## Chapter 14: A view from above

When computing and AI pioneer Alan Turing first proposed his “Imitation Game” in 1950 as a test of artificial intelligence, it was framed in terms of gender recognition:

I propose to consider the question, ‘Can machines think?’ This should begin with definitions of the meaning of the terms ‘machine’ and ‘think’. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. […] Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the ‘imitation game’. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either ‘X is A and Y is B’ or ‘X is B and Y is A’. The interrogator is allowed to put questions to A and B thus:

C: Will X please tell me the length of his or her hair? Now suppose X is actually A, then A must answer. It is A’s object in the game to try and cause C to make the wrong identification. His answer might therefore be

‘My hair is shingled, and the longest strands are about nine inches long.’

In order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten. The ideal arrangement is to have a teleprinter communicating between the two rooms. Alternatively the question and answers can be repeated by an intermediary. The object of the game for the third player (B) is to help the interrogator. The best strategy for her is probably to give truthful answers. She can add such things as ‘I am the woman, don’t listen to him!’ to her answers, but it will avail nothing as the man can make similar remarks.

We now ask the question, ‘What will happen when a machine takes the part of A in this game?’ Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman? These questions replace our original, ‘Can machines think?’

What we now call the Turing Test takes a much simpler, gender-neutral form; it simply involves a human judge trying to determine whether the entity at the other end of an online chat is another human or a machine. It’s telling, though, that in his original formulation, Turing’s challenge involved subverting what he regarded as the epitome of human judgment: making a gender determination without relying on physiological cues.

There’s also something very meta about a closeted gay mathematician in postwar Britain proposing to test whether a machine can pass for a man trying to pass for a woman.

Despite the increasingly common injunction never to make assumptions about the gender of someone we’ve just met, we do it all the time. But how? Hopefully you’re convinced by now that gender isn’t something that can be worked out by pure logic. It has to be a more holistic (and so, necessarily imperfect) judgment, based on the whole “fuse box” of gendered attributes and behaviors.

In Chapter 2, we explored handedness from a similarly multidimensional perspective, using techniques that allowed us to correlate dozens of questions about behaviors and presentation with personal identity as a left- or right-handed person. In this chapter, we’ll explore what happens when we carry out the same kind of multidimensional analysis for gender and sexual orientation. This also lets us zoom out from the “trees” of individual responses, which we’ve focused on so far, to see the whole forest.

When we build an “optimal linear estimator” to predict the answer to the question “Are you female?” based on answers to 37 of the survey’s behavioral questions, the following weights emerge, representing the importance of each response relative to the overall average:

The question weights for optimally predicting “Are you male?” are virtually identical, but with the positive and negative signs flipped. (Running the same analysis to predict pronoun— exclusively “she” or exclusively “he”— also yields a near-identical result.) There are no great surprises here. Which bathroom someone uses is the single most predictive variable, while long or short hair, manicures, and waxing are more weakly predictive.

Just as with handedness, we can also build a linear estimator to predict *ambiguous* responses to the gender questions— those who answer either “yes” to both, or “no” to both of “Are you female?” and “Are you male?” Now, hair length becomes far more relevant; short hair is the single most correlated question, while long hair is the most anti-correlated. Bathroom use is no longer so significant, though use of the women’s bathroom is somewhat more common than the men’s, and the use of urinals is nearly as uncommon as long hair; these typical answers suggest a desire for greater privacy in the bathroom among the gender-ambiguous. Other signs of androgyny are evident too, such as negative correlations with gendered names.

Clearly there are people in the middle, but is the middle the exception, or the norm, behaviorally speaking? Is the gender binary still alive and kicking? And are “both or neither” people a “third gender,” as has sometimes been claimed?

One way to tackle this question is to use a mathematical technique called Principal Component Analysis, or PCA. This involves taking the set of questions above— which are purely about socially visible behavior and presentation, and don’t include “Are you female?” or “Are you male?”— and finding the internal correlations among them.

It works as follows. Let’s define a “component” as just the kind of numerical score we’ve been calculating: a weighted sum of the answers to the 37 yes/no questions about gender presentation. The “first component” of the PCA is the single most informative score we can calculate, meaning that knowing this score for any individual would do the best possible job of estimating the answers to *all* of the questions. In theory, it could be the case that this one number would suffice to fully describe the data. Imagine, for instance, that we were working with a much smaller set of yes/no questions:

* Do you use the women’s bathroom?
* Do you use the men’s bathroom?
* Do you have long hair?
* Do you have short hair?

As we’ve seen over and over, we can never truly assume that a person’s response to one of these questions will determine the answer to any other one. There really are four variables here, with two potential answers each, allowing for 2×2×2×2=16 possibilities of varying popularity. But let’s suppose that we were administering this four-question quiz in an imaginary town— call it Stepford, USA— where everyone answers in only one of two equally likely ways: (+1, -1, +1, -1) for Stepford women, and (-1, +1, -1, +1) for Stepford men, where +1 represents “yes” and -1 is “no.” If we carry out a PCA,[[1]](#footnote-21) the first principal component would then have weights corresponding to (+1, -1, +1, -1), or equivalently, (-1, +1, -1, +1), or any other fixed multiple of these numbers, like (+0.5, -0.5, +0.5, -0.5); only the proportions of the weights relative to each other matter. For a stereotypical woman’s responses, this first principal component value would then work out to +0.5×1+(-0.5)×(-1) + 0.5×1 + (-0.5)×(-1) = +2, while the first principal component value for a stereotypical man’s responses would be +0.5×(-1)+(-0.5)×1+0.5×(-1)+(-0.5)×1 = -2. This single number, +2 or -2, would tell us how any given respondent answered all four questions.

