognition, and is connected with broader cognitive systems; indeed, one might have a person identification node for a recognized individual such as a composer or author whose face would be unfamiliar.) It might seem inevitable that any computationally tractable model of face recognition would have to work along these lines, but decades of experimental work failed to extract the crucial set of descriptive dimensions that could individuate faces in general. What was established was that any such approach would require a very large set of dimensions. For example, a model that begins by identifying twenty-seven facial landmarks through their distinctive local texture might still need to track 100,000 sets of relationships among these in order to achieve results that still fall short of human performance; sparser adaptations of such a model can ease the computational cost, but only with some loss of accuracy.[36]

A new approach emerged with the application of deep learning to large datasets such as Facebook's Social Face Classification (SFC) dataset, in which many individuals have multiple pictures of themselves, taken over some span of time, with different settings and expressions. To train one important model ("DeepFace"), researchers were able to use 4.4 million SFC face images from 4,030 people who each had 800–1,200 posted photos of themselves, exempting the most recent 5 percent of these images for use in the testing phase.[37] Rather than attempting to specify the crucial characteristic dimensions in advance, DeepFace is simply driven to find whatever patterns it can extract from these images to optimize its performance in identifying the corresponding individuals. An initial stage prepares uniform 152 x 152-pixel images of the millions of faces, resizing and cropping raw photos on the basis of simple facial landmarks, and rotating all faces into a uniform frontal alignment with some warp to save perspectivally distinctive characteristics. Each standardized image is then fed through a nine-layer deep neural network, in which early layers pick out low-level characteristics like edges and textures, and hierarchically organize them into patterns, which are fed into three locally connected middle layers, to support specialization in detecting local patterns found in different facial regions. Two fully connected final layers reintegrate the results to take advantage of larger relationships in the whole face, and deliver a probability distribution over the possible identities in the dataset. The input layer has 69,312 nodes or neurons (152 squared, multiplied by the three RGB pixel color values); the output layer has 4,030 neurons for the 4,030 identities in the dataset, each of which gets a real number value, expressing something like the input image's degree of resemblance to that identity. Those output-layer real numbers are then converted to a probability distribution over the set of identities. There is enormous computational power between these small input and output layers: the model as a whole has 120 million parameters or connections between its layers, parameters whose values are adjusted upwards or downwards in the course of learning, to send a stronger or weaker activation signal to the next layer.

It would in some sense be ideal for the model to select the correct identity for each input photo with perfect certainty. DeepFace learns by failing to do so: in the course of its supervised learning, the error signal (or "loss") of the model is the negative log of the output probability assigned to whatever was in fact the correct identity. This error signal drives a backpropagation algorithm that runs down the layers of the model in stochastic gradient descent, successively updating the parameters to minimize the loss in each iteration of training. As a result, after each training step, DeepFace tunes its 120 million parameters in the direction of ideal certainty about the truth, more radically when the model had delivered a lower probability for the correct identity than when it had come closer to certainty. When there are two close rival identities for an image (for example, because there are two people who look alike from a given angle), the model will correct itself after error feedback to heighten its reliance on whatever characteristics pushed it towards the correct identity for that image and decrease its reliance on the characteristics supporting the rival. The parameters start out as a blank slate, and evolve blindly to fit the goal of assigning as high a probability as possible to the correct identity for each image. Other than breaking up the middle layers to focus on different image regions (which roughly correspond to different facial zones because of the initial alignment step), the architecture of the model does nothing to dictate in advance what characteristics of an image will be used to identify it; the purely reactive updating process will simply end up amplifying reliance on whatever characteristics in fact support successful identification during training.

One final characteristic of the model deserves close attention: its capacity to generalize well to new faces seen after training is supported by a special feature of its penultimate fully connected layer. Known as "dropout," this feature randomly switches off the output of about half of the neurons in that layer (on each training cycle, each neuron has an independent 0.5 probability of being silenced). The researchers who originally developed this technique explain it as follows: "The neurons which are 'dropped out' in this way do not contribute to the forward pass and do not participate in back propagation. So every time an input is presented, the neural network samples a different architecture, but all these architectures share weights. This technique reduces complex co-adaptations of neurons, since a neuron cannot rely on the presence of particular other neurons. It is, therefore, forced to learn more robust features that are useful in conjunction with many different random subsets of the other neurons."[38] Adding 50 percent dropout to a model roughly doubles the training time required for it to converge on ideal performance with its training set, but models trained in this way generally perform better when they shift from training to test. Dropout addresses a vulnerability introduced by the sheer computational power of neural networks: because they have so many parameters, they are at risk of explosively overfitting their data during training, using their vast computational power to memorize the right answers for training set images on the basis of spurious correlations. To take a (somewhat unrealistic[39]) toy example, if all and only photos of Robert Redford in the training data happen to have a particular color combination of two particular pixels, an overfit model could use that coincidence to recognize him, and collapse during the test phase when presented with new photos. If roughly half of the final-layer neurons are randomly dropped on each training cycle, then identifications will tend to be made in some more robust way, and only genuine patterns should contribute to identification.[40] Dropout is one of dozens of "regularization" techniques applied in deep learning to improve a model's ability to generalize to unseen data; other common methods include randomly adding a bit of noise to the inputs, and cutting off training early, before the model has approached perfection on its training set.[41] Overfit models are brittle: by adding those two stray Redford pixels to a picture of someone else, we could induce an overfit model to misidentify him as Redford. To judge in that way is to judge in a manner that could easily err; a well-regularized model will be impervious to small changes in its inputs that are irrelevant to the target phenomenon, generalizing well from past experience to new cases.[42]

DeepFace generalizes well. Whatever characteristics it extracts in the course of its training to recognize faces, these characteristics do seem to have broader validity when applied to face verification problems involving new faces not encountered in training. Faces captured on video stills are generally harder to judge than those in photographs. However, a version of DeepFace set up for verification tasks[43] could classify pairs of new (untrained) faces from YouTube videos as the same or different at an accuracy level of 91.4 percent, where previous state-of-the-art systems were all below 80 percent accuracy.[44] DeepFace's verification performance on the easier "Labelled Faces in the Wild" (LFW) dataset hits 97.0 percent,[45] close to the human level of 97.5 percent accuracy for the same task.[46]

