...e of the creators of Photosynth at Microsoft.

May 6, 2017 · 38 min read

Physiognomy's New Clothes

by Blaise Agüera y Arcas, Margaret Mitchell and Alexander Todorov

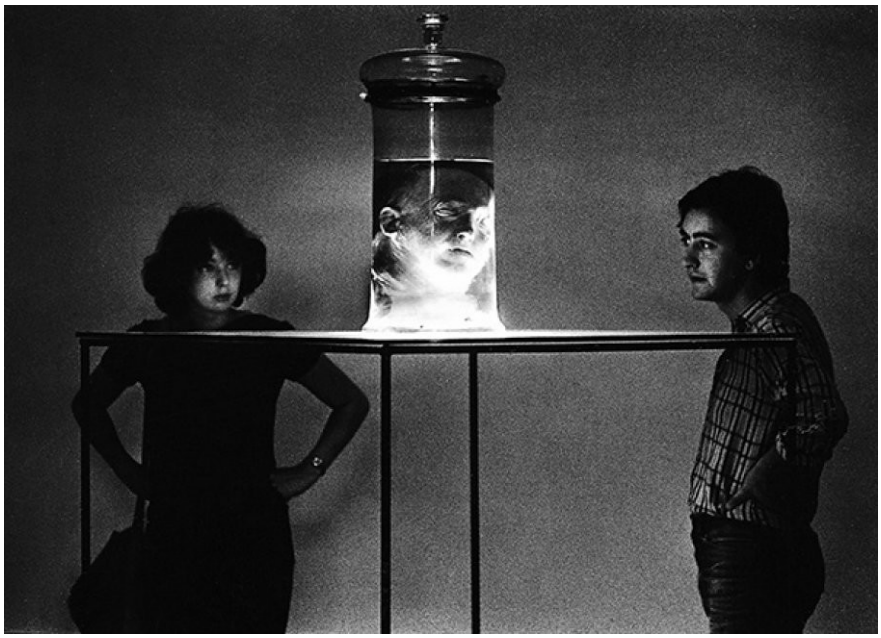

Figure 1. A couple viewing the head of Italian criminologist Cesare Lombroso preserved in a jar of formalin at an exhibition in Bologna, 1978. (Photo by Romano Cagnoni/Hulton Archive/Getty Images)

Introduction

In 1844, a laborer from a small town in southern Italy was put on trial for stealing "five ricottas, a hard cheese, two loaves of bread […] and two kid goats". The laborer, Giuseppe Villella, was reportedly convicted of being a *brigante* (bandit), at a time when brigandage—banditry and state insurrection—was seen as endemic. Villella died in prison in Pavia, northern Italy, in 1864.

Villella's death led to the birth of modern criminology. Nearby lived a scientist and surgeon named Cesare Lombroso, who believed that *brigantes* were a primitive type of people, prone to crime. Examining Villella's remains, Lombroso found "evidence" confirming his belief: a

depression on the occiput of the skull reminiscent of the skulls of "savages and apes".

Using precise measurements, Lombroso recorded further physical traits he found indicative of derangement, including an "asymmetric face". Criminals, Lombroso wrote, were "born criminals". He held that criminality is inherited, and carries with it inherited physical characteristics that can be measured with instruments like calipers and craniographs [1]. This belief conveniently justified his a priori assumption that southern Italians were racially inferior to northern Italians.

The practice of using people's outer appearance to infer inner character is called *physiognomy*. While today it is understood to be pseudoscience, the folk belief that there are inferior "types" of people, identifiable by their facial features and body measurements, has at various times been codified into country-wide law, providing a basis to acquire land, block immigration, justify slavery, and permit genocide. When put into practice, the pseudoscience of physiognomy becomes the pseudoscience of scientific racism.

Rapid developments in artificial intelligence and machine learning have enabled scientific racism to enter a new era, in which machine-learned models embed biases present in the human behavior used for model development. Whether intentional or not, this "laundering" of human prejudice through computer algorithms can make those biases appear to be justified objectively.

A recent case in point is Xiaolin Wu and Xi Zhang's paper, "Automated Inference on Criminality Using Face Images", submitted to arXiv (a popular online repository for physics and machine learning researchers) in November 2016. Wu and Zhang's claim is that machine learning techniques can predict the likelihood that a person is a convicted criminal with nearly 90% accuracy using nothing but a driver's license-style face photo. Although the paper was not peer-reviewed, its provocative findings generated a range of press coverage. [2]

Many of us in the research community found Wu and Zhang's analysis deeply problematic, both ethically and scientifically. In one sense, it's nothing new. However, the use of modern machine learning (which is

both powerful and, to many, mysterious) can lend these old claims new credibility.

In an era of pervasive cameras and big data, machine-learned physiognomy can also be applied at unprecedented scale. Given society's increasing reliance on machine learning for the automation of routine cognitive tasks, it is urgent that developers, critics, and users of artificial intelligence understand both the limits of the technology and the history of physiognomy, a set of practices and beliefs now being dressed in modern clothes. Hence, we are writing both in depth and for a wide audience: not only for researchers, engineers, journalists, and policymakers, but for anyone concerned about making sure AI technologies are a force for good.

We will begin by reviewing how the underlying machine learning technology works, then turn to a discussion of how machine learning can perpetuate human biases.

Machine learning for understanding images

Computers can analyze the physical features of a person by making calculations based on their picture. This is an example of the more general problem of image understanding: a computer program analyzes a photo, makes a determination about the photo, then emits some kind of meaningful judgement (say, "the person in this photo is likely between the ages of 18 and 23").

The relationship between the photo and the response is determined by a set of parameters, which are tuned during a learning phase—hence "machine learning". The most common approach is *supervised learning*, which involves working through a large number of labelled examples—that is, example images paired with the desired output for each. When the parameters are set to random values, the machine will only get the answer right by pure chance; but even given a random starting point, one can slowly vary one or more parameters and ask, "is this variation better, or worse?" In this way, by playing a game of Marco Polo with parameters, a computer can optimize itself to learn the task. A typical training program involves trying millions, billions, or trillions of parameter choices, all the while steadily improving performance on the task. Eventually the improvement levels off, telling us that the accuracy

has probably gotten as good as it's going to get, given the inherent difficulty of the task and the limitations of the machine and the data.

One technical pitfall to guard against is *overfitting*. This happens when the machine is able to memorize the right answers to individual training examples without *generalizing*, meaning learning an underlying pattern that will hold when tested on different data. The simplest way to avoid overfitting is simply to test the performance of the system on a random subset of the labelled data that is "held out", meaning not used during training. If the system's performance on this test data is roughly as good as on the training data, then one can feel confident that the system really has learned how to see a general pattern in the data, and hasn't just memorized the training examples. This is the same as the rationale for giving students a midterm exam with questions they haven't seen before, rather than just reusing examples that have been worked through in class.

Every machine learning system has parameters—or there is nothing to learn. Simple systems may have only a handful. Increasing the number of parameters can allow a system to learn more complex relationships, making for a more powerful learner and, if the relationships between input and output are complex, a lower error rate. On the other hand, more parameters also allow a system to memorize more of the training data, hence overfit more easily. This means that there is a relationship between the number of parameters and the amount of training data needed.

Modern, sophisticated machine learning techniques like convolutional neural networks (CNNs) have many millions of parameters, hence need a great deal of training data to avoid overfitting. Obtaining enough labelled data to both train and test a system is often the greatest practical challenge facing a machine learning researcher.