But maybe not! Even in Stepford, a woman might decide to get a bob at the hairdresser’s. The occasional man might even grow his hair out. This would cause a *second* principal component to emerge, with weights (0.5, -0.5, -0.5, 0.5). As you can check for yourself, this component will be zero for long-haired women *or* short-haired men, but will produce a +2 for women with bobs, and a -2 for men with long hair. So in this slightly more complex (and slightly more realistic) situation where hair length doesn’t always coincide with the gender norm, our original four questions need to be represented by *two* numbers, not just one. The first, and most informative, still encodes gender, while the second reveals any gender-atypical variation in hair length.

So that’s Principal Component Analysis. Let’s now leave Stepford and enter the real world, where I’ve asked [[9,XXX]] Americans all 37 gender presentation questions. The first principal component that emerges looks almost indistinguishable from the gender prediction weights. Maybe this isn’t surprising, but it is meaningful, since gender identification didn’t come into this analysis. It tells us that in the real world, all of these behavioral markers of gender co-vary in a way that closely mirrors people’s identification with a gender.

Next, let’s look at a histogram of everyone’s first principal component value, that is, the sum of their answers multiplied by these weights.

There are two humps— it looks a lot like a gender binary! Though on closer inspection, we can see that this isn’t quite the full story. For something to be described as truly binary, it would have to be the case that it can *only* assume two values, like the +2 and -2 of our Stepford example. Here, we instead see a whole range or spectrum of values along this “gender axis,” from roughly -0.015 to +0.015, with no empty regions along that range. However, there are clearly two peaks, with a less-populated valley in between. The curve looks like a two-humped camel’s back.

Let’s compare this to a histogram of respondents’ heights. Men are on average taller than women, so overall, the distribution of human height also looks like a sum of two superimposed bell curves; however, because the random variations *within* the male and female distributions are large compared to the difference between them, there isn’t a noticeable dip in between: the 67” or so midpoint (5’7”) remains a very common height overall. Hence the sum looks more like a one-hump dromedary than a two-hump camel.

Still, while the gender presentation histogram is two-humped, those humps do overlap. So the question of whether gender is binary or not doesn’t have a clear answer. The debate is itself a false binary. Many people, in their presentation, appear to fall clearly into one peak or the other, and the very presence of those distinct peaks suggests categories; being great pattern recognizers and namers of things, we’d make up names for them if they didn’t already exist. However, because the peaks aren’t cleanly separated, those categories won’t apply to everyone.

But let’s check our assumptions. If you’re paying close attention, you may have noticed that we haven’t confirmed yet whether those two humps in the first principal component histogram actually correspond to gender identification. We can answer that question, and more, by visualizing respondents as points in [[2D/3D]] space, with their coordinates given by their first [[and second/, second, and third]] principal components— giving us something like a god’s eye view of every respondent in [[2D/3D]] “gender presentation space.” Then, we can color these points based on gender identification— female, male, both, or neither.

The histogram of respondents’ first principal components is what we’d get by letting all of these points fall onto the horizontal axis, and counting them up in bins. It’s easy to see that the overwhelming majority of the right-hand hump would consist of orange dots— people who identify exclusively as female— while the left-hand hump would consist almost entirely of blue dots— people who identify exclusively as male. In [[2D/3D]], these humps look like well-defined clusters. However, if we look more closely, we can see people whose gender identity we wouldn’t predict correctly based on how they present. There are orange dots surrounded by blue ones, and blue dots deep in orange territory. Moreover, there’s a smattering of “both” and “neither” across the entire range of presentations (these are drawn with bigger dots for extra clarity, since they’re scarcer).

Visualizations like these illustrate why the idea of a “third gender,” as a distinct cluster with specific social markers, has never really taken off. “Both” and “neither” points are scattered throughout the whole space, including— but not limited to— the “excluded middle” between conventionally masculine and feminine presentations. We can confirm this impression by graphing the likelihoods of female, male, or “both or neither” identification as a function of their score using the optimal linear predictors, as we did for left-, right-, and ambiguous-handedness. Female and male predictors are mirror images of each other, and show near-perfect performance at the extreme ends of the scale, with a transition in between where a greater proportion of “both or neither” people fall. However, significantly, there’s no score, even at the most ambiguous point on the scale where male or female identification are equally likely, where “both or neither” is the *likeliest* identification.

