Chapter 11

AI Bias

Nijeer Parks stands at a Western Union located inside a pharmacy in Haledon, New Jersey. Thirty miles away, another man who the facial algorithm decides looks like him, is shoplifting snacks at a Hampton Inn. The police are called to the hotel where they confront the suspect, a black man. They run his identification and find it to be fraudulent. Officers report that they spot a bag in his jacket pocket that they suspect to be marijuana. At that time, marijuana was illegal in New Jersey, and as they try to handcuff him, he makes a run for it and gets away in his rental car, sideswiping the police cruiser in the process. One officer reports having to jump out of the way to avoid being hit as the car sped by and scraped a column of the hotel. Later the police locate the vehicle, abandoned, about a mile away. The man has committed felonies. The police have an ID—the details are bogus, but the picture appears correct. The police run it through facial recognition software. The next day, the software comes up with a match: Nijeer Parks.

Parks is arrested. He spends 10 days in jail and pays around $5,000 to defend himself even though at the time of the crime he was 30 miles away—a fact confirmed by multiple data points, including Western Union records. Although he knows he is innocent, Parks also knows that a decade earlier he had two arrests and convictions for selling drugs. Per sentencing guidelines, a third conviction could result in a 10-year prison sentence for him. Overall, the conviction rate for a third offense in the U.S. is more than 90 percent, according to the Justice Department. He considers taking a plea deal. "I sat down with my family and discussed it," Parks said. "I was afraid to go to trial. I knew I would get 10 years if I lost."

Eventually Parks was able to get proof from Western Union that he had been sending money inside a pharmacy more than 30 miles away when the incident happened. At his last court hearing he told the judge he was willing to go to trial to defend himself. A few months later, his case was dismissed. "I was locked up for no reason," Parks said. "I've seen it happen to other people. I've seen it on the news. I just never thought it would happen to me. It was a very scary ordeal."

In January 2020, Robert Williams returned home to his family, his wife Melissa and his two young daughters, in a peaceful suburb in Farmington Hills, Michigan. An hour earlier at work, he had received a strange call. The person on the line said they were from the Detroit Police Department and he needed to present himself at the station to be arrested. He thought it was a prank. But as he pulled into his driveway, a police cruiser pulled up behind his car, blocking him in. Moments later, two officers handcuffed Williams on his front lawn, in front of his wife and daughters. His daughters are distraught, especially his youngest. He told her he'll be right back. The police wouldn't tell him or his wife why he is being arrested. They showed a piece of paper with his photo and the words "felony warrant" and "larceny." They won't tell his wife where they are taking him. They tell her, "Google it."

According to Kashmir Hill's newspaper reporting in June 2020, the police drove Mr. Williams to a detention center. He had his mug shot, fingerprints, and DNA taken and was held overnight. Around noon the next day, two detectives took him to an interrogation room and placed three pieces of paper on the table, face down.

One detective turned over the first piece of paper. It was a still image from a surveillance video, showing a heavyset man, dressed in black and wearing a red St. Louis Cardinals cap, standing in front of a watch display from which five timepieces, worth $3,800 in total, were shoplifted. "Is this you?" asked the detective. The detective turned over a second piece of paper—a close-up. The photo was blurry, but it was clearly not Williams. He picked up the image and held it next to his face. “No, this is not me,” Williams said. “You think all black men look alike?”

The detective turned over the third piece of paper, which was another photo of the man from the watch store next to Williams’s driver’s license. Williams again pointed out that the pictures were not of the same person.

Per Williams’s recollection, after he held the surveillance video still photo next to his face, the two detectives leaned back in their chairs and looked at one another. One detective, seeming chagrined, said to his partner: “I guess the computer got it wrong.”

Williams asked if he was free to go. “Unfortunately not,” one detective said. Williams was kept in custody until that evening, 30 hours after being arrested, and released on a $1,000 personal bond. He waited outside in the rain for 30 minutes until his wife could pick him up. When he got home at 10 p.m., his five-year-old daughter was still awake. She said she was waiting for him because he had said, while being arrested, that he’d be right back.

The police-required interrogation caused Williams to miss work—breaking his four-year perfect-attendance record. Two weeks after his arrest, Williams took a vacation day to appear in a Wayne County court for arraignment. When the case was called, the prosecutor moved to dismiss. No apology or restitution was made.

The experience has been so traumatic for his daughter, he is considering therapy for her. She has since taken to playing “cops and robbers” and accuses her father of stealing things, insisting on “locking him up” in the living room.

In response to Hill’s article, the Wayne County prosecutor’s office issued a two-page statement noting their new policy regarding facial recognition, as well as an apology to Williams. The last paragraph reads:

> Prosecutor Worthy said: In the summer of 2019, the Detroit Police Department asked me personally to adopt their Facial Recognition Policy. I declined and cited studies regarding the unreliability of the software, especially as it relates to people of color. They are well aware of my stance and my position remains the same. Any case presented to my office that has utilized this technology must be presented to a supervisor and must have corroborative evidence outside of this technology. This present case occurred prior to this policy. Nevertheless, this case should not have been issued based on the DPD investigation, and for that we apologize. Thankfully, it was dismissed on our office's own motion. This does not in any way make up for the hours that Mr. Williams spent in jail.

A national study in 2019 of over 100 facial recognition algorithms found that they were from 10 times to 100 times more likely to be wrong on Black and Asian faces than faces of white subjects. Several academic studies have documented the pitfalls of using AI facial recognition. Clare Garvie, a Distinguished Fellow at the Center on Privacy and Technology at Georgetown, has written extensively about the problems with facial recognition. She argues that low-quality search images, such as a still image from a grainy surveillance video, should be banned—and that the systems currently in use should be tested rigorously for accuracy and bias.

Because of these concerns, some jurisdictions, such as San Francisco, California, have banned the use of AI facial recognition for law enforcement. Others, such as the state of Florida, continue to make extensive use of it daily. Florida's system has been in operation for two decades, and just this year, in 2020, the state will process about 55,200 queries of its facial recognition database for Florida law enforcement officers. According to Florida's official tally, there are a little more than 400 success stories based on their two decades of use. There is no tally for the number of false arrests resulting from its use.

Garvie notes "It's really being sold as this tool, accurate enough to do all sorts of crazy stuff…. It's not there yet." Among the documented problems is law enforcement's over-reliance on the system to make arrests. For example, although Florida officials said investigators could ***not*** rely on facial recognition results to make an arrest, *The New York Times* reports that "...documents suggested that on occasion officers gathered no other evidence."

