Jennifer Nagel, University of Toronto
June 12, 2023

Lecture 5: Knowledge possession and knowledge transmission

Vācaspatimiśra (c. 900CE): "a knowledge source does not deviate from its object. A knowledge source's non-deviation amounts to the fact that there will never be disagreement anywhere, anytime, in any condition, between the nature of the object and what we are taught by the knowledge source." (Nyāya-sūtra, c.200CE/2017, p.15). Williamson (2000, p. 33): Knowledge is "the most general factive mental state."

1. Types of transmission

Testimony is not our only way of transmitting knowledge. Knowledge can be transmitted from one agent to another by manipulation of perceptual attention, without the need for the receiving agent to trust the sender. Demonstrative reasoning can also bypass trust, as illustrated by Augustine's example of the eavesdropper learning that the soul is immortal by recognizing the soundness of the proof that the atheist is reciting, where the atheist is "unaware that he's stating truths" (Augustine, 389 CE/1995, 142).

In testimonial transmission, you need to trust the speaker to take their word at face value. If your trust is irrational, you will not gain knowledge even if the speaker knows (although you will gain the truth). Knowledge gain is possible if you have knowledgeable trust in the speaker. If you know that I have knowledge on the point in question, then your complete way of judging this problem gets you safely to the truth, and my knowledge will be transmitted to you. But how do we know who is knowledgeable on a point of interest?

Sociologists of conversation observe that we treat our conversational partners as if we had a sense of the zones within which their words can be safely taken at face value, their "epistemic territory" in John Heritage's terminology. "While it may be thought that the notion of epistemic territory introduces a contingency of daunting difficulty and complexity into the study of interaction, in fact relative access to particular epistemic domains is treated as a more or less settled matter in the large bulk of ordinary interaction" (Heritage, 2012, p.6).

Today I will argue that epistemic territory is generally mapped by the same domain-general learning mechanisms responsible for other types of intuitive cognition, such as face recognition. I also aim to shed light on the safety condition on knowledge, according to which "you know only if your belief is safe, i.e. it must be that you would so believe only if your belief were true" (Sosa, 1999, 381). Another formulation of safety: "If in case α one knows p on a basis *b,* then in any case close to α in which one believes a proposition p* close to p on a basis close to *b*, p* is true" (Williamson, 2009, 325). Q: what does it mean to "so believe" or to have a basis close enough for knowledge?

2. An overview of face recognition

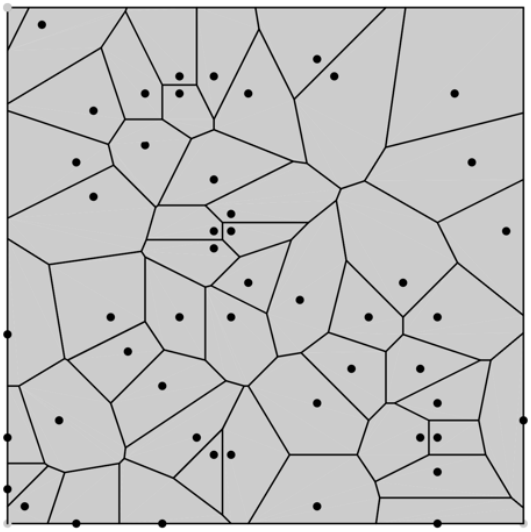
The average adult can recognize the faces of roughly 5000 people (Jenkins, Dowsett, & Burton, 2018). Human faces are broadly similar in their basic configuration, so telling them apart requires subtle sensitivity to small differences. To complicate matters further, photographs or percepts arising from a single individual exhibit large differences arising from changing facial expressions, hairstyles, aging, blemishes, makeup, facial hair, illumination, angle of view, distance, and partial occlusion. We need to explain "not only how we tell people apart, but also how we tell people together" (Jenkins, White, Van Montfort, & Burton, 2011, p.321).

Q: Can you cluster these photos into groups by identity, so that in each group you have photos of the same person?

(Image Source: Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), p.316)

Individualized variation. 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, etc. What makes this hard: "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" (Young & Burton, 2018, p.106).

For any given face, experience with similar faces is helpful. People will do better with ethnicities they have experienced more (Chiroro & Valentine, 1995; Malpass & Kravitz, 1969), assuming they have needed to track individual identities, as opposed to mere group membership (Sporer, 2001). Young children recognize caregiver-aged faces better, but after the age of five, we are more accurate at recognizing the faces of our own age cohort (Rhodes & Anastasi, 2012; Wiese, Komes, & Schweinberger, 2013).



The qualities 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 (Valentine, 1991). We can imagine this as a **Voronoi diagram** (see left), 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 (Lewis & Johnston, 1999).

3. Face recognition in artificial intelligence

Description-driven approaches. Bruce & Young: "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" (Bruce & Young, 1986, p.308). These approaches seemed to demand a very large set of dimensions. Tracking 100,000 sets of relationships among 27 facial landmarks achieves results that still fall well short of human performance (Chen, Cao, Wen, & Sun, 2013).