We don’t see anything like this strong performance when we build an optimal linear predictor for “both or neither”-ness. Even for a person maxing out on this score, the likelihood of identifying as “both or neither” is only 5% or so, with the balance of the probability more or less evenly split between female and male identification. An ambiguous “femaleness vs. maleness” score implies much higher odds of being “both or neither”— closer to 20%— but only because there are a lot fewer binary respondents in that excluded middle. Seen another way, there are more “both or neither” people who present the way traditionally female or male people do than there are in the excluded middle; it’s just that, in the ambiguous area, they stand out more clearly against the sparser binary background. So, “both or neither” people can look any which way. Most aren’t androgynous.

You may be wondering how to interpret “Principal component 2,” the vertical axis in the scatter plot. This second component seems relevant to gender identity too, though less straightforwardly. Notice how people whose first component is near zero— that is, who are most ambiguous on this dimension— are likelier to identify as female, and much likelier to identify as “neither,” if their second component is less than zero (toward the bottom), and likelier to identify as male if the second component is greater than zero (toward the top).

High values for this second component are correlated with living in the city and working in tech, while low values are correlated with living in the suburbs and having long hair. What does it all mean? There may be an interesting story here, but interpretation of these second-order effects can start to look more like data art than data science.

If we were able to create a dataset including all of the sex-relevant biological, physiological, and anatomical variables we discussed in Chapter 11 (on intersexuality), we’d almost certainly see the same two-hump pattern in the first principal component. Further, many of those variables (such as having larger breasts) are strongly correlated with behavioral variables (such as wearing a bra), for obvious reasons— though correlations like these can weaken in the disembodied, online world we increasingly inhabit. So, as we all relocate to the internet and gender fades in importance, the two-hump camel pattern may start to look more like the height histogram’s one-hump dromedary.

As long as we can see each other’s real life faces, though, even in a tightly cropped Zoom window, the overlapping but still clearly two-humped pattern of facial sexual dimorphism— the different average shapes of men’s and women’s faces— will remain obvious. We’ll also continue to make occasional misgendering mistakes, just as if we were to guess at the colors of dots on the survey PCA based on their position alone.

Recognizing gender and other attributes based on facial images is also a classic computer science problem. In 1990, Alex “Sandy” Pentland, a computer science professor at MIT and serial entrepreneur, dedicated his lab’s attention to face recognition. They used precisely the techniques we’ve just described for their analysis, simply replacing survey responses with image pixels. They began by aligning carefully posed and lit digital images of volunteers’ faces, and reducing the resolution to something manageable— 256×256 monochrome pixels, represented as numbers between 0 (black) and 1 (white). Averaging those images yields a somewhat blurry, indistinct “average face,” since everyone’s eyes, nose, and mouth are in slightly different positions.

Averages can also be made using smaller sets of more similar faces, to generate composites images of specific groups of people. This is what South African artist Mike Mike did years later in his itinerant photography project *The Face of Tomorrow*,[[2]](#footnote-22) which digitally superimposed portraits of men and women grouped by age and, albeit somewhat arbitrarily, by geography. By manually aligning the pupils of the eyes and other facial features, he was able to make his composites a lot sharper. The results are striking, and the faces are beautiful in their way, though also bland, since all asymmetries and distinctive features (including anything one might consider a blemish) have been averaged away.

Though it may not be immediately obvious, these composites are visualizations of exactly the kind of optimal linear estimators we’ve described for survey responses. Hence, the relative similarity of the face of a young Sydneysider to one or another of the gendered composites for young people in Sydney is a gender predictor.[[3]](#footnote-24)

This was the reasoning that led Victorian polymath Francis Galton (1822–1911), who also pioneered the statistical method of correlation, to invent the facial compositing technique— long before we had computers. Galton was less interested in gender, though, than in law and order. Like his contemporary and fellow traveler of Cesare Lombroso, Galton sought to characterize criminality. So, he set out to visually characterize “criminal types” by superimposing exposures of convicts on the same photographic plate.[[4]](#footnote-25) But is there really such a thing as an “average criminal face”?

Although there was real mathematical insight behind Galton’s idea, it suffered from many of the same problems as Lombroso’s pseudo-scientific studies of criminality. For starters, superimposing just a handful of exposures can’t yield anything like a real population average. A more serious issue, though, is that the selection of people included will determine what the resulting model “predicts”; as a statistician today might say, garbage in, garbage out. If, for instance, the 19th century Irish were economically disadvantaged, likelier to turn to crime, and on top of that likelier to be convicted by the British legal system, then the average “criminal face” might simply look Irish. By a feat of circular logic like many we’ve already encountered, this “scientific observation” could then be used justify anti-Irish prejudice. Predictive policing systems today have been built on the same flawed logic.[[5]](#footnote-26)

With computers, it’s possible to go beyond the averaged faces Galton could approximate using multiple exposures on a photographic plate (or Mike Mike’s more refined digital composites) and run a full Principal Component Analysis. The equivalent of the “question weights” for each principal component are a set of positive and negative weights on face pixels, each of which, when applied to an individual face image and added up, produces a number. Visualizing these weights, using black for the most negative, white for the most positive, and medium gray for zero, yields ghostly face-like images, which one can think of as systematically tweaking the average face— adding or subtracting hair in one case, making the forehead and cheekbones more or less prominent in another, and so on. In this way, just as we’ve shown how a questionnaire can be reduced to one or two numbers, a face photo can also be reduced to one or a handful of numbers. Forty numbers (corresponding to the first forty principal components) are plenty for uniquely identifying a face, and can easily distinguish feminine-looking faces from masculine-looking ones, much as the first PCA component of the survey does.[[6]](#footnote-27) Now we have not just a gender prediction model, but a general system for facial recognition. What could possibly go wrong?