There is risk in using AI, but there is also value. Law enforcement and security organizations are reporting value and time savings from AI facial recognition. The Security Industry Association (SIA) noted that, since 2015, the nonprofit group Thorn has provided a tool called Spotlight, which uses facial recognition among other technologies to help investigators find underage sex trafficking victims in online ads. Spotlight has reportedly been used in 40,000 cases in North America, helping rescue 15,000 children and identify 17,000 traffickers. In addition to fighting human trafficking, SIA notes successful cases in airport security catching those travelling with fraudulent passports, counterterrorism, health care, and law enforcement investigations, citing specific cases of success.

Facial recognition is but one example of where bias in AI becomes apparent. Every decision maker should appreciate how deeply bias permeates AI applications, and where this bias can produce harm.

The Zoom-in Problem

In considering the value of AI vs. the risks, it is helpful to understand the zoom-in and zoom-out problems with AI. At the crux of what Parks and Williams experienced is what we call the *zoom-in* problem. It occurs when taking a large population of data and *zooming-in* to the matched individual. When AI generates a false positive and matches the wrong individual, such as in a criminal justice scenario, it is a serious problem that requires strong countermeasures to offset possible harms. We offer countermeasures in the next chapter.

The Zoom-out Problem

Beyond the false matches resulting from zoom-in challenges, another manifestation of AI bias is what we call the *zoom-out* problem. In the *zoom-out* situation, we notice bias when we add up the individual AI decisions and see that using AI to make individual decisions has, in aggregate, produced something fundamentally different (perhaps unfair) compared to what we, reasonably, expected.

For example, what would happen if an AI handled the recruitment of people for different jobs? We would expect AI-based recruitment to generate a diverse list of candidates that are qualified for the job—the candidates would be consistent with the law that forbids discrimination by age, sex, race, religion, and disabilities. In the U.S., the Civil Rights Act of 1964 bans employment discrimination based on race, color, religion, sex, or national origin. The law defines these as "protected classes." Many other countries have similar laws. The U.K., for example, extended its anti-discrimenation protections with the Equality Act of 2010 to guard against discrimination on the basis of sex/gender, age, race, religion/beliefs, pregnancy, maternity, disability, sexual orientation, marriage/civil partnership, and gender reassignment. But Today's AI does not have a way to embed such laws in its behavior.

When Global Witness, a non-profit focused on protecting human rights, conducted a test of Facebook's recruitment advertising, it found that Facebook's AI did not generate a list of diverse candidates consistent with the law. Global Witness requested that Facebook serve ads for four real job openings (mechanics, preschool nurses, pilots, and psychologists) in the U.K. It used the "Traffic/Link Clicks" objective—which, according to Facebook, ensures the ads are delivered to "the people who are most likely to click on them." They specified no further targeting criteria. Facebook's algorithm was completely in control of who was being shown the ads. The results demonstrated significant gender bias:

- 96% of the people shown the ad for mechanic jobs were men.
- 95% of those shown the ad for preschool nurse jobs were women.
- 75% of those shown the ad for pilot jobs were men.
- 77% of those shown the ad for psychologist jobs were women.

According to the September 2021 report by Global Witness, these ads violated Facebook's own rules and are part of broader bias concerns with the Facebook Algorithm.

"Our evidence tallies with the findings of Algorithm Watch and academics who have also shown that Facebook's ad delivery algorithm is highly discriminatory in delivering job ads in France, Germany, Switzerland, and the U.S. Recent investigations in the U.S. have shown that Facebook's ad delivery system is skewed by gender, potentially excluding swathes of women from seeing job opportunities, even when they are equally as qualified as the men that are being selected by Facebook to see certain ads," the report noted. In addition to gender bias, the report showed how Facebook's own ads discriminated against those over 55 years of age—even though they were qualified.

Facebook has already faced and settled five discrimination lawsuits from civil rights and labor organizations, workers, and individuals. The allegations claimed Facebook's ad systems *excluded* certain people from seeing housing, employment, and credit ads based on their age, gender, or race. *CNBC* reported, in 2019, that Facebook said it would launch a different advertising portal for housing, employment, and credit ads on Facebook, Instagram, and Messenger. The portal would offer significantly *fewer* targeting options. For example, advertisers would ***not*** be able to target users by characteristics like gender, age, religion, race, ethnicity, or zip code. By the time of the Global Witness test, Facebook/Meta had made good on this commitment. But, ironically, removing the ability of advertisers to target messages by age, gender, religion, race, and ethnicity did not solve the bias problem; it simply made it less visible on the surface. Sometimes the solution to bias isn't excluding or ignoring the variables of bias but rather acknowledging the variables and using AI to offset them. As we *zoom out* from the AI deciding which individuals will get which ads, we can see, in aggregate, severe bias in the outcome.

In advertising, there are two elements that can manifest bias. First is *delivery*. This occurs when AI systematically serves ads to one group over another. This is the problem Facebook was sued over and settled with the U.S. government in 2019. This led Facebook to remove certain ad targeting criteria, e.g. age and sex, for certain types of advertising, including employment and housing ads, in their ad manager portal. It was still possible to let the AI optimize the campaign based on who clicked on the ads, and this led to the problem that Global Witness found in 2021. The AI could act in biased ways because there was no parameter in the AI algorithm to ensure a *fair balance* of advertising delivery to the protected classes. A solution is to include variables related to protected classes so that the AI can ensure fairness in the delivery.

It is important to consider that there may be deeply embedded biases that are not as easy to observe or fix. In the book, *The Alignment Problem*, Brian Christian provides a wonderfully written detailed account of deeply embedded bias in Today's AI. He details how a tech company (Amazon) used AI to screen candidates for its company. It loaded in a decade's worth of employee performance and set the AI on the task of finding candidates that could perform well. But, because the tech firm had hired and promoted mostly males in the past, the AI perpetuated this pattern. In his example of gender bias in filtering job applications, he notes that the initial solution was to remove gender as a declared variable and remove the names—yet the AI still acted in a biased way because there are other signifiers embedded in the word selection or activities that correlate with gender. This bias was more deeply embedded than could be seen by the human eye.