Deep learning with large data. DeepFace was developed using 4.4 million face images from 4030 people who each had 800-1200 photos of themselves (Taigman, Yang, Ranzato, & Wolf, 2014). This nine-layer model has 120 million parameters (connections between nodes in successive layers), whose values are adjusted in the course of learning, to send a stronger or weaker activation signal to the next layer, on the basis of an error signal at each round of training. Ideally, the model would select the correct identity for each input photo with perfect certainty. It learns by failing to do so: 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. After each step of its supervised learning, the model tunes its parameters in the direction of ideal certainty about the truth. DeepFace approaches human performance levels.

Regularization. DeepFace's capacity to generalize beyond its training data 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 (on each training cycle, each neuron has an independent 0.5 probability of being silenced) (Krizhevsky, Sutskever, & Hinton, 2012). Dropout is one of many regularization methods used to ensure that the model tracks robust features rather than spurious correlations in the training data (Kukačka, Golkov, & Cremers, 2017). For a more nuanced picture of regularization, see (Ma, Bassily, & Belkin, 2018; Zhang, Bengio, Hardt, Recht, & Vinyals, 2021).

Google's FaceNet (Schroff, Kalenichenko, & Philbin, 2015) was trained on dataset of 200 million images drawn from 8 million identities. This model is organized in 22 layers with 140M parameters. The final output layer delivers a compact 128-dimensional embedding, a vector representing the image in "face space". The goal of the model is to refine these output vectors so that they cluster in identities. This model delivers superhuman performance in face verification (scoring 99.63% on LFW).

Figure 3. The **Triplet Loss** minimizes the distance between an *anchor* and a *positive*, both of which have the same identity, and maximizes the distance between the *anchor* and a *negative* of a different identity.

Epistemic safety in FaceNet. 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.

Voronoi tessellations 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 (Douven & Gärdenfors, 2020, 320). There is a payoff in learnability here: 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.

4. The basic structure of human face recognition

How could our face recognition be as data-driven as FaceNet's? We have more than enough parameters, but we don't learn facial identities by looking at 1,000 labeled photos of a person. However, live action affords rich exposure: 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.

What you see is what you get: "the development of face processing is guided by the same ubiquitous rules that guide the development of cortex in general" (Arcaro, Schade, & Livingstone, 2019, p.341). Primate face recognition relies not on innate face-specific mechanisms, but on the same statistical learning processes that govern object recognition across the board.

5. Over-parameterization, overfitting, and direct fit

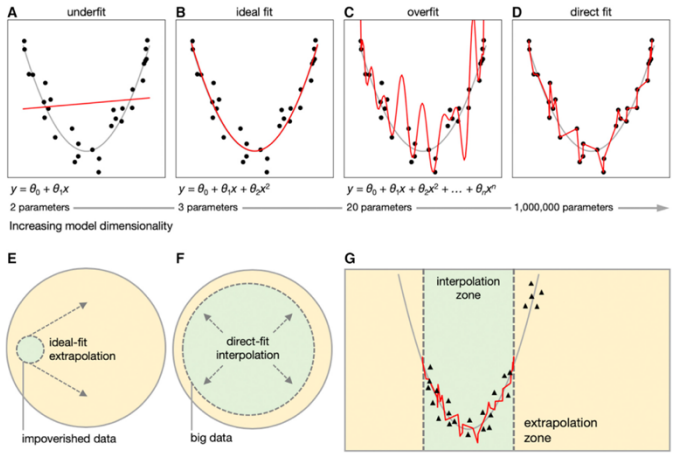


Figure 1. Direct-Fit Learning with Dense Sampling Supports Interpolation-Based Generalization
(A) An overly simplistic model will fail to fit the data.
(B) The ideal-fit model will yield a good fit with few parameters in the context of data relying on a relatively simple generative process; in fact, this is the model used to generate the synthetic data (with noise) shown here.
(C) An overly complex (i.e., over-parameterized) model may fixate on noise and yield an explosive overfit. (A)–(C) capture the "textbook" description of underfitting and overfitting.
(D) Complex models, such as ANNs, however, can nonetheless yield a fit that both captures the training data and generalizes well to novel data within the scope of the training sample (see G and Bansal et al., 2018 for a related discussion).
(E) Traditional experimentalists typically use highly controlled data to construct rule-based, ideal-fit models with the hope that such models will generalize beyond the scope of the training set into the extrapolation zone (real-life data).
(F) Direct-fit models—like ANNs and, we argue, BNNs—rely on dense sampling to generalize using simple interpolation. Dense, exhaustive sampling of real-life events (which the field colloquially refers to as "big data") effectively expands the interpolation zone to mimic idealized extrapolation.
(G) A direct-fit model will generalize well to novel examples (black triangles) in the interpolation zone but will not generalize well in the extrapolation zone.

Image: (Hasson et al., 2020, p.419).