[[ANIMATE ONLINE]]

Face recognition has legitimate uses. Our own brains (or most of them) certainly do it; it powers the capacity for individual recognition described in the Introduction. There’s nothing sinister about a smart camera that detects faces in the viewfinder to keep them in focus, or a smartphone with a face unlocking feature uniquely keyed to the phone’s owner.

However, early funding for face recognition didn’t come from phone or camera manufacturers, but from the Defense Department, which coordinated a number of institutions to accelerate its development, starting in 1993. Based on his lab’s promising early results, Sandy Pentland was tapped to help lead this effort. The goal was much the same as that of the revolutionary French government’s *Comité de Surveillance*, established in every municipality during the Reign of Terror: to develop techniques whereby the state could monitor the movements of outsiders, dissidents, and suspect individuals in general.[[7]](#footnote-28)

This French bureaucratic term, *surveillance*, from *sur* (over) and *veiller* (watch), has since entered the English language too. But whereas blanket surveillance of public spaces used to require vast, shadowy networks of informants, face recognition allows it to be automated. Today, the authorities’ “view from above” of citizen’s movements in security camera-saturated cities like Beijing or London far outdoes the analog police states of the 18th, 19th, and 20th centuries.

The technology has gotten a lot better, too. Manhattan-based startup Clearview AI uses deep learning, based on large neural networks rather than “linear” methods like PCA,[[8]](#footnote-30) to provide the FBI, Immigration and Customs Enforcement (ICE), the Department of Homeland Security (DHS), Interpol, and many local police departments (as well as various corporate customers) with a truly powerful face recognition engine.[[9]](#footnote-31) The many successive layers of a deep learning system make it far more robust, so that face images don’t need to be aligned, uniformly posed or lit; such systems work surprisingly well even when faces are partially hidden, made up, or disguised.[[10]](#footnote-33) Trained on billions of face images scraped from the web and social media, Clearview’s system is said to be able to reliably recognize anyone with an online presence from surveillance photos or video.[[11]](#footnote-34)

Perhaps even more troubling are applications of face recognition that don’t purport to recognize individuals, but rather to make inferences about them based on how they look. Some of these systems revive Francis Galton’s long-dormant fantasy of recognizing “criminal types.” In 2016, for instance, two machine learning researchers, Xiaolin Wu and Xi Zhang, put a paper online, *Automated Inference on Criminality Using Face Images*,[[12]](#footnote-36) claiming that their neural net could predict the likelihood that a person was a convicted criminal with nearly 90% accuracy using nothing but a driver’s license-style face photo. Their training and testing data consisted of 730 images of men wanted for or convicted of a range of crimes by provincial and city authorities in China, as well as 1,126 face images of “non-criminals […] acquired from Internet using [a] web spider tool.”

Although we can’t be sure what their neural net is *really* picking up on, given opaque data sources, some clues are offered by the three examples each of “criminals” and “non-criminals” shown in the paper. For one, the “non-criminals” all wear white-collar shirts, while the “criminals” don’t. The main difference, though, is likely to be the frowny faces of the “criminals,” and the smiley faces of the “non-criminals” (this is also consistent with measurements the authors offer of average differences in facial landmark positions between the two groups). That’s right: the “criminals” look meaner, and the “non-criminals” look nicer.

We develop biases against mean-looking people early in childhood. Even 3- and 4-year olds can reliably distinguish “nice” from “mean” face images.[[13]](#footnote-37) Physiognomists, too, have long held the perfectly reasonable-sounding position that nice people have nice faces, and mean people have mean faces.

What do these stereotypically “nice” and “mean” faces look like? Researchers studying the social perception of faces in recent years[[14]](#footnote-38) have used our old friend, Principal Component Analysis, to explore not only this question, but how people’s first impressions of faces work in general. In one experiment, a video game-like face renderer produces a wide range of artificial faces from 50 numerical measurements; experimental participants then describe their impressions of randomly generated faces, or rate them according to some perceived trait. PCA can capture most of the variability in people’s first impressions using just a few components, three of which the researchers have named “dominance,” “attractiveness,” and “valence.” The first two are fairly self-explanatory. The third, “valence,” is associated with positive impressions like “trustworthy” and “sociable.”

An optimal linear predictor lets us visualize what a stereotypically “trustworthy” or “untrustworthy” face looks like, as shown here for white males. Notice how the “trustworthy” face has a more smiley expression— and reads as more feminine— while the “untrustworthy” face is more frowny and more masculine. It’s likely no coincidence that the supposedly “criminal” faces in Wu and Zhang’s paper look a lot like the “untrustworthy” face, while the “non-criminal” faces are more like the “trustworthy” face. Galton’s composites of “criminal types” also look more like the “untrustworthy” face— as does the downturned mouth of “Prendergast, the Murderer.”