Christian also cites examples of deeply embedded bias in large language models like GPT-3. The bias exists in the body of content fed to the AI and becomes apparent when using vector math to query the AI. He recounts how researchers found biased associations such as asking the AI to compute “Pilot” minus “Male.” The AI’s answer seemed sexist: According to the AI,“Pilot” minus “Male” = *“Flight Attendant.”* Asking the AI to compute “Doctor” minus “Male” produces the answer *“Nurse.”* Christian illustrates in multiple spheres, including hiring, healthcare, and parole decisions that there can be an alignment problem with Today's AI because of deeply embedded bias. I recommend reading his book for an appreciation of the scope of the problem. In our experience, the solution to AI bias is not to remove the use of variables that show bias in delivery. Rather, the solution is to use the variable to train the AI to overcome the bias, as we will discuss in the next chapter.

The second element of advertising that can manifest bias is*presentation*. Since people respond to the message itself, one of the established facts of advertising effectiveness is that messages that contain people who look like the viewer—or that the viewer can relate to—are more likely to generate a response than do ads with people who look different—or that the viewer can’t relate to. A career ad featuring an older gray-haired white male George Clooney-like executive may not signal to a younger black female prospective employee, “This is an inclusive workplace that will value you.” AI that optimizes for click responses, but fails to adjust the message *presentation* for different groups, will result in the *zoom-out* problem because the response rate will be lower for the people that are not represented in the presentation of the advertisement—even if the ad delivery is equitably delivered. To be specific, consider the pilot advertisements. Global Witness found 75 percent of the ads were shown to men. If this was addressed in the delivery to ensure equal distribution to both men and women that were qualified for the job, but the advertisement features a George-Clooney like pilot, the ad is more likely to get a response from older white males. However, if the advertisement presentation has an equal representation of women and men, the advertisement is more likely to result in a more balanced response rate from women and men. Moreover, if AI is used to match which message presentation are most likely to generate a response from women, and which for men, the AI can improve the equitable recruitment in occupations that skew male or female. The zoom-out problem can be fixed with AI—AI can help increase inclusivity, if AI is used in the right ways to adjust the presentation of the message.

The Bias That Lurks Beneath the Surface

The technical reason for bias is typically because the distribution of data the AI uses for training isn't representative of the population. For example, one study found a root cause in facial recognition's poor performance for women, particularly women of color, was that there weren't enough women of color in the database used to train the facial recognition software to perform as it should when compared to the accuracy for white males. The problem with the distribution of data the AI uses for training may be more complex, such as the way in which cameras capture the color of skin, which can introduce bias, or the way compression algorithms discussed in Part 1 lose detail and introduce bias. Perhaps the compression is more problematic for darker vs. lighter pixels and that creates a skin tone bias. These more complex reasons for bias are not as easy to find or fix.

Here's another example of simple vs. complex issues of representativeness: In Large Language Models (LLMs) one could find sexist bias if the training data set is composed of writings from the 1950s. It's a simple problem that the LLM training data isn't representative of today's society. It may be relatively easy to fix by expunging old writing or down-weighting it. But there may be complex reasons for bias that are harder to fix. There may be bias in the way the AI associates femaleness and maleness in a language model with a number of words, including occupations. As humans, we know that both a pilot and a flight attendant can be either male or female because we understand occupations are not gender specific. A representative association of pilot would be male, female, or non-binary. However, the AI may observe pronouns that are associated with male more often in close proximity with the word "Pilot." The AI may observe words associated with "female" more often with "Flight Attendant." When the AI performs the vector calculation it finds the association "Pilot" = "Male." The AI doesn't understand that the word "pilot" and "flight attendant" are occupations. The AI doesn't understand that occupations ought to be orthogonal (the mathematical term that can be interpreted, in lay terms, as independent) to gender identity. Teaching AI not to be sexist may prove challenging because, as explained in Part 1, even when AI can define the words, Today's AI doesn't understand what words mean the to the same extent that humans understand their meaning.

On the surface, the problem of representativeness may appear easier to solve in some cases, because we can define what the distribution of certain variables *should* look like. We can state we want the representation of male and female job applicants to be equal (or any other ratio). We can partition the AI to generate a model for the best male candidates and another AI model for the best female candidates. There are many adjustments we can make to teach the AI to produce distributions we want—at least the distribution we want on variables that are defined, labeled, and counted. We can make sure we have an equal distribution of age, race, and gender when building a facial recognition training data set—and that should improve the matching accuracy. However, it may be that the distribution that matters is at the pixel-level in ways that are hard for us to fix—the problem occurs the moment a digital camera captures the image. Different digital camera manufacturers have different ways of processing the image and for facial recognition, the way colors are compressed by digital cameras can create bias for skin tone. This is just one of many examples of complex bias that can occur in datasets. As a result, as we apply countermeasures, we should always remember that there may be bias lurking beneath the surface that we may not be able to fathom, and, therefore, offset.

Chapter 12

Experiments As A Countermeasures to Offset AI Bias

The painter George Seurat pioneered a technique of painting called pointillism. It's a technique of painting in which small, distinct dots of color are applied in patterns to form an image. If you stand far enough back, the distinct dots blur together so that the painting looks similar to brush strokes of other impressionist painters of that era. However, move closer and one can see the painting isn't generated by brush strokes, rather there are distinct dots that create the overall image. It's quite remarkable to see in person. Imagine, if you will, that each data point used by the AI is like a dot of color on the canvas. One problem in AI is when the overall picture becomes distorted because the individual dots used to create the image are not representative of the overall image one is trying to create. Instead of having the right balance of blue, green, and red to achieve the image, far too many of the dots are red, creating a distorted view when one *zooms out* to take in the overall image. This can occur when building facial recognition software and not starting with a representative population. It can occur in building a team and using a flawed AI dataset for hiring people to that team. It can occur in advertising, when starting with a message that might appeal more to one group than another. It can occur in large language models where the words and phrases added to the dataset represent old biases rather than current views. Experiments are one countermeasure for improving representativeness—just keep in mind that bias may be much deeper than what we can easily observe.

Experiments are also a countermeasure for the *zoom-in* problem. The *zoom-in* problem occurs when a large set of data, such as the 640 million images in the FBI's facial recognition database, zooms into a single dot on the canvas—and it is the *wrong* dot. The challenge is it can be wrong, but those using AI (such as law enforcement) are likely to have no indication that it is wrong. They will only have a confidence score produced by the AI facial recognition software, which may be biased. Recall from the introduction of the book how the recognition software identified a plant as a person with near 100 percent confidence. Experiments are more reliable than the confidence score generated by the AI. In addition, we present experiments that can act as a countermeasure for the zoom-out problem, where the individual decisions the AI makes might look reasonable in isolation, but in aggregate, show a pattern of bias.