Example: dating a photo

Convolutional neural networks are very general and very powerful. As an example, consider Ilya Kostrikov and Tobias Weyand's ChronoNet, a CNN that guesses the year in which a photo was taken. Since public sources can provide large numbers of digitally archived photos taken over the past century with known dates, it's relatively straightforward to obtain labeled data (dated photos, in this case) with which to train this network.

Once the network is trained, we can feed in a photo, and we get out the year in which the system guesses it was taken. For example, for the following two photos ChronoNet guesses 1951 (left) and 1971 (right):

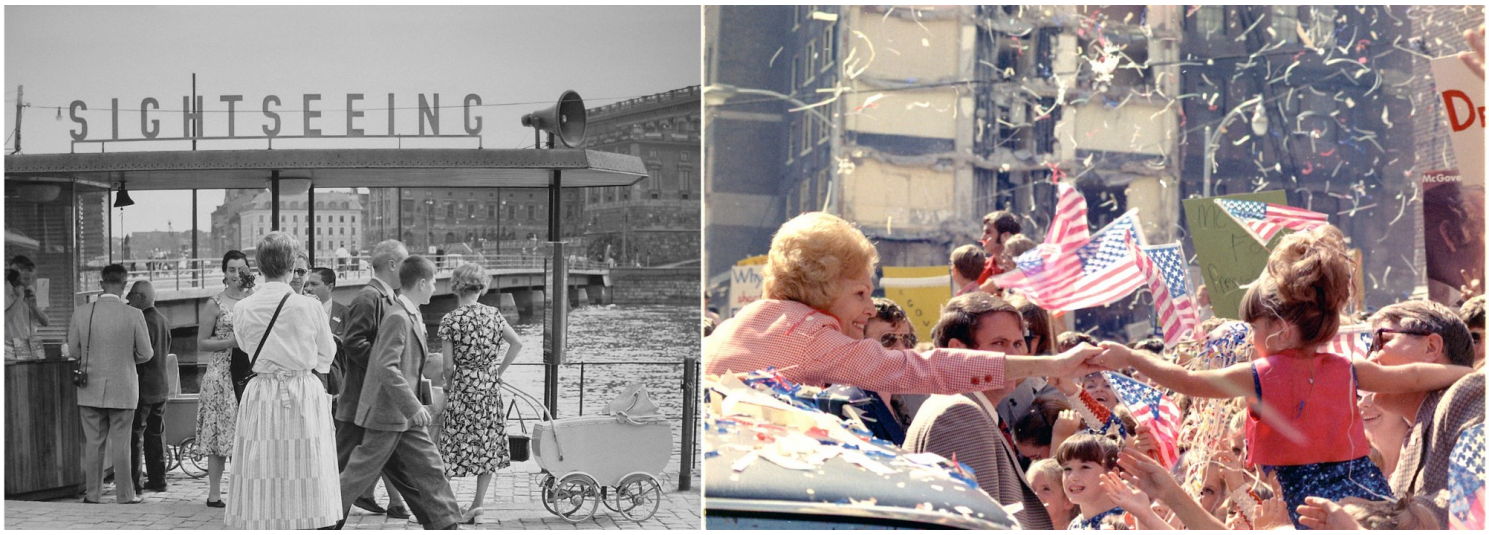

Figure 2. Image dating with deep learning. ChronoNet guesses 1951 (left) and 1971 (right).

These are good guesses. The photo on the left was taken on the Stockholm waterfront in 1950, and the one on the right is of Mrs. Nixon in a 1972 campaign motorcade in Atlanta.

How does the network actually figure it out? From a mechanistic point of view, the millions of learned parameters are just the weights used in a series of weighted average calculations. Starting from the original pixel values, weighted averages are combined, then used as input for a similar set of calculations, which are then used as input for a similar set of calculations, and so on—creating a cascade of weighted average calculations in many layers. [3] In ChronoNet, the final layer outputs values corresponding to probabilities for possible years. While technically correct, this "explanation" is of course no explanation at all; a human expert in dating photographs could equally well say "I answered this way because it's the way my neurons are wired together".

In fact, like a human expert, the artificial neural network has likely learned to be sensitive to a variety of cues, from low-level properties like the film grain and color gamut (as film processing evolved quite a bit during the 20th century) to clothes and hairstyles, car models and fonts. The loudspeaker and the style of pram in the Stockholm photo

might also be clues. The remarkable thing about so-called *deep learning*, which has powered rapid advances in AI since 2006, is that the features relevant to the task (colors, car models, and so on) can be learned implicitly in the service of a higher-level goal (guessing the year). [4]

Previous approaches to machine learning might also have achieved the high-level goal of guessing the year, but would have needed manually written computer code to extract features like fonts and hairstyles from the raw image. Being able to ask a computer to learn a complex problem from end to end, without such custom work, both greatly speeds development and often dramatically improves the result.

This is both the power and the peril of machine learning, and especially deep learning. The power is clear: a general approach can discover implicit relationships in a wide variety of different problems; the system itself learns what to look for. The peril comes from the fact that a scientist or engineer can easily design a classification task that the machine can learn to perform well—without understanding what the task is actually measuring, or what patterns the system is actually finding. This is problematic when the "how" or "why" of such a system's judgments matter, as they certainly would if the judgment purported to be of a person's character or criminal status.

Learning a "criminal type"

"Automated Inference on Criminality Using Face Images" aims to do what ChronoNet does, except that instead of arbitrary photographs it operates on images of faces, and instead of guessing a year, Wu and Zhang's system guesses whether the face belongs to a person with a criminal record or not; thus they claim to "produce evidence for the validity of automated face-induced inference on criminality" for the first time. To understand why this claim is problematic, we need to study the methods and results more closely.

Methods and results

Wu and Zhang begin with a set of 1,856 closely cropped, 80x80 pixel images of Chinese men's faces from government-issued IDs. The men are all between 18 and 55 years old, lack facial hair, and lack facial scars or other obvious markings. 730 of the images are labeled "criminals", or more precisely,

> *"[…] 330 are published as wanted suspects by the ministry of public security of China and by the departments of public security for the provinces of Guangdong, Jiangsu, Liaoning, etc.; the others are provided by a city police department in China under a confidentiality agreement. […] Out of the 730 criminals 235 committed violent crimes including murder, rape, assault, kidnap and robbery; the remaining 536 are convicted of non-violent crimes, such as theft, fraud, abuse of trust (corruption), forgery and racketeering."*

The other 1,126 face images are of

> *"non-criminals that are acquired from Internet using the web spider tool; they are from a wide gamut of professions and social status, including waiters, construction workers, taxi and truck drivers, real estate agents, doctors, lawyers and professors; roughly half of the individuals […] have university degrees."*

It is worth re-emphasizing that all of the face images are from government-issued IDs—the "criminal" images are not mugshots. Otherwise what comes next would be unsurprising.

Wu and Zhang use these labeled examples to do supervised learning. They train the computer to look at a face image and produce a one-bit yes/no answer: did this image come from the "criminals" group or the "non-criminals" group? They try out four different machine learning techniques of varying sophistication, in the sense described earlier—more sophisticated techniques have more parameters and are thus able to learn subtler relationships. One of the less sophisticated techniques involves preprocessing the images with custom code to extract the locations of specific known facial features, like the corners of the eyes and the mouth, then using older methods to learn patterns relating the positions of these facial features. The authors also try a convolutional neural net, AlexNet, based on a widely cited 2012 paper by Google researchers Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton. Its architecture is similar to that of ChronoNet. This CNN, which is both the most modern model and the one with the largest number of parameters, is the strongest performer, achieving a classification accuracy of nearly 90%. Even the older methods, though, have accuracies well above 75%.