People make instinctive, automatic judgments about character traits like trustworthiness after seeing a face for less than one tenth of a second. Author Malcolm Gladwell wrote effusively about snap judgments like these in his 2005 bestseller *Blink: The Power of Thinking Without Thinking*. It does seem that such judgments do a remarkable job of predicting important life outcomes,[[15]](#footnote-39) ranging from political elections to economic transactions to legal decisions.

The question is, are these *good* judgments? Either answer would be interesting. If they *are* good judgments, then the physiognomists are right after all— we really do advertise our character traits on our faces. On the other hand, if those judgments are arbitrary, then the fact the they do such a good job of predicting outcomes tells an even more disturbing story: that unwarranted prejudice based on resting facial appearance plays a decisive role in human affairs.

A large body of research points to the latter.[[16]](#footnote-40) For example, in 2015 Brian Holtz of Temple University published the results of a series of experiments[[17]](#footnote-41) in which face “trustworthiness” was shown to strongly influence experimental participants’ judgment. The participants were asked to decide, after reading a 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. In another study,[[18]](#footnote-42) 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 more objective reputational information about the past investment behavior of their partners. Even more chillingly, another study[[19]](#footnote-43) found that among prisoners convicted of first degree murder, the unlucky defendants 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.

So, we’re not endowed with some *Blink*-like genius for making snap judgments about people’s character based on how they look. Unfortunately, we make those snap judgments anyway. This is sometimes to our own detriment; for instance, studies in which the trustworthiness of economic behavior was measured show that relying on face judgments can make our decisions *less* reliable than if we were making them, literally, blind.[[20]](#footnote-44) The worst consequences, though, fall on anyone who happens to have an “untrustworthy” face. For example, it seems that they’ll be significantly likelier, all things being equal, to end up convicted of a crime, while a guilty party with a “trustworthy” face is likelier to get away with their misdeeds. This also means that a facial “criminality detector” tells us more about our own biases than about the subject’s character.

In the Fall of 2017, a year after the “resting criminal face” paper came out, a higher-profile study began making the rounds claiming that sexual orientation could be reliably guessed based on a facial image. The Economist featured this work on the cover of their September 9th magazine; on the other hand two major LGBTQ organizations, The Human Rights Campaign and GLAAD, immediately labeled it “junk science.” Michal Kosinski, who co-authored the study with fellow researcher Yilun Wang, initially expressed surprise, calling the critiques “knee-jerk” reactions. However, he then doubled down with even bolder claims: that similar approaches will soon be able to measure the intelligence, political orientation, and criminal inclinations of people from their faces alone.[[21]](#footnote-45)

Once again, this echoes the claims of Cesare Lombroso and Richard von Krafft-Ebing: physiognomy, now dressed in the “new clothes” of machine learning. However, while Wu and Zhang seemed well outside the mainstream in tech and academia, Kosinski is a faculty member of Stanford’s Graduate School of Business, and this new study was accepted for publication in the respected Journal of Personality and Social Psychology. Much of the ensuing scrutiny focused on ethics, implicitly assuming that the science was valid. As with the previous year’s result, however, there’s reason to believe otherwise.

The authors trained and tested their “sexual orientation detector” using 35,326 images from public profiles on a US dating website. Composite images of the lesbian, gay, and straight men and women in the sample,[[22]](#footnote-46) reminiscent of Galton’s composites, reveal a great deal about the information available to the algorithm.

According to Wang and Kosinski, the key differences between these composite faces are in their physiognomy, meaning that sexual orientation tends to correlate with a characteristic facial structure. However, you’ll notice right away that some of these differences are more superficial. For example, the “average” straight woman appears to wear eyeshadow, while the “average” lesbian doesn’t. Glasses are clearly visible on the gay man, and to a lesser extent on the lesbian, while they seem absent in the heterosexual composites. Might it be the case that the algorithm’s ability to detect orientation has less to do with facial structure than with grooming, presentation, and lifestyle?

I was already running the Mechanical Turk surveys that year that would eventually turn into this book, so I teamed up with a couple of colleagues, Margaret Mitchell, then at Google, and Alex Todorov, then faculty at Princeton’s Psychology Department, to run an extra survey that might shed some light on this.[[23]](#footnote-47) In addition to the usual questions about sexual orientation and attraction, we added a few, like “Do you wear eyeshadow?” and “Do you wear glasses?” “Do you have a beard or moustache?” and “Do you ever use makeup?” were already on there. These graphs break down responses by sexual orientation, based both on behavior and identity. For instance, “Opposite-sex attracted women” are people who answer “yes” to “Are you female?,” “no” to “Are you male?,” and “yes” to both “Are you sexually attracted to men?” and “Are you romantically attracted to men?”; a nearly identical subset of women answer “yes” to “Are you heterosexual or straight?” “Same-sex attracted women” are those who answer “yes” to either or both of “Are you sexually attracted to women?” and “Are you romantically attracted to women?”