Experiments

Compared to AI, experiments are a much older and simpler approach to finding patterns. One of the oldest recorded experiments to influence public health policy dates back more than 2500 years ago to King Nebuchadnezzar, ruler of Babylon. Nebuchadnezzar ordered his people to eat only meat and drink only wine, a diet he believed would keep them in sound physical condition. But several young men of royal blood, who preferred to eat vegetables, objected. The king allowed these men to follow a diet of legumes and water for 10 days. "When Nebuchadnezzar's experiment ended, the bean-loving teetotallers appeared better nourished than the mandated meat-eaters, so the king allowed them to continue their diet. Not exactly a randomized, double-blinded, placebo-controlled clinical trial, but the modest experiment may have been one of the first times in human history that a medical test, however rudimentary, guided a decision about public health," notes Roger Collier in his article, "Legumes, Lemons and Streptomycin: A Short History of the Clinical Trial."

Over 400 years ago, Galileo suggested using an experiment to compare the speed of falling objects. By 1750, randomized trials began to emerge, as noted in the 1747 study of treatments for Scurvy. A population is randomized into different treatment groups.

In the modern version of clinical trial experiments, the treatment group gets the medication and the control group gets a placebo, which appears identical, but does not contain the drug that is being tested. Since the assignment to the two groups was randomized, the two groups are, overall, identical in all respects, except for the fact one received the placebo and the other the treatment. The math can be quite simple: Treatment Group - Control Group = Effect. It is easy to analyze the data to measure the confidence levels and statistical significance of the results.

RA Fisher advanced experimentation design and analysis with his crop rotation studies in Rothamsted, in the 1920s. He improved the design of experiments and analysis by developing an approach to evaluate multiple variables at the same time, and the statistics to precisely measure the incremental contributions of each variable.

Experiments aren't perfect. Experiments can miss the important interactions that can amplify or blunt the variable under consideration. For example, if a treatment only works among women, then a random sample with an equal number of men and women will reduce the size of the effect by half. It is still likely to be clear that the drug treatment has an effect, but the experiment alone won't reveal the interactions of different variables beneath the surface.

Using Experiments to Help AI Learn

While experiments are not perfect, they are, perhaps, the best way to offset the problems with AI using spurious relationships to fit patterns. Here's an example from advertising: A marketer may start a new TV campaign at the same time as activating a digital advertising campaign, out of home billboards, direct mail and audio campaign. Within the audio campaign, the marketer may have podcast, streaming audio, satellite and local radio. At the same time, the marketer may have three different messages encouraging consumers to consider the brand through different appeals. The marketer wants to know how each contributed to sales so they can make adjustments by reallocating budget to the media and messages that are most profitable. In the background, there may be seasonal effects—recall the ice cream sales example from chapter 2, where sales increase when it is warmer outside (during the summer season) and decline when it is colder (during the winter season). In this particular marketing example, if the product is bought as a gift, there will be more sales in the lead up to Christmas than during other times of the year. A marketer will time the advertising to coincide with holiday shopping—but is the advertising causing people to buy the product, or are the sales up because of the holiday? Another background variable is competitive spending. If a competitor launches a campaign at the same time, it is competing for consumer's preference and purchase. It is difficult to conduct experiments to isolate all of these variables. AI can be used to disentangle the contributions of each advertising element, but a problem known as multicollinearity can make it difficult to know if AI has the patterns right. Multicollinearity occurs when the patterns essentially overlap—recall the regression example from chapter 2, and imagine the regression lines for the different variables move in concert with one another because the variables are highly correlated with one another. When radio starts at the same time as television, and a person is exposed to both, then subsequently buys the product, how does the AI tell for certain which one caused the sale, or what the ratio of contribution should be for TV vs. radio? To help the AI learn an accurate relationship, an experiment could be run where radio ads are turned-on in certain cities that are randomly selected and turned off in others. To further help the AI learn, the experiment could include doing the same with TV—TV ads could be bought locally in a randomized set of cities, or, TV considering the higher expense of buying TV ads locally, TV could be started one week later than radio, so there is a period of time when only radio ads are running. Now, the AI has a better chance to accurately learn the ways in which TV and radio interact to influence a sale.

	Radio Ads On	**Radio Adds Off**
TV Ads On	Both Radio & TV's influence	TV Ads Influence
TV Ads Off	Radio Ads Influence	Baseline Sales (without either radio or TV)

Validating AI: The Experiment Kept Secret From The AI

A second way to use experiments is to run them in a way that is invisible to the AI but is known by a human in the loop. In one use case, we found the AI attributed most of the sales impact to digital ads that were re-targeted to people that had visited a webpage previously. We wondered if it was a bit like the wolf and husky gradient descent classification problem in chapter 3, where the AI was taking a shortcut to classifying the influence on purchase because those that had visited the web page (and were thus retargeted) were more likely to make a purchase whether they saw an advertisement or not. The delivery of the ad, we conjecture, might be coincidental (like the snow in the background of the wolf images) and not causal. Recall that gradient descent means the AI will take the path that results in the most immediate improvement in classification—thus erroneously attributing sales to ads that are opportunistically targeted, but not contributing as much value as the AI indicated. So, we ran an experiment and we used the same retargeting to deliver placebo ads—ads for a Smokey The Bear, which had nothing to do with the advertised product. There was no branding whatsoever. We were, in essence, tricking the AI by labeling these "ads" just to see how the AI would calculate their impact. If the AI was working properly, it should show zero impact on sales from the placebo ads. However, the AI was taking a short-cut, and valuing ads that were re-targeted without really measuring if the ads caused an incremental contribution to sales. The experiment gave us insight on how we could adjust the AI by periodically using placebo ads.

Experiments To Test AI vs. Randomized Experiments To Feed The AI

More broadly speaking, one can construct an experiment to test how the AI performs, and learn whether the AI meets our standards. One could develop a randomized list of professions, feed it through the AI and observe in which circumstances, if any, bias appeared. Here, there is a nuance between an experiment and a randomized experiment. Global Witness performed an experiment in which they had a hypothesis that the AI was biased in the delivery of certain ads to men and women. They tested that hypothesis and found it was correct, the AI showed bias for certain professions (pilot, nurse, etc) by sex. This experiment tests the AI but it doesn't teach the AI anything—teaching the AI wasn't the point. Global Watch was trying to teach humans about the flaws in the AI used for job recruitment. In general, we are interested in how we teach AI to learn patterns with less bias. We think of randomized experiments as more comprehensive tests, which can feedback more valuable data to the AI to help it learn how to offset bias.