This raises several questions, perhaps the first of which is "could this result possibly be real?". More precisely,

1. Are these numbers believable?
2. What is the machine learning picking up on?
3. How does this relate to criminal behavior and criminal judgment?

Likely artifacts

To put into perspective just how extraordinary a 90% accuracy claim is, consider that a well-controlled 2015 paper by computer vision researchers Gil Levi and Tal Hassner find that a convolutional neural net with the same architecture (AlexNet) is only able to guess the *gender* [5] of a face in a snapshot with 86.8% accuracy. [6] Consider, also, that Wu and Zhang's claimed "false alarm rate" (meaning, the incorrect assignment of a "non-criminal" to the "criminal" group) for the CNN-based method is just over 6%—comparable to a workplace drug test.

There are likely issues with the analysis making the claimed accuracy unrealistically high. One technical problem is that fewer than 2000 examples are insufficient to train and test a CNN like AlexNet without overfitting. The lower (though still highly significant) accuracy numbers given by the older non-deep learning methods are likely more realistic.

One should also note that the authors cannot reliably infer that their web-mined government ID images are all of "non-criminals"; on the contrary, if we presume that they are a good random sample of the general population, statistically some fraction of them will also have engaged in criminal activity.

On the other hand there may easily be other systematic differences between the "criminal" and "non-criminal" datasets that a judge would presumably not want to consider evidence of guilt or innocence. The "criminals" may tend to be younger, for example, even if both populations only include people between 18 and 55.

Also in this vein, the three sample images of "non-criminals" shown in the paper (see below) all appear to be wearing white-collared shirts while none of the three "criminals" are. Of course with only three

examples of each we don't know if this is representative of the entire dataset. We do know that deep learning techniques are powerful and will pick up on any cues present, just as ChronoNet picks up on subtle details like film grain in addition to differences in image content. Machine learning does not distinguish between correlations that are causally meaningful and ones that are incidental.

What is the machine learning picking up on?

Setting aside technical errors and confounds that may influence the claimed accuracy, there is probably a real correlation between facial appearance as captured in the images and membership in the "criminal" set. What specific features distinguish purportedly "criminal" faces?

Wu and Zhang are able to use a variety of techniques to explore this in detail. This is especially tractable for the simpler machine learning approaches that involve measuring relationships between standard facial landmarks. They summarize,

> *"[…] the angle* θ *from nose tip to two mouth corners is on average 19.6% smaller for criminals than for non-criminals and has a larger variance. Also, the upper lip curvature* ρ *is on average 23.4% larger for criminals than for noncriminals. On the other hand, the distance d between two eye inner corners for criminals is slightly narrower (5.6%) than for non-criminals."* [7]

We may be able to get an intuitive sense of what this looks like by comparing the top row of "criminal" examples with the bottom row of "non-criminal" examples, shown in the paper's Figure 1:

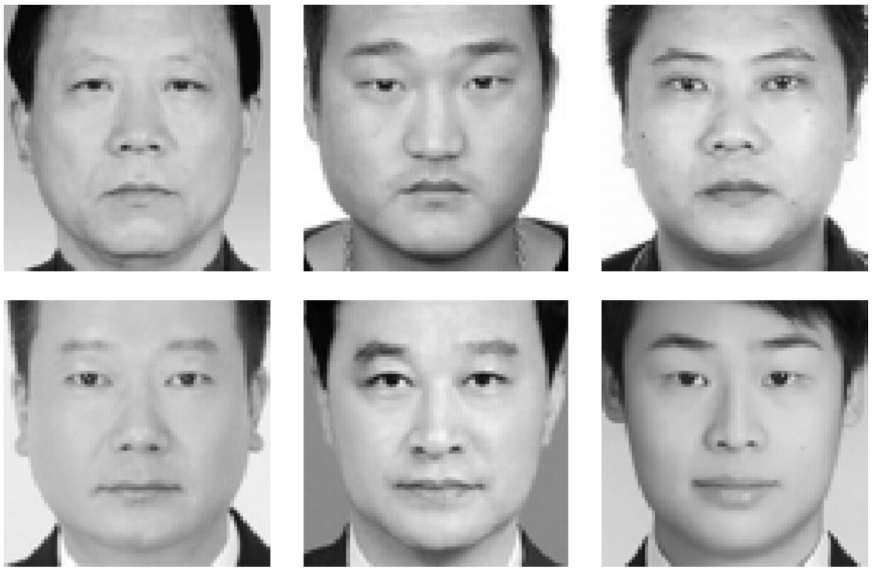

Figure 3. Wu and Zhang's "criminal" images (top) and "non-criminal" images (bottom). In the top images, the people are frowning. In the bottom, they are not. These types of superficial differences can be picked up by a deep learning system.

These are the only six examples the researchers have made public, and there may be a cherrypicking effect at work in the selection of these particular six, but a quick hallway survey (of both Chinese and Western colleagues) suggests that many people, if forced to choose, also find the three bottom photos less likely to be of criminals. For one, although the authors claim to have controlled for facial expression, the three bottom images do all appear to be smiling slightly, while the top ones appear to be frowning.

If these six images are indeed typical, we suspect that asking a human judge to sort the images in order from smiling to frowning would also do a fairly effective job of segregating purportedly "non-criminal" from "criminal". We will return to this point.

What do humans pick up on?

It is worth emphasizing that there is no superhuman magic in this (or any) application of machine learning. While non-experts are only able to date a photo very approximately, most people [8] are exquisitely attuned to faces. We can distinguish hundreds or thousands of acquaintances at a glance and from some distance, register nuances of gaze and expression, and do all of this in well under one tenth of a second. [9]

Wu and Zhang do not claim that their machine learning techniques are recognizing subtler facial cues than people can discern without any help from computers. On the contrary, they connect their work to a 2011 study published in a psychology journal (Valla et al., *The Accuracy of Inferences About Criminality Based on Facial Appearance*) that arrives at the same conclusion using human judgment:

> *"[…] participants, given a set of headshots of criminals and non-criminals, were able to reliably distinguish between these two groups, after controlling for the gender, race, age, attractiveness, and emotional displays, as well as any potential clues of picture origin."*

While Wu and Zhang use ID photos and not mugshots, we should note that in Valla et al.'s paper (despite their claims to have controlled for photographic conditions), the authors compared mugshots of convicted people with pictures of students taken on campus. It is reasonable to assume that mugshots taken in the threatening and humiliating context of arrest look different from pictures taken on a college campus, making the result questionable.

Wu and Zhang also relate their work to a 2014 paper in *Psychological Science* (Cogsdill et al., *Inferring Character From Faces: A Developmental Study*), which was co-authored by one of us. This paper finds that even 3- and 4-year olds can reliably distinguish "nice" from "mean" face images, but critically, there is no claim that these impressions correspond to a person's character. The paper is about the acquisition of facial stereotypes early in development and is based on work that visualizes these stereotypes.