It will surprise no one that straight women tend to wear more makeup and eyeshadow than same-sex attracted and (even more so) lesbian-identifying women— although seen another way, the curves also show us that these stereotypes are often violated. That same-sex attracted people of most ages wear glasses significantly more than exclusively opposite-sex attracted people do might be less obvious, but it is so. A physiognomist might cook up a theory about gay people having worse eyesight, but answers to the question “Do you like how you look in glasses?” reveals that this is probably more of a fashion statement. The pattern holds both for women and for men.

Gay and same-sex attracted men under the age of 45, who contributed the majority of the facial images in the dataset, are also less likely to have beards than those who are opposite-sex attracted. In their paper, Wang and Kosinski speculate that the relative faintness of the beard and moustache in their gay male composite might be due to feminization due to prenatal underexposure to androgens (male hormones). But the fact that older gay men have beards *more often* than older straight men tells a different story, in which fashion trends and cultural norms are the determining factor.

Similarly, the paper notes that the heterosexual male composite has darker skin than the other three composites. Once again, the authors reach for a hormonal explanation: “While the brightness of the facial image might be driven by many factors, previous research found that testosterone stimulates melanocyte structure and function leading to a darker skin.” A simpler explanation: opposite-sex attracted men are 29% more likely to work outdoors, and among men under 31, this rises to 39%; of course, spending time in the sun darkens skin.

None of this proves that there’s no physiological basis for sexual orientation; on the contrary, lots of evidence shows us that orientation runs much deeper than a presentation choice or a “lifestyle.” (For one thing, if it *were* simply a “lifestyle,” gay conversion therapy wouldn’t be such a spectacular failure.) It follows that if we dig deeply enough into human physiology and neuroscience, we may eventually find factors that reliably correlate with sexual orientation. The survey has little to offer here, but there’s one tantalizing clue: very tall women are significantly overrepresented among lesbian-identifying respondents. However, while this is interesting, it’s very far from a good *predictor* of women’s sexual orientation.

To get a sense of why, we should look at the numbers. The way Wang and Kosinski measure the performance of their “AI gaydar” is equivalent to choosing a straight and a gay or lesbian face image, both from data “held out” during the training process, and asking how often the algorithm correctly guesses which is which. 50% performance would be no better than random chance. For women, guessing that the taller of the two is the lesbian achieves only 51% accuracy— barely above random chance. This is because, despite the statistically meaningful overrepresentation of tall women among the lesbian population, the great majority of lesbians are *not* unusually tall.

By contrast, the performance measures in the paper, 81% for gay men and 71% for lesbian women, seem impressive.[[24]](#footnote-48) However, we can get comparable results simply by using a handful of yes/no survey questions about presentation. For example, for pairs of women, one of whom is lesbian, the following trivial algorithm is 63% accurate: if neither or both women wear eyeshadow, flip a coin; otherwise guess that the one who wears eyeshadow is straight, and the other lesbian. Making an optimal linear estimator using six more yes/no questions about presentation (“Do you ever use makeup?,” “Do you have long hair?,” “Do you have short hair?,” “Do you ever use colored lipstick?,” “Do you like how you look in glasses?,” and “Do you work outdoors?”) as additional signals raises the performance to 70%. Given how many more details about presentation are available in a face image, 71% performance no longer seems so impressive.

Several studies[[25]](#footnote-49) have shown that human judges’ “gaydar” is no more reliable than a coin flip when the judgment is based on pictures taken under well-controlled conditions (head pose, lighting, glasses, makeup, etc.). It’s well above chance if these variables *aren’t* controlled for, because a person’s presentation— especially if that person is out— involves social signaling.

Wang and Kosinski argue against this interpretation on the grounds that their algorithm works on Facebook selfies of users who make their sexual orientation public as well as dating website profile photos. The issue, however, isn’t whether the images come from a dating website or Facebook, but whether they are self-posted or taken under standardized conditions. In one of the earliest “gaydar” studies using social media,[[26]](#footnote-50) participants could categorize gay men with about 58% accuracy in a fraction of a second; but when the researchers used Facebook images of gay and heterosexual men posted by their friends (still an imperfect control), the accuracy dropped to 52%.

From a human rights perspective, this is fortunate, since as of 2022, homosexuality remains a criminal offense in more than 70 countries. In a number of them (Iran, Saudi Arabia, Yemen, Afghanistan, Brunei, Mauritania, Pakistan, Qatar, the United Arab Emirates, Nigeria, Sudan, and Somalia), it may be punishable by execution. Hence, in these countries, an effective “gaydar camera” would in fact amount to a dragnet for rounding up gay people, who are defined as “criminals.” We should remember that homosexuality was also criminalized in most Western countries until very recently. In England, it was punishable by death under the Buggery Act of 1533, which codified “sodomy” as “the detestable and abominable vice of buggery”; the 1861 Offences against the Person Act, modified the punishment to life imprisonment, which remained in force until 1967. Famously, Oscar Wilde and Alan Turing were both convicted under this law. Turing, at the time both a computer science pioneer and a war hero for his seminal work on cracking the Nazis’ Enigma codes, was sentenced to so-called “chemical castration” (estrogen injections, like those David Reimer underwent) as an alternative to prison. He died in his forties, like Oscar Wilde, possibly of suicide. Turing was only officially “pardoned” by Queen Elizabeth in 2013.