In facial recognition, one could create a randomized experiment to systematically test how well facial recognition identifies a wide range of faces, collected independent of the training set and representing the diversity in the population. If such an experiment had been done by the police departments purchasing facial recognition software, the problem of incorrectly classifying people of color would have been identified earlier and might have led to better outcomes.

AI developers should integrate independent randomized experiments into the AI development process. In the case of facial recognition, independent testing across a diverse set of people, a diverse set of image capture modalities (such as different video resolutions, different lighting situations, different camera manufacturers), and at different confidence levels can provide objective perspective on how often the AI will finger the wrong person or fail to match the right person.

To illustrate the importance of testing different confidence levels, consider this experiment: Comparitech, in 2020, fed images of 530 U.S. Senators and Congress members into Amazon's software. Amazon's default setting is 80 percent confidence, and that setting generates 32 incorrect matches. Similar testing was done in the U.K. with their politicians. In both cases, false positives declined as the confidence interval was increased. At 90 percent confidence, there were 12 misidentifications—half of which were not white. Considering that non-white politicians make up only about one-fifth of the population in the test, it reinforced the point that current facial recognition technology has a race bias problem. Setting the confidence interval higher reduces false positives, but also reduces matches. Unfortunately, the same confidence intervals can yield different results in different systems—the interval depends on the proprietary data set and algorithm, but requiring a very high confidence, such as 99 percent might be appropriate for law enforcement, with very few false positives generated, but also far fewer correct matches generated as well. Keep in mind, high confidence from the AI doesn't necessarily mean increased accuracy. In the CS:GO kill bot described in the introduction of this book, the AI had a very high confidence that a plant was an enemy person. In Chapter 2, the AI was 99 percent confident the Panda was a Gibbon, even though the image was imperceptibly hacked to fool the AI. Our recommendation is to perform comprehensive experiments to measure AI's false positives and false negatives. It is an essential step in addressing bias and can be conducted independently by end-users, academics, or others. It is therefore important to conduct experiments that measure the real-world false positives and false negatives.

An example of a comprehensive experiment was reported by *Forbes* in July 2019. Researchers from the University of Essex were given privileged access to six live trials, watching them from the preinstallation discussions through deployment and use by police at the Westfield shopping center in Stratford, London. Before we reveal the results of the experiment, consider what you would deem an appropriate Blackstone ratio of catching the guilty vs. incorrectly catching the innocent. The Blackstone ratio is based on the 18th century English jurist who wrote, “It is better that ten guilty persons escape than one innocent suffers.” (That’s a 10 to 1 ratio). In 1785, a few years before the Bill of Rights, Ben Franklin put the ratio at 100 to 1 when he wrote, “It is better that one hundred guilty persons escape than one innocent suffers.” Consider how many guilty you’d be willing to miss catching so that an innocent person isn’t falsy fingered.

Statisticians and AI data scientists use slightly different terminology compared to 18th century jurists, but great minds think alike—all are formulating a way to address imperfection in the system. Statisticians call it Type I and Type II error. AI data scientists call it “false accept” and “false reject” rates. Those in the legal profession call it the Blackstone Ratio—that is Type II errors (false rejects) to 1 Type I errors (false accepts).

	Association	No Association
Match	Correct	**False Accept (false positive)** Incorrectly Stop an Innocent Person (Type I error)
No-Match	**False Reject (false negative)** Fail to stop a guilty person (Type II error)	Correct

How did AI perform in these live experiments in the UK? Over the six trials, the facial recognition technology matched 42 people walking around Westfield with individuals on a watchlist. In only eight cases could the report authors say with absolute confidence the technology made correct matches. Nonetheless, the police stopped individuals on 26 occasions based on the AI’s conclusions. More than half of these matches (14 or of 26) were found to be incorrect. In four cases, the searches were unsuccessful when the person they were hunting disappeared in the crowds. The report observed, “Overall, in just eight cases was a correct match found.” The author calculated that facial recognition was only successful in 19 percent of cases. For those keeping the Blackstone ratio score, that's five innocent people stopped to catch one guilty person. That is a very high Type I error rate (there was no measurement of the Type II error rate). The study around Westfield was in-line with a separate study that found the police in London used live face recognition to scan 8,600 people's faces—the results were one correct match leading to an arrest, and seven false positives.

False negatives should also be reported. A false negative, in this context, is when AI fails to make a match even though their image is in the database. In essence, this is the other half of the Blackstone ratio, and measures how often a guilty person goes free. False negatives are difficult to observe without the use of a formalized experiment. The London police, for example, could have paid a few hundred people known to be in their database, representing a diverse cross-section to avoid bias, to visit locations that would result in a facial scan so that they could measure the false negative rate. This measurement would complete the Blackstone ratio calculation

Another experiment is measuring how false negatives and false positive rates changed at different confidence levels. The confidence level is a mathematical calculation that many in AI have turned into a setting that can be adjusted in the software, which causes the AI to generate different outputs. Amazon suggested, the confidence interval should be set to 99 percent for law enforcement, at such a setting it may be the case that the system almost never provides any matches because such a setting effectively results in many of the guilty going unmatched. However, these surveillance and matching systems are expensive investments, and the goal is to generate leads—the match itself is not a conviction. If the system rarely produces a match, is it worth the hefty price tag? If the matches are mostly wrong, putting innocent people in the sites of law enforcement, is it worth it? One should measure how useful is the system relative to the investment in reducing crime. Might the same investment be more useful in reducing crime if applied to another area (the next best alternative)?

While these results that the University of Essex reported are concerning, the fact that the authorities invited independent academic researchers to observe and report the findings is laudable. Today's AI is not precise and it is valuable for authorities to encourage experiments measuring the false accept and false reject rates with independent experiments.

The same measurement approach suggested here for facial recognition can be used in a wide range of AI applications in business and government. Law enforcement, business or government should require randomized control testing on an ongoing (or at least periodic) basis for higher risk use of AI. Such testing should cover different modalities (for example, stills and video with a representation of the quality of images fed to the software) at different confidence intervals (starting with the default settings). The test should measure and report false positives and false negatives and the cost/benefit of the system.