A slightly simpler deep learning model trained on a larger dataset does even better, by directly exploiting the "face space" framework. Google's FaceNet was trained on dataset of 200 million images drawn from 8 million identities.[47] This model skips the step of rotating faces to frontal view; each image is minimally prepared by tightly cropping a square around the face in its original orientation, and bringing it to a uniform size. The model is organized in 22 layers with 140 million parameters. The input layer takes a cropped image (220 x 220-RBG pixels), and the final output layer delivers a 128-dimensional embedding, a mathematical vector abstractly representing that image in face space. The goal of the model is to refine these output vectors so that they cluster in identities: the vector for a brightly lit frontal view of some individual needs to be closer to the vector for a dimly lit side view of that very person than it is to the vector of a brightly lit frontal view of anyone else, even someone who superficially resembles the target. Training is driven by a "Triplet Loss" procedure, in which each anchor image of a person X is compared to a distinct positive image of X and a negative image of some other person Y; the loss function minimizes the Euclidean distance between the vectors for the anchor and positive while maximizing the distance between the vectors for the anchor and negative.[48] When supplied with a large enough series of appropriately challenging triplets, FaceNet learns to produce vectors that capture some deep underlying invariance behind various images of a person. The recognitional capacity for a person sends all images of that person, with any viewpoint, pose, expression or illumination, into a distinct region of "face space," duly separated from the image zones of others. Given enough data, the triplet loss procedure will drive the model towards creating a tessellation of its 128-dimensional face space in which the vector for every image is closer to vectors for other images of that person than to vectors for anyone else. This model outperforms DeepFace, and indeed outperforms humans, verifying pairs of faces in the LFW dataset with an accuracy of 99.63 percent.[49]

It is not hard to locate epistemic safety in FaceNet. Given the model's success in testing, its 128-dimensional vectors compose a set of similarity spaces that work well to pick out individuals. Following training on photographs from a domain of suitably facially distinctive individuals, an image can be recognized (known to be of some particular individual) when its vector lies closer to the vectors of other images of that individual than to any vector of an image of anyone else, so that slight deviations from the existing image would still be correctly identified. To judge an image's identity in this manner is to judge in a way that could not easily err. We can also see how, part way through training, judgments could be unsafe near the not-yet-sharply drawn borders between zones for different individuals, even when judgments on central cases are already safe, just as you might recognize the identity of a minor celebrity in a clear frontal image, knowing who it is when seen that way, even if you cannot reliably pick him out from a dimly lit side view.[50] For this partially known face, even a correct identification in that fringe zone is not safely correct, as a slight deviation in the image would map to another identity.

However, safety failures at the blurred boundaries between individuals do not entail that no judgments are safe.

The core idea of tessellating face space is not a new one; what is new is that the dimensions of the tessellation are not calculated strategically or supplied in advance, but rather allowed to evolve in the massive set of model parameters, in response to massive data, through simple local calculations. The triplet loss procedure pushes the model towards a Voronoi tessellation: to the extent that training converges on this outcome, the resultant face recognition zones have some powerful properties. In particular, they will satisfy the Convexity criterion that Peter Gärdenfors and Igor Douven have argued is a feature of natural concepts: when a region is convex, for any two points in that region, every point on the line between those points also lies in that region.[51] As Douven and Gärdenfors point out, when convexity holds, there is an immediate payoff in learnability: if you learn, of a few vectors, that they all map onto Tom Hanks, then you have automatically learned that everything between those vectors also maps onto him, defining a rich space of possible percepts.

The success of the big data approach makes sense, given the complexity of facial recognition. Trillions of theoretically possible faces might be compared with each other in any number of ways, and there are no simple laws dictating which dimensions will best separate the faces one actually encounters. It is only on the basis of the particular distinctive characteristics these faces actually display that the most successful artificial neural networks learn how to tell them apart, and these networks are able to do so because the plasticity of their millions of parameters enables them to adapt to the actual complexities of face space, at least when driven by a loss function optimized for successful discernment.

3. THE BASIC STRUCTURE OF HUMAN FACE RECOGNITION

If big-data models occasioned a breakthrough to human-level performance in artificial intelligence, it does not follow trivially that human face recognition itself involves similar strategies. There are several ways in which big-data approaches in artificial intelligence might be expected to differ fundamentally from whatever is happening in us. FaceNet and DeepFace had supervised training on millions of images; ordinary humans will tend to see fewer photos of each known person than DeepFace, fewer total people than FaceNet, and only rarely have people's names stipulated when they are seen. Artificial neural networks start with blank-slate parameters, but humans might start with some naturally selected genetic tuning to detect faces. Lastly, the superhuman performance of recent models might look like another sign that their learning is unlike ours. However, a closer look at the differences here will show that they are not as deep as they might initially seem, while simultaneously showing us that lessons learned from face recognition apply more broadly to other kinds of human knowledge.

It is certainly not typical for us to learn to recognize someone by looking at a thousand labeled photos of them. However, we do not have to meet someone a thousand times to have equivalent exposure. In each live meeting (or viewing of a video clip) we see a given face moving through multiple expressions, often from multiple viewpoints, with the fact that these varied presentations are of a single individual given to us by the context. As Cameron Buckner has pointed out, we benefit not only from the provision of live motion over static images (he estimates the relevant human visual frame rate at ten to twelve images per second), but also from our capacity to mull over images in memory, dreams and daydreaming.[52] Meanwhile, in attaching an identity to a percept, we do experience some supervised learning. New personal acquaintances often introduce themselves by name on first encounter, or are introduced by knowledgeable others, and we have ongoing practice with the name in conversational interactions. Celebrity images are often broadcast in contexts where they are named, for example in news reports or in film credits. More subtly, we are kept on track by unsupervised learning (also known as "self-supervised" learning), in which we generate predictions about identity over time and space, using faces as a marker of identity, and receive corrective prediction error feedback from the actual course of our sensations. As you scan the face of someone turning towards you, previous acquaintance generates expectations which may be satisfied by recognition, or create the learning opportunity of evident misrecognition. After you have met him, you do not need to know your waiter's name to have perceptual expectations as he turns back to face your table, and from a different angle you can now see what you know from context to be the same face, in a new light. Even without explicit labeling, we can learn the faces of others through the ongoing practice of anchoring multiple views to a single identity in social interaction.

We might also wonder about the human starting point in face recognition. Progress in deep learning has in some way made this worry more vivid. Earlier approaches to identifying the significant dimensions of facial variation could not secure human-level performance even by positing 100,000 such dimensions. Newer models are much more compact: when judging an identity, DeepFace ultimately relies on no more than 4,096 dimensions, and in practice roughly 1,000. FaceNet boils things down to a 128-dimension vector, which can be quantized to one byte per dimension by the end of training, without loss of accuracy.[53] These 128 dimensions are not characteristics directly describable in natural language (they are more like high-level geometrical abstractions than say, which of eight major nose shapes or eye colours someone has); yet somehow, the set of these specifications is enough to individuate human faces as well as humans can. The compactness of this representation might make us wonder whether humans could have some innate ability to discern such a set of dimensions, naturally selected for its social value. When it seemed that more than 100,000 pre-engineered dimensions were needed to specify a face, it was harder to see how our genome could have encoded the corresponding set of dimensions in our face recognition area; now that it has been demonstrated that the task of discerning faces can be executed more economically, the question has been reinvigorated.