If subtle biases in image quality, expression, and grooming can be picked up on by humans, these biases can also be detected by AI. While Wang and Kosinski acknowledge grooming and style, they believe that the chief differences between their composite images relate to face shape, arguing that gay men’s faces are more “feminine” (narrower jaws, longer noses, larger foreheads) while lesbian faces are more “masculine” (larger jaws, shorter noses, smaller foreheads). As with less facial hair on gay men and darker skin on straight men, they suggest that the mechanism is gender-atypical hormonal exposure during development. This echoes the widely discredited 19th century model of homosexuality, “sexual inversion,” as illustrated by Krafft-Ebing’s “Count Sandor V” case study among many others.

More likely, though, this is a matter of shooting angle. A 2017 paper on the head poses heterosexual people tend to adopt when they take selfies for Tinder profiles is revealing.[[27]](#footnote-51) In this study, women are shown to be about 50% likelier than men to shoot from above, while men are more than twice as likely as women to shoot from below. Shooting from below will have the apparent effect of enlarging the chin, shortening the nose, shrinking the forehead, and attenuating the smile. This view emphasizes dominance— or, maybe more benignly, an expectation that the viewer will be shorter. On the other hand, shooting from above simulates a more submissive posture, while also making the face look younger, the eyes bigger, the face thinner, and the jaw smaller— though again, this can also be interpreted as an expectation or desire[[28]](#footnote-52) for a potential partner (the selfie’s intended audience) to be taller. If you’re seeking a same-sex partner, the partner will tend to be of the same height, so we’d expect the average shot to be neither from below nor from above, but straight on.

And this is just what we see. Heterosexual men on average shoot from below, heterosexual women from above, and gay men and lesbian women from directly in front.[[29]](#footnote-53) Notice that when a face is photographed from below, the nostrils are prominent (as in the heterosexual male face), while higher shooting angles hide them (as in the heterosexual female face). A similar pattern can be seen in the eyebrows: shooting from above makes them look more V-shaped, but they get flatter, and eventually caret-shaped (^) as the camera is lowered. Shooting from below also makes the outer corners of the eyes seem to droop. In short, the changes in the average positions of facial landmarks match what we’d expect to see from differing selfie angles.

[[MORPH]]

Let’s turn this observation on its head: might the human smile and frown have their origins in dominant (taller, looking downward) and submissive (shorter, looking upward) head poses? If so, these expressions may originally have evolved, at least in part, to mimic the way facial appearance varies based on height and posture.[[30]](#footnote-55) Indeed, there’s strong evidence that our smile is the human version of the “bared teeth face” or “fear grin” in monkeys, associated with submission or appeasement.[[31]](#footnote-56) Submissive postures also tend to involve crouching and making one’s body look small, while dominant postures involve looking big and towering over others, or at least using head angle to *seem* taller— hence the expression “looking down your nose at someone.”

Whatever the reasons, when we look at human faces today— especially static, two-dimensional photos of strangers taken under uncontrolled conditions— there’s visual ambiguity between head pose, the shapes of facial features, and affect (i.e. smiling or frowning).

In the heterosexual context, this may also explain the stereotypically more feminine look of the average “nice” or “trustworthy” face, and the more masculine character of the average “mean” or “untrustworthy” face. As researchers put it in a 2004 paper, *Facial Appearance, Gender, and Emotion Expression*,[[32]](#footnote-57)

[A] high forehead, a square jaw, and thicker eyebrows have been linked to perceptions of dominance and are typical for men’s faces […], whereas a rounded baby face is both feminine and perceived as more approachable […] and warm […], aspects of an affiliative or nurturing orientation. This leads to the hypothesis that— regardless of gender— individuals who appear to be more affiliative are expected to show more happiness, and individuals who appear to be more dominant are seen as more anger prone. As cues to gender and cues to affiliation/dominance are highly confounded, this would lead to more women being expected to be happiness prone and more men to be anger prone.

This brings us back to physiognomy. A woman with what appears to be an unsmiling, perhaps defiant expression, photographed in V.G. Rocine’s 1910 physiognomy primer *Heads, Faces, Types, Races*, supposedly illustrates “Degeneracy seen in [the] eyes, eyebrows, nose, lips, jaw, hair and pose.” As with a collection of similar mugshots from Lombroso’s *The criminal woman*, a great majority of the female “crime” in question was either nonviolent and associated with poverty, or just reflected behavior that didn’t conform to the prevailing gender and sexual norms. With so many biases in play, it’s hard to know which was judged first— women’s faces, or their actions.