The "direct fit to nature." In the textbook view of model-building, an ideal-fit model "learns the underlying generative or global structure of the data by exposing a few latent factors or rules" (see Figure 1B) (Hasson, Nastase, & Goldstein, 2020, p.418). But what about domains not governed by a few latent factors? Here Hasson and colleagues argue that our best option will be a *direct fit* model in which generic local calculations interpolate between observations (Figure 1D). Given dense data sampling within some zone, these models can represent extremely complex phenomena within that zone with strong fidelity, even while failing to extrapolate well beyond that zone. "The direct-fit perspective emphasizes the tight link between the structure of the world and the structure of the brain" (Hasson et al., p.430).

If our face recognition works along the lines of a Voronoi tessellation of face space, it is a tessellation in which some anchor points effectively coincide (e.g. indistinguishable twins). In addition, new anchor points may need to be added to the map as we go along; however, 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.

6. Recognizing knowledge

In a context in which there is reward for distinguishing cases of knowledge from cases of ignorance, if you have adequate level of experience of cases of both types and there are learnable regularities here, you can learn to distinguish them. Different social actions will win reward depending on the epistemic state of your target: if you want to know which way the coin in my palm is facing, you know you can ask me; if you want me to know how many coins are in your pocket, you must show or tell me. If a given <agent, proposition> pairing swiftly registers as

closer to a prototype of knowledge than to any prototype of ignorance, the agent will be seen as knowing (and vice-versa for ignorance).

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 dimensions separating types of ignorance from ways of knowing, rather than dissuading us from continuing to apply the infallibilist rule that there is some type of state subjects can have only to truths.

Just as we learn to recognize a face by seeing multiple instances of it, so also we learn to recognize knowledge this way, going up a level. In ordinary reinforcement learning, we get adapted to regularities in the world. One interesting set of the regularities in the world is a set of ways in which creatures like us get adapted to regularities in the world, the set of ways of knowing.

References:

Arcaro, M. J., Schade, P. F., & Livingstone, M. S. (2019). Universal mechanisms and the development of the face network: what you see is what you get. *Annual review of vision science, 5*, 341-372.

Augustine. (389 CE/1995). *Against the Academicians and The Teacher* (P. King, Trans.). Indianapolis: Hackett.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305-327.

Chen, D., Cao, X., Wen, F., & Sun, J. (2013). *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.

Chiroro, P., & Valentine, T. (1995). An investigation of the contact hypothesis of the own-race bias in face recognition. *The Quarterly Journal of Experimental Psychology, 48*(4), 879-894.

Douven, I., & Gärdenfors, P. (2020). What are natural concepts? A design perspective. *Mind & Language, 35*(3), 313-334.

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron, 105*(3), 416-434.

Heritage, J. (2012). Epistemics in action: Action formation and territories of knowledge. *Research on Language & Social Interaction, 45*(1), 1-29.

Jenkins, R., Dowsett, A., & Burton, A. (2018). How many faces do people know? *Proceedings of the Royal Society B, 285*(1888), 20181319.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition, 121*(3), 313-323.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems, 25*, 1097-1105.

Kukačka, J., Golkov, V., & Cremers, D. (2017). Regularization for deep learning: A taxonomy. *arXiv preprint arXiv:1710.10686*.

Lewis, M. B., & Johnston, R. A. (1999). A unified account of the effects of caricaturing faces. *Visual cognition, 6*(1), 1-41.

Ma, S., Bassily, R., & Belkin, M. (2018). *The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning.* Paper presented at the International Conference on Machine Learning.

Malpass, R. S., & Kravitz, J. (1969). Recognition for faces of own and other race. *Journal of personality and social psychology, 13*(4), 330.

Nyāya-sūtra. (c.200CE/2017). The Nyāya-sūtra (M. Dasti & S. Phillips, Trans.). In. Indianapolis: Hackett.

Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: a meta-analytic and theoretical review. *Psychological Bulletin, 138*(1), 146.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). *Facenet: A unified embedding for face recognition and clustering.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Sosa, E. (1999). How must knowledge be modally related to what is known? *Philosophical Topics, 26*(1/2), 373-384.

Sporer, S. L. (2001). Recognizing faces of other ethnic groups: An integration of theories. *Psychology, Public Policy, and Law, 7*(1), 36.

Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). *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.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology, 43*(2), 161-204.

Wiese, H., Komes, J., & Schweinberger, S. R. (2013). Ageing faces in ageing minds: A review on the own-age bias in face recognition. *Visual cognition, 21*(9-10), 1337-1363.

Williamson, T. (2000). *Knowledge and its Limits*. New York: Oxford University Press.

Williamson, T. (2009). Replies to Critics. In P. Greenough & D. Pritchard (Eds.), *Williamson on Knowledge* (pp. 279-284). New York: Oxford University Press.

Young, A. W., & Burton, A. M. (2018). Are we face experts? *Trends in cognitive sciences, 22*(2), 100-110.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM, 64*(3), 107-115.