This question is part of a larger set of questions about the extent to which the human capacity to discern faces is shaped by experience as opposed to being genetically hardwired. The primate cortex has some areas that are selectively activated by the sight of faces, particularly in the inferotemporal cortex (IT). Face recognition is just one specialization within the IT; other areas with corresponding characteristic locations across primates include areas responsive to places, buildings, tools, hands, and bodies; literate humans also have an area for text (the "Visual Word Form Area"), activated by written words, but also by braille, and by fingerspelling for those who know sign language. Evidence for these localizations comes from fMRI activation and from studies of selective impairments caused by lesions in the brain.[54]

The fact that there is cortical specialization for face recognition in a similar location in humans and other primates might suggest that this must be genetically specified; however, advocates of genetic explanations need to explain the existence of similar stereotyped areas for buildings, tools, and text, which emerged too recently in our history to have had an inherited impact. One recent review examines the options here.[55] According to the "neuronal recycling" theory, primates have genetically determined zones to recognize naturally important categories, and novel categories such as printed text settle opportunistically into whatever natural zone is visually closest, perhaps branching structures in the case of text.[56] Other theories contend that local specialization is driven by the timeline of development, with faces ending up where they do because of their prevalence in the early experience of infants,[57] or that objects end up in locations appropriate to the degree of complexity they present.[58] A final class of theory suggests that the ratio between curvature and rectilinearity is crucial in localization.[59]

Hypotheses about the importance of developmental timing and visual shape have been tested with experiments on groups of monkeys, who were exposed to different textual symbol systems (a blocky set and an alphanumeric set) at different ages. The monkeys did develop distinct and localized visual sensitivity to these unnatural symbol systems, and these localizations were driven by stimulus shape, rather than the age of exposure.[60] Evidence for an underlying shape-based architecture also comes from the observation that face domains in IT are somewhat activated by similarly shaped non-face objects, such as apples and clock faces.[61]

A deeper understanding of the relationship between face specialization and experience came from a study in which three macaques were deprived of early face experience, initially hand-raised for several months by human carers wearing welders' masks, in an otherwise stimulating and visually rich environment.[62] For ordinary control macaques, the face domain is naturally formed by five to six months of age, at which point there is strong, consistent activation for faces as opposed to other objects, localized in a stereotyped patch of the superior temporal sulcus (STS). However, face-deprived monkeys did not show the typical pattern when presented with images of faces at six months; there was very weak or absent activation in those typical face areas (and elsewhere) for face stimuli, although there was ordinary specialization for the perception of scenes and bodies, and indeed enhanced recognition of hands. Even more strikingly, the face-deprived monkeys did not show preferential looking towards faces in images. Preferential looking actually precedes the formation of the face area in normal monkeys: as early as ninety days, normal monkeys will look preferentially at faces, but the face-deprived monkeys displayed no signs of this. Arcaro and colleagues concluded that even preferential looking towards faces is learned rather than innate. "We propose that IT domains develop as follows: very early in development, reinforcement affects how infants interact with their environment, including what they look at."[63]

Assuming that there is some reward for detecting faces, sensitivity for faces is not genetically predetermined beyond the proto-organization for shape: it is driven and shaped by the set of faces that you see, or to quote the subtitle of Arcaro's 2019 review, "what you see is what you get." So, rather than being the product of some idiosyncratic innate module, face recognition works like the recognition of anything else: "the development of face processing is guided by the same ubiquitous rules that guide the development of cortex in general."[64] The basic statistical learning engine of the cortex works to make whatever discriminations will generate reward, developing whatever categorizations of objects and events can prove their value in the animal's pursuit of its ends.

The superhuman performance of models like FaceNet marks another possible point of contrast with human face recognition, but this performance is no doubt raised above ours by its exposure to eight million identities in its training data. Restrictions in large data work to create human-like limitations: for example, in a version of the own-ethnicity-bias, models show poorer performance in verification tasks on ethnicities that were under-represented in their training data.[65] It has been proposed that models intensively trained on about five thousand identities should show human-like performance;[66] as far as their basic structure is concerned, systems like FaceNet are considered to be "a promising model of neurally-inspired face representations in high-level visual cortex."[67]

4. OVER-PARAMETERIZATION, OVERFITTING, AND DIRECT FIT

One thing that we have in common with systems like DeepFace and FaceNet is a wealth of parameters: they have 120–140 million, and we doubtless have more, given overall estimates of the number of neurons in our brains.[68] This raises a puzzle, most obviously in the cases of DeepFace and FaceNet, whose parameters and data can easily be counted: they are clearly "over-parameterized," in the sense of having more parameters than training data points. Because these models could simply use their vast resources to memorize the labels for their training data, there is a mystery about how they succeed in generalizing so well.[69]

Classical methods impose considerable restraint in adding parameters to a model. A first-order polynomial (two parameters) will draw a straight (perhaps sloped) line through a noisy cloud of data; a second-order polynomial (three parameters) will draw a parabolic curve, which might in some cases come closer to capturing the shape of the cloud. We can keep adding parameters, raising the degree of the linear function until it oscillates as much as needed to come arbitrarily close to each of the given data points; however, if there is noise in this initial data, the high-degree polynomial that laces its way through all those noisy data points will have very weak predictive power both between and beyond these points. The complex theory represented by our high-degree polynomial will also be unstable, requiring radical revision each time a new data point is added (with its measure of random noise needing to be captured by the addition of further parameters, plus extensive adjustments to the coefficients of existing parameters). To avoid such "explosive overfitting," traditional statistical methods restrict the number of parameters to make the model less sensitive to peculiarities such as noise in the data: optimally, traditional models have just enough parameters to ensure that their expressive power matches the complexity of the phenomenon being modelled. With too few parameters, the model will underfit the data, missing the subtlety of the phenomenon (this is the problem of 'bias'); with too many parameters, the model will overfit, sidetracked by irrelevant characteristics of the given data in a way that will make it a poor representation of the underlying phenomenon, and a poor predictive guide to future data (this is the problem of "variance"). It was traditionally thought that a trade-off between bias and variance was inevitable, and best balanced by adding parameters to reduce bias just until the point where variance starts to rise ("ideal fit"). The mystery is how massively over-parameterized models avoid problematic overfitting, especially given that they are in practice trained to achieve virtually zero error on their training data.