What do supposedly "nice" and "mean" faces look like? Research on social perception of faces in the last decade has shown that one's impression of a face can be reduced to a few basic dimensions, including dominance, attractiveness, and valence. ("Valence" is associated with positive evaluations like "trustworthy" and "sociable".) Various methods have been developed to visualize the facial stereotypes that map onto these dimensions. In one, participants rate randomly generated synthetic faces on traits like trustworthiness and dominance. Because the faces are generated from a statistical model that varies the relative sizes or positions of different facial features, it's possible to calculate average features representing a "trustworthy" or "untrustworthy" face; for white males, such faces look like this:

Figure 4. Stereotypically "nice" (left) and "mean" (right) faces, according to both children and adults.

The "untrustworthy" face and Wu and Zhang's "criminal" face (Figure 3) look related.

The fallacy of objectivity

Wu and Zhang don't use scare quotes in asserting a relationship between people's impressions (e.g., "untrustworthy") and purportedly objective reality (e.g., "criminal"), instead claiming that the kinds of facial features we see on the right *imply* criminality. This incorrect assertion rests on the presumed objectivity and independence of the inputs, outputs, and the algorithm in between.

Because the algorithm they use is based on a highly general deep learning technique that can learn patterns from any kind of image data —a convolutional neural network—it is reasonable to call it objective, that is, it does not in itself embody biases about facial appearance or criminality.

The input is presumed to be objective because it is a standardized ID photo. The output is presumed to be objective because it is a legal judgment—and independent of the input because justice is presumed to be, in the most literal sense, blind. As the authors put it,

> *"We are the first to study automated face-induced inference on criminality free of any biases of subjective judgments of human observers."*

The claims to objectivity in the inputs and outputs are misleading, as we will see, but what is most troubling about this work is its invocation of two different forms of authority, scientific and legal, to once again

resurrect and "prove" the existence of a hierarchy of virtue among "types" of people. Those with more curved upper lips and eyes closer together are of a lower social order, prone to (as Wu and Zhang put it) "a host of abnormal (outlier) personal traits" ultimately leading to a legal diagnosis of "criminality" with high probability.

This language closely echoes that of Cesare Lombroso. Before exploring the likely reasons for correlations between facial appearance inputs and criminal judgment outputs, it's worth pausing and reviewing the history of such claims.

Scientific racism

Physiognomy and the theory of "types" [10]

The roots of physiognomy lie in the human propensity to interpret a person's appearance associatively, metaphorically, and even poetically. This kind of thinking, dating back at least to the ancient Greeks, [11] is evident in the Renaissance polymath Giambattista della Porta's book *De humana physiognomonia*, which makes the case visually that a piggish-looking person *is* piggish: [12]

Figure 5. Like man, like swine: From Giambattista della Porta's De humana physiognomonia (Naples, 1586).

To make such ideas respectable in the Enlightenment, it was necessary to excise the poetry from them and concentrate on more specific

physical and behavioral features. In the 1700s, the Swiss theologian Johann Caspar Lavater analyzed character based on the shape and positions of the eyes, brows, mouth, and nose to determine whether a person was, among other characteristics, "deceitful", "full of malice, "incurably stupid", or a "madman".

In this vein, Victorian polymath Francis Galton (1822–1911) tried to empirically characterize "criminal" types by superimposing exposures of convicts on the same photographic plate. Around the same time, Lombroso took physiognomic measurement further with his more "scientific" criminological approach. [13] While Lombroso can be credited as one of the first to attempt to systematically study criminality, he can also be credited as one of the first to use modern science to lend authority to his own stereotypes about lesser "types" of humans.

Figure 6. Francis Galton's attempt to reconstruct an "average criminal face".

Scientific rigor tends to weed out incorrect hypotheses given time, peer review, and iteration; but using scientific language and measurement doesn't prevent a researcher from conducting flawed experiments and drawing wrong conclusions—especially when they confirm preconceptions. Such preconceptions are as old as racism itself.

Scientific racism from 1850–1950

The beliefs Lombroso appears to have harbored with respect to "southerners" in Italy suggested a racial hierarchy with political implications, but 19th century American physiognomists had even more compelling reasons to rationalize such a hierarchy: they were slave-owners. Samuel Morton used cranial measurements and ethnological arguments to make a case for white supremacy; as his

followers Josiah Nott and George Gliddon quoted in their 1854 tribute *Types of Mankind*,

> *"Intelligence, activity, ambition, progression, high anatomical development, characterize some races; stupidity, indolence, immobility, savagism, low anatomical development distinguish others. Lofty civilization, in all cases, has been achieved solely by the "Caucasian" group."*

Despite this book's scholarly pretensions and the intervening centuries, its figures (typical of the period) illustrate the same kind of fanciful visual "reasoning" and animal analogies evident in della Porta's treatise, though in the modern context yet more offensively:

Figure 6. The idea that there are inferior types of humans has historically been linked to the scientifically invalid idea that some humans are more like animals than others. From Nott and Gliddon, Types of Mankind, 1854.

Later in the 19th century, Darwinian evolutionary theory refuted the argument made in *Types of Mankind* that the races are so different that they must have been created independently by God. However, by making it clear that humans *are* in fact animals, and moreover closely related to the other great apes, it provided fertile ground for Morton's discrete racial hierarchy to be reimagined in shades of grey, differentiating humans who are "more human" (more evolved, physically, intellectually and behaviorally) and "less human" (less evolved, physically closer to the other great apes, less intelligent, and less "civilized"). [14] Darwin wrote in his 1871 book *The Descent of Man*:

> *"[…] man bears in his bodily structure clear traces of his descent from some lower form; […] [n]or is the difference slight in moral disposition between a barbarian, such as the man described by the old navigator Byron, who dashed his child on the rocks for dropping a basket of sea-urchins, and a Howard or Clarkson; and in intellect, between a savage who does not use any abstract terms, and a Newton or Shakspeare. Differences of this kind between the highest men of the highest races and the lowest savages, are connected by the finest gradations."*

Unsurprisingly, Darwin's apex of humanity is peopled by the physicist Isaac Newton, the playwright William Shakespeare, the abolitionist Thomas Clarkson, and the philanthropist John Howard: all were English, Christian, white, male, and from the educated classes—that is, much like Darwin himself. Darwin's views were in step with (and, in some ways, more progressive than) those of his peers; more generally, they illustrate *homophily*, the pervasive cognitive bias causing people to identify with and prefer people similar to themselves.

This combination of homophily, rationalization of a racial hierarchy based on heritable physical and behavioral traits, and a resulting theory of physiognomic "types" survived far into the 20th century. The details of the hierarchy shifted depending on the beliefs and sympathies of the theorizer. For the German evolutionary biologist Ernst Haeckel (1834–1919), Jews shared a high place alongside Germans and English people in this hierarchy; [15] but in the Nazi era, such hierarchies were used to cartoon and vilify Jews just as Haeckel and his precursors had done for "Papuans", "Hottentots", and other foreigners with whom they had no social ties. For example, the 1938 children's book *Der Giftpilz* (*The Toadstool*), used as a state-sponsored school textbook, cautioned that

> *"Just as it is often difficult to tell a toadstool from an edible mushroom, so too it is often hard to recognize the Jew as a swindler and criminal […] How to tell a Jew: the Jewish nose is bent. It looks like the number six […]".*

Figure 7. From Vaught's Practical Character Reader, 1902, p. 80.