Post Hoc Experiments

In the case of advertising, randomized testing can help identify and reduce bias in *delivery* and *presentation*. A post hoc experiment takes a population, such as those that use a particular website over a certain period of time and have an equal opportunity to see advertisements, and compare the groups that received ads for a particular advertiser and those that did not to determine if there are any statistically significant differences between the groups. If, after accounting for the difference in the opportunity to see advertisements based on website usage (heavier users have a greater chance of exposure to any given advertisement) there are statistically significant differences, that is evidence of bias in the delivery of the advertisement. Sometimes this bias is intentional and legal, such as delivering an ad for a running shoe to those that are interested in running. Other times, such as with employment advertisements, the difference in delivery to protected classes is problematic.

Another approach to randomized testing is to set up an experiment, up front. These are often referred to as Randomized Controlled Tests, and they are the gold standard in establishing causality. A population is sorted into different groups at random so that the groups are identical overall. The groups are then treated differently in that they receive different advertisements. Next, the groups are measured to determine if any meaningful differences emerge as a result of advertising exposure. The classic experiment in advertising is to expose one group to advertisements to see if they buy the product at a higher rate than the other group that did not receive an advertisement for the product (instead, they were given a placebo advertisement with an entirely unrelated message). In the scenario of measuring the effectiveness of advertisements for a running shoe company, the treatment group receives advertisements for the running shoe while the other group, known as the control group, receives public service advertisements, such as advertisements for the American Red Cross. Next, at some point in the future, the purchase rate of the two groups are compared. The control group represents the baseline level of what occurs without the exposure to the running shoe advertisement. If the group exposed to the advertisement is bought at a higher rate, it establishes a causal relationship between exposure to the advertisements and sales. It is common to see a five to fifteen percent increase in purchase rate after exposure to advertising.

What if we wanted to experiment to test if advertisements that included someone that looked like the person receiving the advertisements were more likely to result in a response. If we have access to cohort data for age, race and gender cohorts, we could create thirty unique combinations of age ranges, race and gender, and then create advertisements that include the representation of each group. We could write code so that the control group received any of the thirty advertisement representations (the ads are effectively randomized) while the treatment group receives the advertisement that matches their cohort for age group, race and gender. Would such an approach offset the *George Clooney problem* (no offense George)? The results can be analyzed to determine if the approach produced higher engagement compared to the control group.

But, maybe the relationship of which representation resonates most with each person is not as straightforward as matching to one's age group, race and gender. Perhaps there are more complex relationships. AI can perform a similar experiment, if it is designed with control groups and experimentation built in. IBM's Advertising Accelerator with Watson works with their creative lab (a human in the loop) to develop a diversity of imagery to feed their AI. Message elements include images of different people and backgrounds, male and female voice-overs, background music, headlines, and more. They view it as a way to increase inclusivity by allowing the AI to embrace diversity and match the message to people. One can imagine a rainbow of different potential applicants supporting diversity with the AI learning how to best match images to people to increase conversions. IBM's automated experiments compare the same messages with AI turned on versus the same messages with AI turned off—our review of their data found that the AI Personalization technology improves overall advertising conversion by about 50 percent over the control group.

Hiring experiments have been discussed in terms of checking for AI bias in recruitment and candidate selection for interviews. Organizations that choose to use AI in the hiring process should perform periodic experiments to analyze for bias and adjust accordingly. It is relatively easy to construct an experiment to evaluate if AI-based hiring software is surfacing equal proportions of protected class candidates. Countermeasure #2 (described below) provides approaches that can be helpful in offsetting bias.

Cautions

Randomized Control Tests can help in many situations. However, they will not work in all situations. Publishers *Gizmodo* and *The Markup* jointly reported that software called PredPhol dispatched police patrols disproportionately to communities of color, because the software predicted that is where crime would occur. If the predictive policing algorithm uses the number of arrests to determine where crime is likely to occur, but arrests depend on police patrols, then without some randomized variability introduced in the data, is it possible the AI may create a self-fulfilling prophecy?

A countermeasure is to inject *randomized control tests* to expand the search space and create a point of comparison to check the AI. Expanding the search space means adding data that the AI might not have currently. For example, the police data is based in part on where patrols circulate, so this can create a self-fulfilling prophecy. If you expect to find crime in a certain neighborhood, so extra patrols are sent to these neighborhoods, and you subsequently find more crime, is that because there was more crime or does it mean that more patrols will catch more crime?

The solution, where possible, is the double-blind randomized experiment. This approach is well known in pharmaceutical research where neither the person giving the medication to be tested nor the person receiving the medication in a drug trial know if they are receiving the placebo (control group) or the actual drug that is being tested.

It is not difficult for an AI system that is predicting crime to apply random selection and send patrols to neighbors. . However, to the extent the patrol officers believe they are in the randomized control group, and the AI did not predict a certain crime in their patrol area, they may behave differently, and be less vigilant in terms of attempting to intervene, thus perpetuating the bias in the system.

As noted, the gold standard for randomized control tests is medical research where the administration of the placebo drug is double-blinded, meaning neither the person giving the drug nor the person receiving it know whether they are in the control group or not. In either case, both the control and the treatment groups are identical, except for the difference in the drug versus placebo. But policing is one of several situations that do not lend themselves to double-blinded testing. Some forms of bias cannot be corrected with randomized control tests alone.

Chapter 13

Quotas, Partitions and Weighting to Offset AI Bias

Experiments can uncover bias, but what does one do to offset it? It is possible to offset bias along *specific* dimensions by taking a page from 80 years of adjusting for bias in survey datasets. The Parks and Williams cases of false positives leading to arrests are problematic in part because law enforcement should have known better—it has been documented that facial recognition is 10 to 100 times more likely to be wrong for Black and Asian Americans. The reason for the disparity in recognition accuracy is likely due to poor representativeness in the underlying training data set. The case of bias in Facebook's advertising run by Global Witness was a failure in representativeness—the AI was biased by sex and age. Race, age, sex, religion, and disability are protected classes for hiring, and it was extremely problematic when the tech company Brian Christian cited entrench gender bias in their hiring process—this, too, was a problem in representativeness of the training data set.

Fortunately, the survey research industry has decades of experience learning how to correct for bias in representativeness. The 1936 *Literary Digest* case study is a textbook example of failure due to selection bias. Here's a description from the University of Pennsylvania's instructional materials:

> The *Literary Digest* was one of the most respected magazines of the time and had a history of accurately predicting the winners of presidential elections that dated back to 1916. For the 1936 election, the *Literary Digest* prediction was that Landon would get 57% of the vote against Roosevelt's 43% (these are the *statistics* that the poll measured). The actual results of the election were 62% for Roosevelt against 38% for Landon (these were the *parameters* the poll was trying to measure). The sampling error in the *Literary Digest* poll was a whopping 19%, the largest ever in a major public opinion poll. Practically all of the sampling error was the result of sample bias.