The answer to this puzzle seems to be that the trade-off between bias and variance holds only up to a point, and something interesting happens on the far side of that point, when very complex models meet very large data.[70] Assuming a fixed set of data and some set of dimensions the learning model must uncover, expanding the complexity of the model by adding parameters will at first improve its performance (as the model becomes complex enough to express the regularities in those dimensions), and then worsen it (as the model becomes too complex, overfitting noise). However, if we continue to increase the complexity of the model enormously, past the point at which its parameters can perfectly fit all the training data (a point known as the "interpolation threshold"), then *both* bias and variance can sink, and then keep sinking: "increasing model complexity past the interpolation threshold can actually result in an *increase* in model performance without succumbing to overfitting."[71]

On the far side of the interpolation threshold, adding parameters is no longer improving the model's fidelity to its training data points, as these can already be captured perfectly. Instead, in an appropriately designed model, additional parameters enable better sampling of the dimensions of the phenomenon that the model is intended to capture. This big-data, big-model strategy has no special value where that number of dimensions is small because the underlying phenomenon is simple. Wherever the observational data is generated by some natural phenomenon which actually has the simplicity of a quadratic equation, the "ideal fit" three-parameter model will be perfect, and big data will not make things better. However, when the underlying phenomenon is very complex, and we have data rich enough to reflect this complexity, over-parameterization becomes attractive.

Uri Hasson and colleagues call the iterative optimization of neural networks a process of "direct fit" (as opposed to ideal fit). In their terms, an ideal-fit model "learns the underlying generative or global structure of the data by exposing a few latent factors or rules"; such a model has the power to extrapolate from a narrow region of data points to the rest of reality.[72] In this kind of modeling, a simple equation (such as $e=mc^2$) can be found to hold very broadly on the basis of restricted local observations. By contrast, a direct-fit model "uses local computations to situate novel observations within the context of past observations; it does not rely on explicit modeling of the over-arching generative principles."[73] Direct-fit models interpolate well between their training data points by means of generic low-level methods such as averaging and nearest-neighbor computations; while they can capture extremely complex patterns within this sampled area, they do not extrapolate meaningfully beyond it.

An ideal-fit model for face recognition would need to incorporate all the laws giving rise to perceptible human facial variation, including laws of genetics, epigenetics, physiology, optics, even sociological rules about facial expressions and cosmetics. To complicate matters further, faces can be nearly replicated in twin formation, or scarred in hard-to-predict accidents. If some ideal-fit model is even possible, from a God's-eye perspective, it would seem to demand a God-like computational capacity: to generate a formula predictive of human facial variation, it seems one would have to simulate a massive array of causes that interact in a nonlinear fashion. A direct-fit model for face recognition, by contrast, can be shaped simply by encountering multiple faces with enough variation to make them serviceable guides to the rewarding goal of differentiating individuals. Generative ideal-fit models work by identifying the laws behind a small patch of data, and extrapolating them to unseen data; by contrast, direct-fit models work by densely sampling a larger range of data and interpolating well within that field, despite not extrapolating well outside it.[74] Where ideal-fit models sample a small zone of data and work well to predict simple phenomena that occur uniformly both within that zone and beyond it, direct-fit models are like sheets of fabric spread over a wide range of data points in some immensely convoluted landscape, pulled tighter to that complex topology by the optimization of their excess parameters.

The adaptive character of direct-fit models makes them sensitive to local complexities. To distinguish two people, I do not need a consciously available list of characteristics separating possible percepts of them; I just have to have an adequate level of experience of their faces, under conditions where I independently know who is who. What will count as an adequate level of experience with a face is not settled by its intrinsic character alone, but also by the context of other available faces: for example, if a person has a twin, one will typically need to have much more extensive experience with the pair of them before one can tell safely them apart. Those familiar with twins can have heightened sensitivity to their small facial asymmetries and skin markings without needing to apply that particular sensitivity to every other face they encounter; driven by local calculations, direct-fit models enable tailored strategies for different identities.

The local flexibility of direct-fit models is attractive, but it should be underscored that it does not constitute an advance guarantee of their success in any particular case. Given limits on our vision and processing, some very similar faces may remain indistinguishable for some of us, no matter how much data we gain. If our face recognition works along the lines of a Voronoi tessellation of face space, it is a tessellation in which some anchor points may effectively coincide. In addition, new anchor points may need to be added to the map as we go along, but again, thanks to the local character of direct-fit models, the need for these new anchor points does not compromise the safety of our identification of individuals further away in face space.

Despite occasional encounters with indistinguishable faces, the learning rule behind human face recognition seems to presuppose that for each new person we encounter, there is some many-to-one mapping defining a zone of percepts that can come only from this individual. Learning what Idris Elba looks like, or coming to be able to recognize his face, going forward, is not the same as matching a single picture of him to the label; it is mapping this distinctive larger zone of possibilities for him. The very idea that there is such a distinctive zone assumes that there is a *type* of percept that can only come from the face of Idris Elba. Classical approaches to face recognition assumed that we needed to start by defining some set of dimensions adequate to support the construction of a set of these types, one for each individual we might encounter. Direct-fit approaches work by first presuming that there is a type unique to each individual, and using this presumption to drive the development and use of a set of dimensions powerful enough to map the contours of this type with the given data. Even if in some cases (truly indistinguishable twins) the presumption fails, as long as it is generally successful enough, it will continue to be applied. In a learning context where Idris Elba is independently labeled, it may not matter whether we can recognize him on the basis of his facial characteristics; however, in more interesting test contexts, Idris Elba will not be independently labeled, and it is exactly because we keep facing such test contexts that we develop the capacity for face recognition. Given the importance of distinguishing individuals, together with the facts about actual human facial distinctiveness, it pays off over time to identify individuals this way, and as we discover through trial and error that we are largely able to do so, this reinforces our tendency to seek to differentiate people according to the rule that there is a type of percept that can only come from the face of each individual. Because we learn from prediction error, the fact that we sometimes get it wrong drives learning of the specifics of faces and the dimensions individuating them, rather than dissuading us from continuing to apply the rule.

This system incorporates some strong presuppositions: human face recognition would collapse in a world with chaotic faces, suddenly rampant use of latex masks, or genetic changes leading to a population of faces whose individual differences were all too subtle for us to perceive. Ultimately, it is a somewhat contingent fit between characteristics of human faces and human cognition that enables the kind of safely individuating judgements of which humans are generally capable. Making a safe judgment of identity on the basis of some percepts is a matter of being sufficiently well-adapted to local faces that these percepts fall within the zone uniquely picking out this identity, but again, because safety is a local matter, it is not threatened by the merely hypothetical possibility that there might fail to be such a zone for this individual, or indeed for any individual, under imaginative extrapolation to different circumstances.