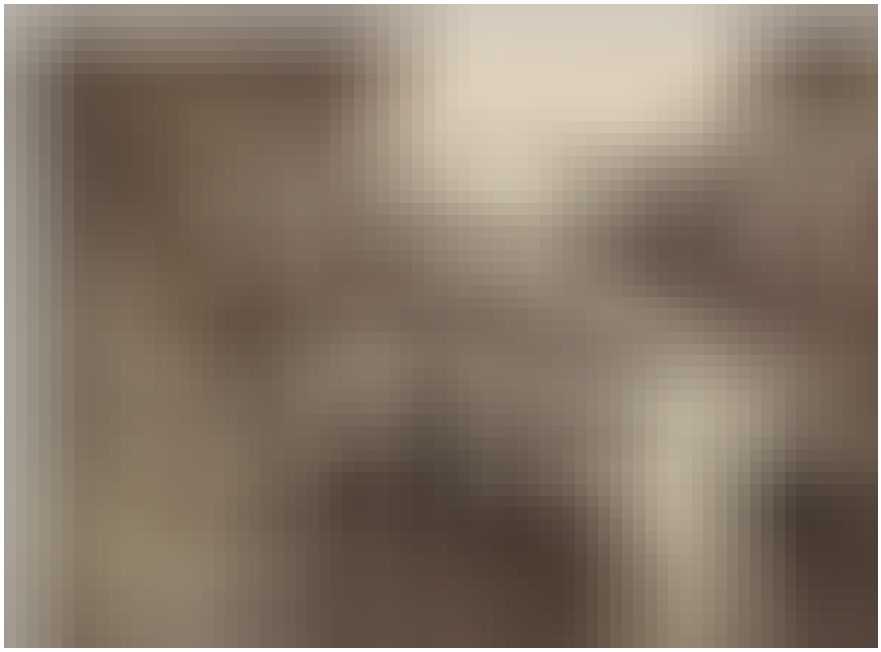

Figure 8. Nazi "race scientists" doing institutionalized physiognomy, 1933.

Scientific racism today

Despite the social and scientific progress of the past half-century, scientific racism is less firmly relegated to the past than many of us would have assumed. Present-day American "pickup artist" and white nationalist James Weidmann, for example, has blogged in support of physiognomy:

> *"There's evidence (re)emerging […] that a person's looks do say something about his politics, smarts, personality, and even his propensity to crime. Stereotypes don't materialize out of thin air, and the historical wisdom that one can divine the measure of a man (or a woman) by the cut of his face has empirical support. […] You CAN judge a book by its cover: ugly people are more crime-prone. […] Physiognomy is real. It needs to come back as a legitimate field of scientific inquiry […]".*

What Wu and Zhang's paper purports to do is precisely that; and while they do not directly suggest applications for their deep learning-based physiognomy, they are excited about its implications for "social psychology, management science, [and] criminology".

An Israeli startup, Faception, has already taken the logical next step, though they have not published any details about their methods, sources of training data, or quantitative results:

> *"Faception is first-to-technology and first-to-market with proprietary computer vision and machine learning technology for profiling people and revealing their personality based only on their facial image."*

The Faception team are not shy about promoting applications of their technology, offering specialized engines for recognizing "High IQ", "White-Collar Offender", "Pedophile", and "Terrorist" from a face image. [16] Their main clients are in homeland security and public safety. Faception is betting that once again governments will be keen to "judge a book by its cover".

Unexamined assumptions

Perhaps unsurprisingly, the present-day researchers whose work on social perception of faces Wu and Zhang cite as an inspiration tend to take a more nuanced view of the phenomena they are studying. On one hand, this work has shown that people can form character impressions such as trustworthiness from facial appearance after seeing a face for less than one tenth of a second and that these impressions predict important social outcomes, ranging from political elections to economic transactions to legal decisions. On the other hand, while we form impressions almost reflexively from facial appearance, this does not imply that these impressions are accurate. The evidence suggests that they are not.

Fundamentally, the idea that there might be some "criminal type", and that this is evident on a person's face, rests on several flawed assumptions:

1. The appearance of a person's face is purely a function of innate properties;
2. "Criminality" is an innate property in a certain group of people;
3. Criminal judgment by a legal system reliably determines "criminality" in a way that is unaffected by facial appearance.

Let's examine each assumption in turn.

Reading character into faces

Facial structure is not purely innate, but it is also powerfully shaped by development, [17] environment, and context. A photograph of a person's face further depends on the setting and conditions during photography. Since all of these additional factors can play important roles in the perception of faces—whether by humans or by machines—they are worth summarizing.

Dorothea Lange's famous Depression-era photos, such as her "Migrant Mother" series from 1936, take as their subject the emotional shaping of the human face and body by a difficult environment. They can be seen as portraits of the Dust Bowl itself, as refracted through the faces of those unlucky enough to have lived on the American prairie in the 30s. In each such image, the viewer is invited to ask, "what would this person look like under different circumstances, in another time and place?" Foreheads are weatherbeaten and grim; facial muscles reconfigure the expression around a baseline of anxiety and despair; "upper lip curvatures" are all large. In this sense Lange's photos can almost be read as a critique of physiognomy in their own right.

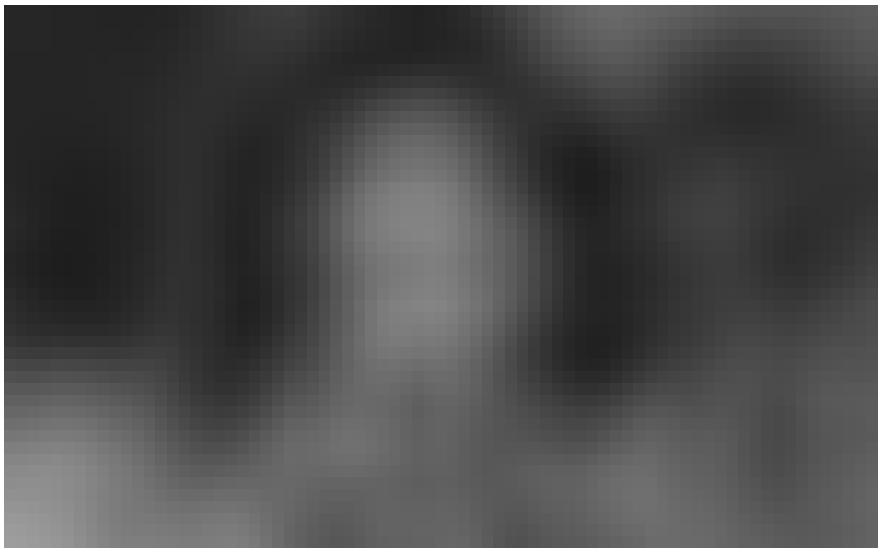

Figure 9. From Dorothea Lange's "Migrant Mother" series. The original caption reads: "Destitute peapickers in California; a 32 year old mother of seven children. February 1936."

Of course, Lange's photos are also artistic statements, and reflect her own vision of the Dust Bowl and its people. We must use caution in assuming that such a portrait can be read as a "pure" representation of its subject.