> The irony of the situation was that the *Literary Digest* poll was also one of the largest and most expensive polls ever conducted, with a sample size of around 2.4 million people! At the same time the *Literary Digest* was making its fateful mistake, George Gallup was able to predict a victory for Roosevelt using a much smaller sample of about 50,000 people.
>
> This illustrates the fact that bad sampling methods cannot be cured by increasing the size of the sample, which in fact just compounds the mistakes. The critical issue in sampling is not sample size but how best to reduce sample bias. There are many different ways that bias can creep into the sample selection process.

In AI, bias is often the result of the training data set (the sample, to use the pollster's term) not being representative of the population. For example, facial recognition software may suffer from under-representation of certain groups of people. Setting quotas to ensure sufficiently large pools of different classifications of people (age, gender, race, etc.) could improve facial recognition. Returning to the employment example cited in the Christian's book, the sample of successful employees was drawn from the past decade of the technology company's engineers, which historically were mostly males. The AI was designed to use this training set to select from job applicants that matched successful employees. Somewhere in the hidden layers of the AI was a selection for maleness. How might one offset the bias effect inherent in job application screening?

One could apply the technique used by George Gallop and apply quotas. Gallop used variables that were likely to contribute to bias—in this case sex, age, race, economic status—and ensured he had a certain amount of each in his sample. In a similar fashion, those using Today's AI may recognize the risk that AI could under-represent certain groups, and, therefore, set quotas that increase these groups in the training set.

One could *partition* the AI to separately evaluate and score within the partitions divided by sex. Rather than having one AI model to select job applicants, the AI company could use two models—one model for males and one for females. We call the process *partitioning*. With partitioning, the AI no longer choses a male over female because of bias in the training set. Instead, there is a partition so that the AI is choosing among females to best match the historically successful females. Within another partition, the AI is choosing among males that best match the historically successful males. One can apply quotas on top of the partitioned models to generate the top 10 females and top 10 male candidates to interview, according to the AI.

When one knows the risk of bias is pronounced, it may be best to either not use Today's AI or partition the population so that AI is not selecting one group over another, but rather creating multiple AI models to score within a population partition. Quotas and partitions are an easy-to-implement solution in many cases. The partition essentially leads to a quota system. The AI is optimizing within the partition, but a *human in the loop* needs to decide on the quota of how many of each type should be selected.

Quotas can be helpful but are not perfect at removing all bias. Three presidential cycles later, in 1948, the quota approach failed to accurately represent the underlying population. The infamous "Dewey Defeats Truman" Newspaper Headline was, in part, because all the major polls at the time used quotas which are less reliable than an approach that weights the data after the fact to remove bias. To quote *Lumenlearning's* case study:

> This election is considered to be the greatest election upset in American history. Virtually every prediction (with or without public opinion polls) indicated that Truman would be defeated by Dewey…. The intent of quota sampling is to ensure that the sample represents the population in all essential respects. This seems like a good method on the surface, but where does one stop? What if a significant criterion was left out–something that deeply affected the way in which people vote? This would cause significant error in the results of the poll…. A probability sampling is one in which every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals, by weighting sampled units according to their probability of selection.

A more sophisticated answer to the limitations in quotas is weighting for representativeness. One approach is a supervised Machine Learning technique known as Matrix Scaling (also known as

Iterative Proportional Fitting). If one wants equal representation of women and men, while also maintaining certain proportions by age, race, and economic status, one can set the desired proportion of each, and the approach adjusts the weights at a personal level to fit the desired outcome. Matrix Scaling allows for representativeness with much more complex interrelationships compared to quotas.

In our experience, we've seen a lot of examples where it is relatively easy to observe what the representative population should be and, therefore, to apply Matrix Scaling to ensure that the AI is working with representative data. We've used Matrix Scaling in advertising attribution to address the match rate bias problem. Here is an example of matrix scaling used to address a march rate problem. For a business that sold directly to consumers, such as a retailer or insurance company, we had a data set of every sales transaction for each product, by day, region, and more. We knew what was bought on sale or with a coupon, and what was bought by customers in the company's database versus bought by people that were unknown to the company. We had a second data set that gave a complete view of how many ads were delivered by each medium by day. We wanted to connect these datasets to calculate the influence that advertising exposure had on sales transactions. There is a problem in that only a portion of people opt-in to have their media data known to marketers. The opt-in rate is uneven across different media exposures. This creates the match rate bias problem for our AI models. To offset this bias, we used Matrix Scaling to ensure representativeness along all the sales and media details so that the subset of people in the matched database perfectly represented the overall census data for the characteristics in the sales database and media characteristics in the media database. Matrix Scaling is an elegant solution in many applications of AI.

Cautions

Quotas, partitions, and weighting can be a countermeasure for bias in some situations this book would classify as *risky*, and where variables related to bias are known and labeled in the data set, but bias can be more subtle and therefore harder (or even impossible) to offset. AI practitioners need to think deeply and methodically about the risk of bias and whether their partitioning, quotas and weighting countermeasures are adequate.

The Matrix Scaling approach helps offset bias, but isn't perfect. First, the variables for potential bias (such as race, age, sex, etc.) must be collected and used in the algorithm. Sometimes that is problematic because the data isn't available—recall how Facebook removed the variable perhaps believing that it would eliminate bias in their AI, but it didn't. Second, when the variable that presents bias is available, someone must decide on the proportions of each variable. This is easy when there is a census view, such as when addressing the match rate problem where the total purchases are known and that is the population we are aiming to represent. But it isn't so easy when making value judgments such as the technology company whose AI was overlooking women—what should the hiring pool be weighted to represent? Just gender? Or a combination of age, gender, race, religion, and disabilities—the full list of protected classes? How should the weights be set? To reflect the total US population? The population of college graduates? The population of college graduates with a STEM degree? Some other target? What about those that identify as non-binary? Matrix Scaling doesn't do well with small population size groups. Is there enough data for non-binary to be its own category? If not, should non-binary be combined into male or female? These are difficult questions to settle. The devil can be in the implementation details. In general, we advise defining, in writing for transparency, the profile of the population that is representative so there is no ambiguity.