5. RECOGNIZING KNOWLEDGE

There is a parallel between the recognition of a face and the recognition of knowledge. Knowledge can be uniquely characterized as the most general factive mental state; on this view, to know is to be in the *type* of mental state whose distinctive character is that it can only be held towards truths.[75] Learning the distinctive contours of this type gives us predictive power, going forward: with sufficient experience, we can encounter a new case of this type and recognize the subject as knowing. In a learning context in which it is independently stipulated that the proposition *p* that the subject judges to be the case is in fact true, there may be no exciting practical difference (with respect to *p*) in recognizing that the subject knows that *p*, as opposed to simply happening to be right that *p*. However, in more interesting test contexts, the key proposition's truth value will not be independently supplied to us, and it is exactly because we keep facing these test contexts that we develop the capacity for knowledge recognition: it pays off over time to recognize situations in which people have a state of mind of a type that one can only have to truths, and as we discover through trial and error that we are largely able to do so, this reinforces our tendency to seek out signs of these truth-anchored mental states in each other. Crucially, we can detect not only knowledge *that p* is the case, where *p* is independently known to be true, but we can also detect knowledge-*wh*. So, for example, from where you are now seated, you can see me as knowing whether this coin I am looking at is heads or tails, because I instantiate a familiar pattern of perceptual access. This is a clear and central enough case that you do not need to know the precise visual input I am receiving to make this determination; you can see me as close enough that small variations in that input will not disrupt my lock on the truth.

As the coin example illustrates, the capacity to differentiate patterns of knowledge and ignorance in our fellow agents enables us to exploit their epistemic access to those parts of reality for which their vantage point is better than ours. If you want to know which way the coin in my palm is facing, you know you can ask me. While many primates show selective social learning from peers recognized as knowledgeable, humans show exceptionally active use of the knowledge of their peers,[76] guided by an exceptionally well-developed sense of what others do and do not know, a sense informed by continual feedback from conversational exchanges[77] and extra-conversational encounters with reality.

This system incorporates some strong presuppositions. Our systematic tendency to attribute knowledge presupposes that there really is a type of state of mind that subjects can only have only to truths. This presupposition would collapse in a world with too much variation either in objective reality or in the cognitive capacities of the subjects who surround us; as it is, the objective and cognitive regularities in our world enable us to have, and subsequently to detect in each other, patterns of knowledge, or successful cognitive adaptation to reality. We can be mistaken about whether someone is successfully adapted on a given point, just as we can be mistaken in identifying someone who turns out to have a secret twin. But because we learn from prediction error, the fact that we sometimes get knowledge attributions wrong drives learning of the specifics of types of interactions between agents and our shared environment, and the dimensions separating types of ignorance from ways of knowing, rather than dissuading us from continuing to apply the rule that there is some type of state subjects can have only to truths.

Our capacity for face recognition is not introspectable because face recognition is a high-dimensional problem that we solve through intuitive visual processing rather than reportable operations on consciously available contents. Face recognition is a subtle capacity, mastered for creatures like us only by grappling with an appropriately massive body of data. But face recognition itself is a type of knowledge. As a consequence, the problem of detecting knowledge inherits the complexity of the problem of detecting faces. Meanwhile, face recognition is just one of a great variety of human ways of knowing: we have all kinds of other perceptual and inferential capacities. We should not expect the contours of this zone to be discoverable in advance through any simple set of rules or hand-engineered features that we could state in a reductive analysis. At the same time, there is an important similarity between examining multiple cases of knowledge and examining multiple views of the same face: something about the actual shape of the target domain becomes clearer as we see many instances of it.

I want to close on the question of what we can still learn from tricky hypothetical cases. I believe there is a difference between extrapolation and interpolation here: if my understanding of knowledge recognition is correct, then the most helpful cases are importantly realistic, like Plato's case of Theodorus mistaken for Theaetetus at a distance,[78] or Sosa's case of his knowledge of his own marital status.[79] If cases start to get extrapolative, lying far outside the zone within which our natural understanding of knowledge is anchored, then I am far from certain that they have the same instructive value. However, the boundary of that zone is itself something we may only be able to learn from trial and error, an ongoing process in which even the most speculative epistemology has much to contribute.

NOTES

1. Plato, *Meno*, 98b.
2. Machery et al., "Gettier Across Cultures"; Nagel, San Juan, and Mar, "Lay Denial of Knowledge for Justified True Beliefs."
3. Horschler, Santos, and MacLean, "Do Non-human Primates Really Represent Others' Ignorance?"; Kaminski, Call, and Tomasello, "Chimpanzees Know What Others Know, But Not What They Believe."
4. Plato, *Meno*, 97ab.
5. Dharmottara, c. 770CE, as reported in Dreyfus, *Recognizing Reality*, 292.
6. Sosa, "How Must Knowledge Be Modally Related to What Is Known?" 381.

7. E.g., Greco, "Safety, Explanation, Iteration"; Pritchard, *Epistemic Luck*; Williamson, *Knowledge and Its Limits*.

8. Comesaña, "Unsafe Knowledge"; Neta and Rohrbaugh, "Luminosity and the Safety of Knowledge"; Zhao, "Knowledge Without Safety."

9. Sosa, *A Virtue Epistemology*; Sosa, *Reflective Knowledge*.

10. Sosa, *Judgment and Agency*; Sosa, *Epistemic Explanations*.

11. E.g., Lackey, "What Luck Is Not."

12. Kelp, "Knowledge and Safety."

13. Grundmann, "Saving Safety from Counterexamples."

14. Sosa, Judgment and Agency; Sosa, Epistemic Explanations.

15. Hirvelä and Paterson, "Need Knowing and Acting Be SSS-Safe?"

16. Jenkins, Dowsett, and Burton, "How Many Faces Do People Know?"

17. Barragan-Jason, Cauchoix, and Barbeau, "Neural Speed."

18. Tanaka, "The Entry Point of Face Recognition."

19. Average competence seems to admit of fairly wide variance: the study estimating average recognition of 5,000 faces aimed just to come within the right order of magnitude, finding a standard deviation of about 2,000 around that mean, but with their twenty-five participants all falling within the 1,000–10,000 range (Jenkins et al., "How Many Faces Do People Know?"). In what follows, I focus on neurotypical performance, setting aside both congenital and acquired prosopagnosia (face-blindness); however, section 3 will argue that the general points about safety do not rely on any peculiarity of face recognition. Similar pattern-detecting capacities such as voice identification (generally unimpaired in those with prosopagnosia: Liu et al., "The Processing of Voice Identity in Developmental Prosopagnosia") have the same basic structure as face recognition, so the epistemic lessons here apply broadly.