Research shows that the photographer's preconceptions and the context in which the photo is taken are as important as the faces themselves; different images of the same person can lead to widely different impressions. It is relatively easy to find a pair of images of two individuals matched with respect to age, race, and gender, such that one of them looks more trustworthy or more attractive, while in a different pair of images of the same people the other looks more trustworthy or more attractive. Consider this example, from a 2011 paper in the journal *Cognition* by Mike Burton and colleagues:

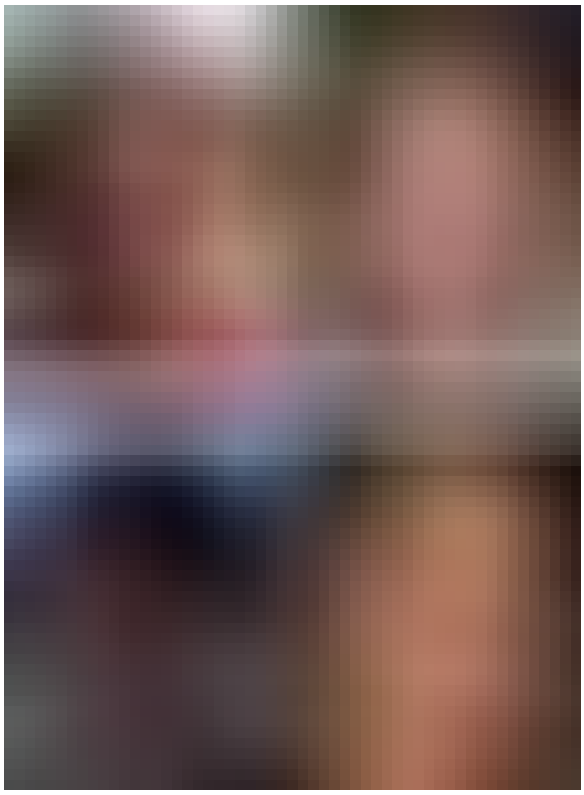

Figure 10. Stimuli in R. Jenkins et al's 2011 paper Variability in photos of the same face.

Most people see the face on the left in the top row as more attractive than the face on the right. Most people also see the face on the left in the bottom row as *less* attractive than the face on the right. However, the two faces on the left are different images of the same person; so are the two faces on the right.

In a recent informal experiment, Canon Lab Australia invited five professional photographers to spend a few minutes with the same man and "register" his essence. Each photographer was given false information about the person, and this false information led to dramatically different photographs. The "self-made millionaire" is staring into the future, whereas the "ex-inmate" looks withdrawn and skeptical. Standard-issue photos such as those used on government IDs are more uniform and presumably more neutral than the Canon Lab Australia photos, but without a carefully controlled experiment, unaccounted biases relating to the setting and the photographer will show up in the data—as they likely do in the 2011 paper by Valla et al. (*The Accuracy of Inferences About Criminality Based on Facial Appearance*) comparing criminal mugshots with photos taken on a college campus.

Overgeneralizing the "resting face"

The idea that there is a perfect correspondence between a person and their image is a psychological illusion fueled by our experience with familiar faces. We instantly recognize images of familiar people, and this recognition evokes our memories and feelings about them. But there is no equivalent process when we look at images of strangers. Each image generates a different and arbitrary impression.

This is in part because it is very difficult to fully separate emotion—even simple impressions like whether a person is smiling or frowning—from the identity of the face itself. Many impressions generated by so-called emotionally neutral faces can be predicted by the similarity of their "neutral" expressions to emotional expressions.

Consider the synthetically generated "trustworthy" and "untrustworthy" faces shown earlier. We can see that trustworthy faces have more positive expressions than untrustworthy faces, and that they are more feminine. That is, impressions of trustworthiness are based on similarity to momentary emotional expressions, which signal behavioral intent, as well as gender stereotypes. In the field of social perception of faces, these impressions of character are understood as an overgeneralization from the here-and-now of the person's possible intentions to what the person is "like" in general. In other words, social intention can be conveyed by moving the face into different configurations, but different people's faces also fall into different spots and cover different gamuts of this same space of configurations—thus our socially useful ability to read intention has a tendency to overgeneralize and wrongly project emotion or intention onto certain people. We can speculate that this effect might be especially pronounced in a snapshot, where the viewer is unable to gauge context or see more of the expressive range of the face.

Essentialism

This kind of facial overgeneralization is an illustration of *essentialism*, the (incorrect) idea that people have an immutable core or essence that fully determines both appearance and behavior. These were the beliefs of Lavater, Lombroso, and Galton—whose life obsession was eugenics. In modern times, genes often play the role of essence, which in earlier periods took on a more philosophical or even mystical character.

Essentialism often seems to color human thought. As Stephen Jay Gould put it in his 1981 book *The Mismeasure of Man*,

> *"The spirit of Plato dies hard. We have been unable to escape the philosophical tradition that what we can see and measure in the world is merely the superficial and imperfect representation of an underlying reality. […] The technique of* correlation *has been particularly subject to such misuse because it seems to provide a path for inferences about causality (and indeed it does, sometimes—but only sometimes)."*

Essentialist reasoning is often circular. For example, in 19th century England, women were commonly held to be essentially incapable of abstract mathematical thought. This was used as a rationale for barring them from higher education in mathematics ("what's the point?"). Of course, without access to higher education, it was exceedingly difficult for Victorian women to break out of this cycle; yet the absence of women doing higher math was the evidence that they could not do it. Even when, against all odds, a woman managed to rise to the top of the heap, as Philippa Fawcett did when she obtained the top score in the prestigious Cambridge Mathematical Tripos exams in 1890, this was regarded as a freak result rather than indicative of a flawed assumption. [18] Although over the past century we have seen many more examples of first-rate female mathematicians, we are still struggling with this kind of confirmation bias and the legacy of gender essentialism in STEM fields.

Criminality

We have seen that facial appearance is influenced by both essential (genetically inherited) and non-essential (environmental, situational, and contextual) factors. What about criminality? Are criminals really a "type"?

The "criminal class"

Like physiognomy itself, the idea of a "criminal type" or "criminal class" held great currency in the 19th century. Historian and cultural critic Robert Hughes colorfully narrates England's 80 year experiment in relocating its criminal class to Australia in his book *The Fatal Shore*. As he describes from the perspective of colonial art, the transported convict was

> *"not so much "brutalized" (in the modern sense: deformed by ill-treatment) as he was "a brute," whose criminal nature was written on his very skin."*

Sending England's criminals to Australia promised to reduce crime in England—though there is no indication that this worked. What it did accomplish was to foster the essentialist anxiety that the "convict stain" in Australia would pass down through generations, resulting in a perpetually criminal and brutal society down under. Yet,

> *"[…] the truly durable legacy of the convict system was not "criminality" but the revulsion from it: the will to be as decent as possible, to sublimate and wipe out the convict stain, even at the cost […] of historical amnesia."* [19]

It perhaps goes without saying that the idea of a "criminal class" was very much bound up in the idea of social class; in practice, the great majority of transported convicts were poor, and many of their crimes—as in any era—were a function of poverty. Probably many of them would not have looked out of place among Dorothea Lange's hard-bitten Dust Bowl migrants. Yet despite the well-documented horrors of penal life, once freed, many of these ex-convicts and their descendants found themselves in a much improved situation compared to 19th century urban poverty and class oppression back in England. Their "criminality" turned out to be circumstantial, not essential. As Georg Christoph Lichtenberg, the person most responsible for unraveling the Lavater's "science", put it,

> *"What do you hope to conclude from the similarity of faces, especially the fixed features, if the same man who has been hanged could, given all of his dispositions, have received laurels rather than the noose in different circumstances? Opportunity does not make thieves alone; it also makes great men."*

Can we, then, make any claims at all about what it might mean for someone to be an intrinsically "criminal type"?

Testosterone

Gender is a good place to start: empirically, people charged with violent crimes tend to be male. Higher testosterone level is likely to be a causal factor, both because it appears to increase aggression and appetite for

risk, and because it increases physical strength. [20] These findings have even been replicated in a range of non-human animals.