1. Calculating the PCA is a bit like long division from hell— it requires a complex, many-step algorithm, which wasn’t practical to do before we had computers. If you’re interested in the mechanics, though, see every old school data scientists’s reference book, Numerical Recipes [[REF]]. [↑](#footnote-ref-21)
2. <https://artbase.rhizome.org/wiki/Q2002> [↑](#footnote-ref-22)
3. Technically, one would need to subtract from a selective composite the overall average face to produce the kind of optimal linear estimator we’ve described. Then, producing an estimate involves multiplying the resulting pixel weights by the difference between an unknown portrait’s pixels and the overall average. [↑](#footnote-ref-24)
4. Galton, *Composite portraits*, Nature 18 (1878): 97-100. [↑](#footnote-ref-25)
5. [[REF]] Blaise Meg Alex AI criminality 2018 Medium. [↑](#footnote-ref-26)
6. [[REF]] Pentland [↑](#footnote-ref-27)
7. Per <https://www.nist.gov/programs-projects/face-recognition-technology-feret>, “The goal of the FERET program was to develop automatic face recognition capabilities that could be employed to assist security, intelligence, and law enforcement.” [↑](#footnote-ref-28)
8. Linear predictors and PCA can be thought of as single-layer neural networks. [↑](#footnote-ref-30)
9. <https://www.cnet.com/news/privacy/clearview-ai-facial-recognition-customers-reportedly-include-ice-justice-department-fbi-macys/> [↑](#footnote-ref-31)
10. [[REF]] [↑](#footnote-ref-33)
11. <https://www.washingtonpost.com/technology/2022/02/16/clearview-expansion-facial-recognition/> [↑](#footnote-ref-34)
12. [[REF]]. [[ALSO THEIR REBUTTAL]]. [↑](#footnote-ref-36)
13. Cogsdill, Todorov, Spelke, Banaji, *Inferring character from faces: A developmental study*, Psychol Sci 25, no. 5 (2014): 1132-1139. [↑](#footnote-ref-37)
14. Oosterhof, Todorov, *The functional basis of face evaluation*, PNAS 105, no. 32 (2008): 11087-11092. [↑](#footnote-ref-38)
15. Olivola, Funk, Todorov, *Social attributions from faces bias human choices*, TICS 18, no. 11 (2014): 566-570. [↑](#footnote-ref-39)
16. Todorov, Olivola, Dotsch, Mende-Siedlecki, *Social attributions from faces: Determinants, consequences, accuracy, and functional significance*, Annu Rev Psychol 66 (2015): 519-545. [↑](#footnote-ref-40)
17. Holtz, *From first impression to fairness perception: Investigating the impact of initial trustworthiness beliefs*, Pers Psychol 68, no. 3 (2015): 499-546. [↑](#footnote-ref-41)
18. Rezlescu, Duchaine, Olivola, Chater, *Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior*, PloS One 7, no. 3 (2012): e34293. [↑](#footnote-ref-42)
19. Wilson, Rule, *Facial trustworthiness predicts extreme criminal-sentencing outcomes*, Psychol Sci 26, no. 8 (2015): 1325-1331. [↑](#footnote-ref-43)
20. Efferson, Vogt, *Viewing men’s faces does not lead to accurate predictions of trustworthiness*, Sci Rep 3, no. 1 (2013): 1-7. [↑](#footnote-ref-44)
21. *Face-reading AI will be able to detect your politics and IQ, professor says*, The Guardian, 12 September 2017. [↑](#footnote-ref-45)
22. Clearly this is an incomplete accounting of sexual orientations, as well as presuming a gender binary. [↑](#footnote-ref-46)
23. Margaret is an AI researcher, and Alex is a world expert on the social perception of faces. He’s an author on several papers cited in this chapter, and published a book on this topic in 2018, *Face Value: The Irresistible Influence of First Impressions*. [↑](#footnote-ref-47)
24. These figures rise to 91% for men and 83% for women if 5 images are considered. [↑](#footnote-ref-48)
25. Cox, Devine, Bischmann, Hyde, *Inferences about sexual orientation: The roles of stereotypes, faces, and the gaydar myth*, J Sex Res 53, no. 2 (2016): 157-171. [↑](#footnote-ref-49)
26. Rule, Ambady, *Brief exposures: Male sexual orientation is accurately perceived at 50 ms*., J Exp Soc Psychol 44, no. 4 (2008): 1100-1105. [↑](#footnote-ref-50)
27. Sedgewick, Flath, Elias, *Presenting your best self (ie): The influence of gender on vertical orientation of selfies on Tinder*, Front Psychol 8 (2017): 604. [↑](#footnote-ref-51)
28. Fink, Neave, Brewer, Pawlowski, *Variable preferences for sexual dimorphism in stature (SDS): Further evidence for an adjustment in relation to own height*, Pers Individ Differ 43, no. 8 (2007): 2249-2257. [↑](#footnote-ref-52)
29. Although the authors use a face recognition engine designed to try to cancel out effects of head pose and expression, we have confirmed experimentally that this doesn’t work, a finding replicated by Tom White, a researcher at Victoria University in New Zealand (<https://twitter.com/dribnet/status/908521750425591808>). [↑](#footnote-ref-53)
30. Parr, Waller, *Understanding chimpanzee facial expression: insights into the evolution of communication*, Soc Cogn Affect Neurosci vol. 1,3 (2006): 221-8. [↑](#footnote-ref-55)
31. Robin Dunbar, *Friends*, 2021. [↑](#footnote-ref-56)
32. Hess, Adams Jr, Kleck, *Facial appearance, gender, and emotion expression*, Emotion 4, no. 4 (2004): 378. [↑](#footnote-ref-57)