20. Lucas and Henneberg, "Are Human Faces Unique?"

21. Jenkins et al., "Variability in Photos of the Same Face," 321.

22. Jenkins et al., "Variability in Photos of the Same Face," 315.

23. Jenkins et al., "Variability in Photos of the Same Face," Experiment 2.

24. Burton et al., "Face Recognition in Poor-Quality Video."

25. Megreya and Burton, "Matching Faces to Photographs."

26. White et al., "Passport Officers' Errors in Face Matching."

27. Young and Burton, "Are We Face Experts?" 106.

28. Malpass and Kravitz, "Recognition for Faces of Own and Other Race."

29. Chiroro and Valentine, "An Investigation of the Contact Hypothesis."

30. Rhodes and Anastasi, "The Own-Age Bias in Face Recognition"; Wiese, Komes, and Schweinberger, "Ageing Faces in Ageing Minds."

31. Sporer, "Recognizing Faces of Other Ethnic Groups."

32. Valentine, "A Unified Account."

33. Lewis and Johnston, "Effects of Caricaturing Faces."

34. Bruce and Young, "Understanding Face Recognition," 308.

35. Bruce and Young, "Understanding Face Recognition," 311.

36. Chen et al., "Blessing of Dimensionality."

37. Taigman et al., "Deepface."

38. Krizhevsky, Sutskever, and Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," 88.

39. Intended just to be illustrative, the example is unrealistic because even without applying an explicit regularization technique such as dropout, a model with DeepFace's architecture would not end up overfitting so crudely when trained on a dataset with the rich structure of a stack of human face photos, not least because stochastic gradient descent itself works as a form of implicit regularization (on this point, see Ma, Bassily, and Belkin, "The Power of Interpolation").

40. Reliance on genuine patterns in the data depends on the actual availability of those patterns. Even with explicit regularization methods such as dropout, it is still possible for a model the size of DeepFace to simply memorize its input, having vastly more parameters than data points in its training set: whether a model actually works by identifying generalizable features of its data depends on the relationship between the level of structure in the data, the architecture, and the learning algorithm of the model (Zhang et al., "Understanding Deep Learning (Still) Requires Rethinking Generalization").

41. Kukačka, Golkov, and Cremers, "Regularization for Deep Learning."

42. The question of what constitutes a small change in input is a complex one. Changes that appear small to humans (to the point of being invisible to the naked eye) may be large for image classification models, for example because of their heavier reliance on texture rather than overall shape in classifications. Together with label noise, this disparity drives the problem of adversarial examples (Goodfellow, Shlens, and Szegedy, "Explaining and Harnessing Adversarial Examples"), a problem beyond the scope of the current paper. For arguments that adversarial examples may reflect epistemic strengths as well as weaknesses, and do not necessarily point to a fundamental difference between human and machine learning, see (Buckner, "Black Boxes or Unflattering Mirrors?"; Zhou and Firestone "Humans Can Decipher Adversarial Images"; Ilyas et al., "Adversarial Examples Are Not Bugs, They Are Features").

43. This version will remove the final classification layer with the specific identities of the training set, but retain the core layers of the model that extract whatever characteristics successfully differentiated those originally trained identities.

44. Taigman et al., "Deepface," Table 4.

45. Taigman et al., "Deepface," section 5.3.

46. Huang et al., "Labeled Faces in the Wild."

47. Schroff, Kalenichenko, and Philbin, "Facenet."

48. Schroff et al., "Facenet," 817.

49. Schroff et al., "Facenet," 822.

50. For a visualization of this effect, see O'Toole et al., "Face Space Representations in Deep Convolutional Neural Networks," Figure 3.

51. Douven and Gärdenfors, "What Are Natural Concepts?" 320.

52. Buckner, "Black Boxes or Unflattering Mirrors?"

53. Schroff et al., "Facenet," 821.

54. Reviewed in Arcaro, Schade, and Livingstone, "Universal Mechanisms and the Development of the Face Network."

55. Arcaro et al., "Universal Mechanisms and the Development of the Face Network."

56. Dehaene and Cohen, "Cultural Recycling of Cortical Maps."

57. Quartz and Sejnowski, "The Neural Basis of Cognitive Development."

58. Gauthier and Palmeri, "Visual Neurons."

59. Nasr, Echavarria, and Tootell, "Thinking Outside the Box."

60. Srihasam et al., "Behavioral and Anatomical Consequences"; Srihasam, Vincent, and Livingstone, "Novel Domain Formation."

61. Tsao et al., "A Cortical Region Consisting Entirely of Face-Selective Cells." This shape-based explanation on its own does not cover all forms of specialization within IT; specifically, it does not cover the shared localization of visual text, Braille and fingerspelling for the Deaf. However, Arcaro and colleagues argue that the shape-based (or scale and curvature-driven) localization of visual objects is itself an instance of a more general principle about computational requirements determining localization. In their view, the key commonality among the various forms of text recognition that are handled in the Visual Word Form Area is that they require rapid and repetitive fine-grained computation, so that ultimately, "these results on cross-modal plasticity indicate commonalities in general computational requirements rather than commonalities in what the computations are used for" ("Universal Mechanisms," 364).

62. Arcaro et al., "Seeing Faces Is Necessary for Face-Domain Formation."

63. Arcaro et al., "Seeing Faces Is Necessary for Face-Domain Formation," 1411.

64. Arcaro et al., "Universal Mechanisms," 341.

65. Wang et al., "Racial Faces in the Wild."

66. Hasson, Nastase, and Goldstein, Direct Fit to Nature," 421.

67. O'Toole et al., "Face Space Representations," 807.

68. The width of those estimates and the varying density of neurons and synapses in different brain regions make it hard to estimate the number of parameters devoted to face recognition with any precision (Azevedo et al., "Equal Numbers of Neuronal and Nonneuronal Cells Make the Human Brain an Isometrically Scaled-up Primate Brain"; Lent et al., "How Many Neurons Do You Have").

69. It has been demonstrated that similarly sized models can simply memorize similarly sized training data, despite having explicit regularization measures: Zhang and colleagues ("Understanding Deep Learning (Still) Requires Rethinking Generalization") were able to train a generally successful image recognition model using completely random image labels, reducing its training error to zero. Trained this way, the model of course cannot generalize beyond its training set; however, when trained with the same set of images with correct labels, the very same model architecture generalizes well.