While testosterone is arguably not strictly "essential"—its blood concentration can vary depending on the situation, and it can be manipulated pharmaceutically—it comes close. There is also evidence that both prenatal testosterone level and responsiveness to testosterone influence aspects of the body plan, including the length ratio of the index and ring finger, as well as some aspects of behavior, again including aggression. This body of work implies that there are developmental variables influencing both body and behavior; modern proponents of physiognomy invariably point to this work in defense of their position.

However, some perspective regarding these findings is useful. The sorts of correlations described in these papers are far from powerful enough to let appearance stand in for a lab test:

> *"In pairs of either natural or composite faces the face higher in testosterone was chosen as more masculine 53% and 57% of the time respectively. The authors argue that only men with very high or very low levels of testosterone may be visually distinguishable in terms of their masculinity. […] other studies find no links between testosterone and masculinity. A study using almost identical methods […] but with a much larger set of men, found no association between perceived facial masculinity and testosterone levels […] Similarly, Neave, Laing, Fink, and Manning (2003) reported links of perceived facial masculinity with second-to-fourth digit ratio (2D:4D), but not with measured baseline testosterone levels; and Ferdenzi, Lemaître, Leongómez, and Roberts (2011) found no association between perceived facial masculinity and 2D:4D ratio."*

In short, studies have shown that body appearance can weakly correlate with behavior in some circumstances—as would, one suspects, many other superficial cues (e.g. white collars on "non-criminals"). But these correlations fall far short of being suitable as proxy variables.

Deep learning can do a better job of extracting nuanced information from an image than simple feature measurements like the face width-to-height ratio. But, as we have pointed out, it is not magic. Many of the papers discussed above use double-blind trials with human judges,

precisely because humans are very good at face perception tasks. Deep learning can't extract information that isn't there, and we should be suspicious of claims that it can reliably extract hidden meaning from images that eludes human judges.

The alternative is that this information does *not* elude human judges, any more than it eludes most of us when we look at Wu and Zhang's three "criminal" and three "non-criminal" sample ID photos.

Judgment

Over the last several years in the US, we have seen increasing attention to the long-running problem of mass incarceration. While the US comprises about 5% of the world's population, it contains about 25% of the global prison population—2.4 million people. Those incarcerated are disproportionately poor and of color; in the US, being a black male makes you nearly seven times likelier to be incarcerated than if you were a white male. [21] This would make a race detector for face images a fairly effective predictor of "criminality" in the US, if by this word we mean—as Wu and Zhang do in China—someone who has been convicted by the legal system.

Are such convictions fair? Due to the long shadow of slavery and systematic discrimination, a disproportionate number of black people in the US live in difficult economic circumstances, and this in itself is associated with increased criminal conviction, as was the case for England's white economic underclass in the 19th century. However, the incarceration disparity is far greater than one would expect from this effect alone.

Many different lines of evidence suggest that black people are arrested more often, judged guilty more often, and sentenced more harshly than white people who have committed the same crime. For example, the black imprisonment rate for drug offenses is about 5.8 times higher than it is for whites, despite roughly comparable prevalence of drug use. People who are black also serve longer sentences. A recently published large-scale longitudinal study finds that even the poorest white children are less likely to go to prison at some point than all but the wealthiest 10% of black children. Once in prison, black people are treated more harshly by the correctional facility. Direct tests of racial bias among trial judges have been conducted using hypothetical cases, and have demonstrated harsher judgment of (hypothetical) black

defendants, especially when the judges harbor high levels of implicit [22] racial bias—which is endemic among judges just as among the general population.

If one controls for race, as Wu and Zhang did in their experiment, [23] do we eliminate these kinds of implicit biases on the part of the judges who establish the experiment's criminality "ground truth"?

A large body of research suggests otherwise. [24] To list a few examples, in 2015 Brian Holtz of Temple University published the results of a series of experiments in which face "trustworthiness" was shown to strongly influence experimental participants' judgment. Specifically, the participants were asked to decide, after reading an extended vignette, whether a hypothetical CEO's actions were fair or unfair. While the judgment varied (as one would hope) depending on how fair or unfair the actions described in the vignette were, it also varied depending on whether a "trustworthy" or "untrustworthy" face was used in the CEO's profile photo. The photos were of faces with high and low "trustworthiness", per Oosterhof and Todorov's 2008 paper. In another study, participants played an online investment game with what they believed were real partners represented by "trustworthy" or "untrustworthy" faces. Participants were more likely to invest in "trustworthy" partners even in the presence of reputational information about the past investment behavior of their partners. Yet more chillingly, a recent study found that among prisoners convicted for first degree murder, the unlucky ones with "untrustworthy" faces were disproportionately more likely to be sentenced to death than to life imprisonment. This was also the case for people who were falsely accused and subsequently exonerated.

Recall that these are the same kinds of "face trustworthiness" judgments (or prejudices) that are already clearly exhibited by 3- and 4-year olds. This does not reflect some inner intuitive genius we are endowed with for accurately judging character at a glance. [25] In fact, the evidence suggests that in many cases, we will do much better if were to ignore the faces and rely on general knowledge about the world. Moreover, studies in which the trustworthiness of economic behavior was measured show that relying on face judgments can make our decisions not more but *less* accurate.

So in summary:

- A machine learned "criminality detector" can pick up on the same things humans pick up on when we look at an image of a face;
- When viewing "criminal" and "non-criminal" face images, what such a detector picks up on is likely related to negative face perceptions;
- Human judges who produce criminality "ground truth" data are themselves strongly influenced by this "untrustworthy" look; and
- The "untrustworthy" look seems not to be a good predictor of actual untrustworthiness—and is unlikely to be predictive of criminality.

This is unfortunate for someone who happens to have an "untrustworthy" face. It is also unfortunate that, rather than finding an efficient and impartial shortcut to making accurate criminal judgments with a computer (perhaps a misguided goal in any case), what Wu and Zhang's experiment likely reveals is the inaccuracy and systematic unfairness of many human judgments, including official ones made in a criminal justice context.

We expect that more research will appear in the coming years that has similar biases, oversights, and false claims to scientific objectivity in order to "launder" human prejudice and discrimination.

Feedback loops

> *"It sucks to be poor, and it sucks to feel that you somehow* deserve *to be poor. You start believing that you're poor because you're stupid and ugly. And then you start believing that you're stupid and ugly because you're Indian. And because you're Indian you start believing you're destined to be poor. It's an ugly circle and* there's nothing you can do about it*."*
>
> — Sherman Alexie, *The Absolutely True Diary of a Part-Time Indian*

There are already many feedback loops in society that create compounding effects for disadvantage. This has been written about extensively in the context of race, disability, and other categories that have historically been associated with identity.

In addition to the psychological weight of internalized negativity that Sherman Alexie points out, there are powerful pragmatic consequences

arising from the same biases being applied to a person repeatedly. If something about one's appearance causes teachers to suspect cheating, schoolmates to avoid sitting at the same lunch table, strangers to avoid striking up conversation, potential employers to refrain from making an offer, and police officers to "stop and frisk" more often, it would be surprising not to find significant long-term consequences.