70. Belkin, Hsu, and Mitra, "Overfitting or Perfect Fitting?"; Loog et al., "A Brief Prehistory of Double Descent."

71. Rocks and Mehta, "Memorizing without Overfitting," 14.

72. Hasson et al., "Direct Fit to Nature," 418.

73. Hasson et al., “Direct Fit to Nature,” 418–19.

74. Hasson et al., “Direct Fit to Nature,” Figure 1.

75. Williamson, *Knowledge and Its Limits*, ch. 1.

76. Tomasello, *Becoming Human*.

77. Westra and Nagel, “Mindreading in Conversation.”

78. Plato, *The Theaetetus of Plato*, 193b.

79. Sosa, “The Analysis of ‘Knowledge that p’,” 3.

REFERENCES

Arcaro, M. J., P. F. Schade, and M. S. Livingstone. “Universal Mechanisms and the Development of the Face Network: What You See Is What You Get.” *Annual Review of Vision Science* 5 (2019): 341–72.

Arcaro, M. J., P. F. Schade, J. L. Vincent, C. R. Ponce, and M. S. Livingstone. “Seeing Faces Is Necessary for Face-Domain Formation.” *Nature Neuroscience* 20, no. 10 (2017): 1404–12.

Azevedo, F. A., L. R. Carvalho, L. T. Grinberg, J. M. Farfel, R. E. L. Ferretti, R. E. P. Leite, W. J. Filho, R. Lent, and S. Herculano-Houzel. “Equal Numbers of Neuronal and Nonneuronal Cells Make the Human Brain an Isometrically Scaled-up Primate Brain.” *Journal of Comparative Neurology* 513, no. 5 (2009): 532–41.

Barragan-Jason, G., M. Cauchoix, and E. Barbeau. “The Neural Speed of Familiar Face Recognition.” *Neuropsychologia* 75 (2015): 390–401.

Belkin, M., D. J. Hsu, and P. Mitra. “Overfitting or Perfect Fitting? Risk Bounds for Classification and Regression Rules that Interpolate.” *Advances in Neural Information Processing Systems* 31 (2018).

Bruce, V., and A. Young. “Understanding Face Recognition.” *British Journal of Psychology* 77, no. 3 (1986): 305–27.

Buckner, C. “Black Boxes or Unflattering Mirrors? Comparative Bias in the Science of Machine Behaviour.” *British Journal for the Philosophy of Science*, forthcoming.

Burton, A. M., S. Wilson, M. Cowan, and V. Bruce. “Face Recognition in Poor-Quality Video: Evidence from Security Surveillance.” *Psychological Science* 10, no. 3 (1999): 243–48.

Chen, D., X. Cao, F. Wen, and J. Sun. “Blessing of Dimensionality: High-Dimensional Feature and Its Efficient Compression for Face Verification.” Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2013.

Chiroro, P., and T. Valentine. “An Investigation of the Contact Hypothesis of the Own-Race Bias in Face Recognition.” *The Quarterly Journal of Experimental Psychology* 48, no. 4 (1995): 879–94.

Comesaña, J. “Unsafe Knowledge.” *Synthese* 146 (2005): 395–404.

Dehaene, S., and L. Cohen. “Cultural Recycling of Cortical Maps.” *Neuron* 56, no. 2 (2007): 384–98.

Douven, I., and P. Gärdenfors. “What Are Natural Concepts? A Design Perspective.” *Mind and Language* 35, no. 3 (2020): 313–34.

Dreyfus, G. B. *Recognizing Reality: Dharmakirti's Philosophy and Its Tibetan Interpretations*. Albany: Suny Press, 1997.

Gauthier, I., and T. J. Palmeri. "Visual Neurons: Categorization-based Selectivity." *Current Biology* 12, no. 8 (2002): R282–R284.

Goodfellow, I. J., J. Shlens, and C. Szegedy. "Explaining and Harnessing Adversarial Examples." *arXiv preprint arXiv:1412.6572* (2014).

Greco, D. "Safety, Explanation, Iteration." *Philosophical Issues* 26, no. 1 (2016): 187–208.

Grundmann, T. "Saving Safety from Counterexamples." *Synthese* 197, no. 12 (2020): 5161–85.

Hasson, U., S. A. Nastase, and A. Goldstein. "Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks." *Neuron* 105, no. 3 (2020): 416–34.

Hirvelä, J., and N. Paterson. "Need Knowing and Acting Be SSS-Safe?" *Thought: A Journal of Philosophy* 10, no. 2 (2021): 127–34.

Horschler, D. J., L. R. Santos, and E. L. MacLean. "Do Non-human Primates Really Represent Others' Ignorance? A Test of the Awareness Rlations Hypothesis." *Cognition* 190 (2019): 72–80.

Huang, G. B., M. Mattar, T. Berg, and E. Learned-Miller. "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, 2008.

Ilyas, A., S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. "Adversarial Examples Are Not Bugs, They Are Features." *Advances in Neural Information Processing Systems* 32 (2019).

Jenkins, R., A. Dowsett, and A. Burton. "How Many Faces Do People Know?" *Proceedings of the Royal Society B* 285, no. 1888 (2018): 20181319.

Jenkins, R., D. White, X. Van Montfort, and A. M. Burton. "Variability in Photos of the Same Face." *Cognition* 121, no. 3 (2011): 313–23.

Kaminski, J., J. Call, and M. Tomasello. "Chimpanzees Know What Others Know, But Not What They Believe." *Cognition* 109, no. 2 (2008): 224–34.

Kelp, C. "Knowledge and Safety." *Journal of Philosophical Research* 34 (2009): 21–31.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. "Imagenet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems* 25 (2012): 1097–1105.

Kukačka, J., V. Golkov, and D. Cremers. "Regularization for Deep Learning: A Taxonomy." *arXiv preprint arXiv:1710.10686* (2017).

Lackey, J. "What Luck Is Not." *Australasian Journal of Philosophy* 86, no. 2 (2008): 255–67.

Lent, R., F. A. Azevedo, C. H. Andrade-Moraes, and A. V. Pinto. "How Many Neurons Do You Have? Some Dogmas of Quantitative Neuroscience Under Revision." *European Journal of Neuroscience* 35, no. 1 (2012): 1–9.

Lewis, M. B., and R. A. Johnston. "A Unified Account of the Effects of Caricaturing Faces." *Visual Cognition* 6, no. 1 (1999): 1–41.

Liu, R. R., S. L. Corrow, R. Pancaroglu, B. Duchaine, and J. J. Barton. "The Processing of Voice Identity in Developmental Prosopagnosia." *Cortex* 71 (2015): 390–97.

Loog, M., T. Viering, A. Mey, J. H. Krijthe, and D. M. Tax. "A Brief Prehistory of Double Descent." *Proceedings of the National Academy of Sciences* 117, no. 20 (2020): 10625–26.

Lucas, T., and M. Henneberg. "Are Human Faces Unique? A Metric Approach to Finding Single Individuals without Duplicates in Large Samples." *Forensic Science International* 257 (2015): 514.e1–514.e6.

Ma, S., R. Bassily, and M. Belkin. "The Power of Interpolation: Understanding the Effectiveness of SGD in Modern Over-Parametrized Learning." Paper presented at the International Conference on Machine Learning, 2018.

Machery, E., S. Stich, D. Rose, A. Chatterjee, K. Karasawa, N. Struchiner, S. Sirker, N. Usui, and T. Hashimoto. "Gettier Across Cultures." *Nous* (2015).

Malpass, R. S., and J. Kravitz. "Recognition for Faces of Own and Other Race." *Journal of Personality and Social Psychology* 13, no. 4 (1969): 330.

Megreya, A. M., and A. M. Burton. "Matching Faces to Photographs: Poor Performance in Eyewitness Memory (without the Memory)." *Journal of Experimental Psychology: Applied* 14, no. 4 (2008): 364.

Nagel, J., V. San Juan, and R. Mar. "Lay Denial of Knowledge for Justified True Beliefs." *Cognition* 129, no. 3 (2013): 652–61.

Nasr, S., C. E. Echavarria, and R. B. Tootell. "Thinking Outside the Box: Rectilinear Shapes Selectively Activate Scene-Selective Cortex." *Journal of Neuroscience* 34, no. 20 (2014): 6721–35.

Neta, R., and G. Rohrbaugh. "Luminosity and the Safety of Knowledge." *Pacific Philosophical Quarterly* 85, no. 4 (2004): 396–406.

O'Toole, A. J., C. D. Castillo, C. J. Parde, M. Q. Hill, and R. Chellappa. "Face Space Representations in Deep Convolutional Neural Networks." *Trends in Cognitive Sciences* 22, no. 9 (2018): 794–809.

Plato. *Meno*. Translated by B. Jowett. New York: Liberal Arts Press, 1949.

———. *The Theaetetus of Plato*. Translated by M. J. Levett. Indianapolis: Hackett, 1990.

Pritchard, D. *Epistemic Luck*. Oxford: Oxford University Press, 2005.

Quartz, S. R., and T. J. Sejnowski. "The Neural Basis of Cognitive Development: A Constructivist Manifesto." *Behavioral and Brain Sciences* 20, no. 4 (1997): 537–56.

Rhodes, M. G., and J. S. Anastasi. "The Own-Age Bias in Face Recognition: A Meta-Analytic and Theoretical Review." *Psychological Bulletin* 138, no. 1 (2012): 146.

Rocks, J. W., and P. Mehta. "Memorizing without Overfitting: Bias, Variance, and Interpolation in Overparameterized Models." *Physical Review Research* 4, no. 1 (2022): 013201.

Schroff, F., D. Kalenichenko, and J. Philbin. "Facenet: A Unified Embedding for Face Recognition and Clustering." Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.

Sosa, E. "The Analysis of 'Knowledge that p'." *Analysis*, 25, no. 1 (1964): 1–8.

———. "How Must Knowledge Be Modally Related to What Is Known?" *Philosophical Topics* 26, no. 1/2 (1999): 373–84.

———. *A Virtue Epistemology: Apt Belief and Reflective Knowledge, volume I*. Oxford: Oxford University Press, 2007.

———. *Reflective Knowledge: Apt Belief and Reflective Knowledge, volume II*. Oxford: Oxford University Press, 2009.

———. *Judgment and Agency*. New York: Oxford University Press, 2015.

———. *Epistemic Explanations: A Theory of Telic Normativity, and What It Explains*. Oxford: Oxford University Press, 2021.

Sporer, S. L. "Recognizing Faces of Other Ethnic Groups: An Integration of Theories." *Psychology, Public Policy, and Law* 7, no. 1 (2001): 36.

Srihasam, K., J. B. Mandeville, I. A. Morocz, K. J. Sullivan, and M. S. Livingstone. "Behavioral and Anatomical Consequences of Early Versus Late Symbol Training in Macaques." *Neuron* 73, no. 3 (2012): 608–19.

Srihasam, K., J. L. Vincent, and M. S. Livingstone. "Novel Domain Formation Reveals Proto-Architecture in Inferotemporal Cortex." *Nature Neuroscience* 17, no. 12 (2014): 1776–83.

Taigman, Y., M. Yang, M. A. Ranzato, and L. Wolf. "Deepface: Closing the Gap to Human-Level Performance in Face Verification." Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition, 2014.

Tanaka, J. W. "The Entry Point of Face Recognition: Evidence for Face Expertise." *Journal of Experimental Psychology: General* 130, no. 3 (2001): 534.

Tomasello, M. *Becoming Human: A Theory of Ontogeny.* Cambridge: Belknap Press, 2019.

Tsao, D. Y., W. A. Freiwald, R. B. Tootell, and M. S. Livingstone. "A Cortical Region Consisting Entirely of Face-Selective Cells." *Science* 311, no. 5761 (2006): 670–74.

Valentine, T. "A Unified Account of the Effects of Distinctiveness, Inversion, and Race in Face Recognition." *The Quarterly Journal of Experimental Psychology* 43, no. 2 (1991): 161–204.

Wang, M., W. Deng, J. Hu, X. Tao, and Y. Huang. "Racial Faces in the Wild: Reducing Racial Bias by Information Maximization Adaptation Network." Paper presented at the Proceedings of the ieee/cvf international conference on computer vision, 2019.

Westra, E., and J. Nagel. "Mindreading in Conversation." *Cognition* 210 (2021): 1–15.

White, D., R. I. Kemp, R. Jenkins, M. Matheson, and A. M. Burton. "Passport Officers' Errors in Face Matching." *PloS One* 9, no. 8 (2014): e103510.

Wiese, H., J. Komes, and S. R. Schweinberger. "Ageing Faces in Ageing Minds: A Review on the Own-Age Bias in Face Recognition." *Visual Cognition* 21, no. 9-10 (2013): 1337–63.

Williamson, T. *Knowledge and Its Limits*. New York: Oxford University Press, 2000.

Young, A. W., and A. M. Burton. "Are We Face Experts?" *Trends in Cognitive Sciences* 22, no. 2 (2018): 100–10.

Zhang, C., S. Bengio, M. Hardt, B. Recht, and O. Vinyals. "Understanding Deep Learning (Still) Requires Rethinking Generalization." *Communications of the ACM* 64, no. 3 (2021): 107–15.

Zhao, H. "Knowledge Without Safety." *Synthese* 197, no. 8 (2020): 3261–78.

Zhou, Z., and C. Firestone. "Humans Can Decipher Adversarial Images." *Nature Communications* 10, no. 1 (2019): 1334.