What is most alarming about the prospect of Wu and Zhang's work being used as a tool for police and security applications, as the Faception startup does, is that it "scientifically" legitimizes a correlation that itself emerges from training data with embedded social bias. Wu and Zhang get their own result exactly wrong when they write,

> *"Unlike a human examiner/judge, a computer vision algorithm or classifier has absolutely no subjective baggages, having no emotions, no biases whatsoever due to past experience, race, religion, political doctrine, gender, age, etc., no mental fatigue, no preconditioning of a bad sleep or meal. The automated inference on criminality eliminates the variable of meta-accuracy (the competence of the human judge/examiner) all together."*

This kind of rhetoric advocates for replacing biased human judgment with a machine learning technique that embeds the same bias—and more reliably. Worse, however, it argues that introducing machine learning into an environment where it can augment or scale up human judgment of criminality can help to make things fairer. In fact it will do the opposite, because humans will assume that the machine's "judgment" is not only consistently fair on average but independent of their personal biases. They will thus read agreement of its conclusions with their intuition as independent corroboration. Over time it will train human judges who use it to gain confidence in their ability to recognize criminality in the same manner. Our existing implicit biases will be legitimized, normalized, and amplified. We can even imagine a runaway effect if subsequent versions of the machine learning algorithm are trained with criminal convictions in which the algorithm itself played a causal role.

"Predictive policing" (listed as one of TIME Magazine's 50 best inventions of 2011) is an early example of such a feedback loop. The idea is to use machine learning to allocate police resources to likely crime spots. Believing in machine learning's objectivity, several US

states implemented this policing approach. However, many noticed that the system was learning from previous data. If police were patrolling black neighborhoods more than white neighborhoods, this would lead to more arrests of black people; the system then learns that arrests are more likely in black neighborhoods, leading to reinforcement of the original human bias. It does not result in optimal policing with respect to actual incidence of crime.

Conclusion

On a scientific level, machine learning can give us an unprecedented window into nature and human behavior, allowing us to introspect and systematically analyze patterns that used to be in the domain of intuition or folk wisdom. Seen through this lens, Wu and Zhang's result is consistent with and extends a body of research that reveals some uncomfortable truths about how we tend to judge people.

On a practical level, machine learning technologies will increasingly become a part of all of our lives, and like many powerful tools they can and often will be used for good—including to make judgments based on data faster and fairer.

Machine learning can also be misused, often unintentionally. Such misuse tends to arise from an overly narrow focus on the technical problem, hence:

- Lack of insight into sources of bias in the training data;
- Lack of a careful review of existing research in the area, especially outside the field of machine learning;
- Not considering the various causal relationships that can produce a measured correlation;
- Not thinking through how the machine learning system might actually be used, and what societal effects that might have in practice.

Wu and Zhang's paper illustrates all of the above traps. This is especially unfortunate given that the correlation they measure—assuming that it remains significant under more rigorous treatment—may actually be an important addition to the already significant body of

research revealing pervasive bias in criminal judgment. Deep learning based on superficial features is decidedly *not* a tool that should be deployed to "accelerate" criminal justice; attempts to do so, like Faception's, will instead perpetuate injustice.

Thanks

- Charina Choi, Google
- Jason Friedenfelds, Google
- Tobias Weyand, Google
- Tim Freeman, Google
- Alison Lentz, Google
- Jac de Haan, Google
- Meredith Whittaker, Google
- Kathryn Hume, Fast Forward Labs

Notes

[1] The craniograph, for measuring the silhouette of a skull, was one of a number of instruments developed specifically for such applications.

[2] Around the same time, a paper on predicting first impressions from faces using deep learning, which correctly identified that they were measuring subjective impressions—not objective character—received less attention.

[3] This layered architecture is loosely modeled on the brain's visual cortex, with each weight corresponding to the strength of a synapse, or electrochemical connection from one neuron to another.

[4] Many convolutional neural networks, including ChronoNet, fall into the category of deep learning. The "deep" means that there are many layers of consecutive operations (hence many parameters).

[5] In this paper gender is modeled as binary and ground truth is based on self-declared gender identity.

[6] In fairness, the real-world corpus of snapshots analyzed in this paper includes some blurry images, people facing away or wearing large sunglasses, and other difficult cases not found in ID photos.

[7] This is reminiscent of the "facial angle" measurement used by Dutch scholar Pieter Camper (1722–89) to "infer" intelligence.

[8] There are specific cognitive disorders that impair some people's performance at this task, just as dyslexia impairs reading. "Face blindness" or prosopagnosia may affect ~2.5% of the population, including some surprising cases like the portrait artist Chuck Close.

[9] One of the great achievements of machine learning over the past few years has been to finally—after decades of effort by many research labs—match human acuity in face recognition.

[10] This is a link to the first chapter of the new book *Face Value: The Irresistible Influence of First Impressions*, written by one of us (Todorov), which includes a more thorough review of the history of physiognomy.

[11] Similar analogical thinking was behind the "theory of humors", also Greek in origin, which held that the balance of blood, phlegm, black bile and yellow bile determined both health and personality. A number of English words still in use derive from this theory: sanguine, phlegmatic, bilious, choleric, melancholic.

[12] What it might mean for a human personality to be "piggish" is of course a second analogical leap.

[13] Neither his methods nor his analysis would pass muster today. His measurements were selective, his datasets small, and his samples biased.

[14] In *The Descent of Man* Darwin wrote of slavery that it was "the great sin", though in the next breath noted, "Some savages take a horrid pleasure in cruelty to animals, and humanity with them is an unknown virtue." For Darwin the sin of slavery was thus one of cruelty, not of inequality.

[15] Haeckel had Jewish friends and colleagues, and was Germany's leading popularizer of Darwin, so it is perhaps unsurprising to find both Jewish and English people favored in his racial hierarchy.

[16] On their website they also annotate these "types" with pop-psychological descriptions that read very much like typologies in a vintage book on physiognomy, as in their "Bingo Player" description (yes, they have a Bingo Player detector): "Endowed with a high mental ceiling, high concentration, adventurousness, and strong analytical abilities. Tends to be creative, with a high originality and imagination, high conservation and sharp senses."

[17] For example: an identical twin study finds that sun exposure, smoking, and body mass index (determined in large part by food and exercise habits), significantly affect facial aging.

[18] "Her score was 13 per cent higher than the second highest score, but she did not receive the title of senior wrangler, as only men were then ranked, with women listed separately." (Wikipedia.)

[19] See *The Fatal Shore*, Chapter 10, endnote 48.

[20] Violent crimes are especially associated with men. In the US, men are convicted of 98% of rape, 90% of murder, and 78% of aggravated assault, but only 57% of larceny-theft and 51% of embezzlement. Canadian data are similar. This suggests that nonviolent or white collar crime may not be strongly gendered.

[21] In this statistic "black" and "white" both exclude populations also identifying as Hispanic/Latino.

[22] Racial implicit bias is measured using reaction time on a simple classification test (white and black faces together with good and bad words). It reveals positive or negative associations with race, independently of a subject's conscious beliefs. It is exceedingly common for a person to harbor implicit bias even if that person is not consciously or explicitly racist; in fact, black subjects typically also exhibit implicit anti-black bias.

[23] Their subjects are all Chinese men; there may be questions worth asking regarding finer-grained ethnic, economic and educational distinctions between the "criminal" and "non-criminal" sets that are beyond the scope of this analysis.

[24] This research, as well as the lack of evidence for the accuracy of first impressions, is also reviewed in *Face Value: The Irresistible Influence*

of First Impressions.

[25] Malcolm Gladwell's book *Blink* popularized the idea that snap judgments (also referred to as "thin slices") can be just as accurate as rational consideration. While intuitively appealing, this view has limited validity, as the book itself acknowledges, and has been widely critiqued.