Be aware of the source of the AI's data, and consider ways in which it might be biased. Be cautious with open source and commercial libraries, as it may be difficult to know where bias may creep into these data sets and it may not be possible to weight this data. In Christian's book, he relates an example where a third-party facial recognition library was used in an application—the AI developer incorporating this Facial Recognition Library had no visibility into the dataset for this AI service and whether it was fundamentally biased. Experiments revealed the outcome was biased and that the data set probably did not represent many people of color, leading to problems for those using the recognition service. But, without source data, there was no way to correct the bias.

Bias is not easily offset. It is safer to assume the AI is biased and will not produce a precise answer. Keep in mind Part 1's description of the weakness of AI in*precision*—assume the AI will be wrong some percentage of the time. Consider different failure rates (both false positives and false negatives), and the implications for those experiencing AI's failure—define what is an acceptable ratio. Then, test and measure if AI's failure rate is acceptable.

Let's examine a *WIRED* story about AI developed by Microsoft for the Argentine government to predict which young girls are likely to become pregnant as teenagers. What could go wrong? According to a national television appearance by Juan Manuel Urtubey, a governor in Argentina, The stated goal was to use the algorithm to predict which girls from low-income areas would become pregnant in the next five years. "With technology you can foresee five or six years in advance, with first name, last name, and address, which girl—future teenager—is 86 percent predestined to have an adolescent pregnancy," Utrubey claimed to his TV audience. As *WIRED* reported, "It was never made clear what would happen once a girl or young woman was labeled as 'predestined' for motherhood or how this information would help prevent adolescent pregnancy. The social theories informing the AI system, like its algorithms, were opaque."

Note that the government did not ask which men or boys were likely to impregnate the girls. The framing of the question asked of the AI was biased to begin with. Even if the question asked of the AI wasn't biased, the fact of the matter is that AI isn't*precise* and the underlying data can be biased in some way. AI should not have been applied to classify individual girls who are likely to become pregnant in their teen years. Considering the lack of precision and lack of transparent rationale for the classification, this book would classify this application of AI as *very risky* with potentially serious consequences for those who were incorrectly classified as destined to be pregnant as a teen.

A closer examination of the dataset reveals that the poorest in this region of Argentina rarely have hot running water in their homes. Similar to the classification of wolves and huskies based on the presence of snow, this AI largely relied on the absence of hot running water in the home to classify which girls would become pregnant. The fact that Today's AI simply produces an answer without identifying if the underlying data is biased and without indicating the factors influencing the classification creates risk. Quotas, partitions, and weighting can't always overcome the risk of bias, especially when AI lacks a clear rationale for its decisions. Fortunately, as discussed in chapter 14, researchers are progressing toward making AI rationale more human readable. This would help surface when the AI appears to be using factors we would consider flimsy, problematic, or biased. In this case, additional research and post hoc experimentation revealed that a single variable, whether the home a girl lives in has hot running water or not, seems to be the variable the AI is most relying on to make its dubious prediction.

Chapter 14

Training, Governance and Accountability to Offset AI Bias

At the heart of Parks' and Williams' false positives leading to arrest are biased datasets used in AI Facial recognition combined with *humans in the loop* that did a poor job in acting as a check on known weaknesses of AI. A lack of training on the weaknesses of AI; a lack of *governance* to ensure that facial recognition is used as a clue, not a basis for arrest; a lack of *transparent* data on its uses so that success and failures can be calculated; and a lack of *rigorous testing* further compounds the problems. As far as I can tell, there is no *accountability* when AI is wrong—those responsible for choosing to apply AI seem to take no action to fix the underlying problem and bear no consequence for their mistakes.

Today, there are ways to ensure fairness in the use of AI. IBM supported an effort called AI Fairness 360. It is a free open-source tool kit to check for bias and apply techniques to mitigate it. To quote the white paper:

> "The initial AIF360 Python package implements techniques from 8 published papers from the broader algorithm fairness community. This includes over 71 bias detection metrics, 9 bias mitigation algorithms, and a unique extensible metric explanations facility to help consumers of the system understand the meaning of bias detection results."

We have seen the lack of training on AI weaknesses, a lack of governance, lack of transparency, lack of testing, and lack of accountability in nearly every place where we have seen AI applied. When AI is applied in an area defined in Part 1 of this book as *low risk*, perhaps that is OK. But when AI is applied in domains this book classifies as *risky*, it is essential to have checks in place.

This section will provide examples of training, governance, transparency, and accountability as they relate to AI facial recognition in law enforcement to illustrate the point. One could adapt this list to apply to the use of AI in any domain.

Training: The ideal training gives people a reference point to understand the fundamental weaknesses of AI and then makes that understanding visceral. Hands-on exercises that bring people into direct contact with AI mistakes should drive the point home that AI is fallible, and, therefore, it is essential to implement checks, balances, and other safeguards. For those responsible and accountable for the AI systems, training on how to apply AI Fairness 360 and other procedures for testing AI should round out the training session.

One of the authors is a member of the Washoe County Honorary Sheriff Association where he donates time to help the department make better use of data to evaluate bias and improve policing. In this capacity, we contacted the Sheriff to share this book and offered to run a free training session. In the training, we proposed to capture a range of images that are similar to those used in law enforcement—stills, grainy surveillance video footage, etc. However, this time, the images captured were of the Sheriff's deputies and staff. We proposed to run the images through the software to see if they are matched. The point of this training is for detectives to appreciate that the AI will attempt to make a match and may get the match wrong sometimes. The next part of the training we proposed is to review how the department's data collection can help identify bias. The training shows the evidence that AI is most likely to get matches wrong for certain groups of people. It is better to present the confidence scores as the percent of time the AI will be wrong. Training concludes with brainstorming appropriate checks and balances and then compares the list with current department policy. As Sheriff Darin Balaam shared with me, his department is aware of the weaknesses and decided not to implement AI facial recognition at this time. Considering AI's speed and labor-saving advantages, we expect many will adopt AI, so it is important to plan a governance framework, with training, in advance of AI adoption.

Governance: In Mr. Williams' misidentification case, the Detroit Police department had poor governance related to AI facial recognition technology. They simply asked a staff member (not an eyewitness) if the match the software had produced was the same person they saw in the video. That was enough to issue a warrant for an arrest and upend Williams' life. The police did not subpoena mobile location records for Williams, which likely would have ruled him out as a suspect. They did not go to the eyewitness who was working in the store when the watches were stolen and get a positive identification.

Detroit is not alone in their poor governance of AI technology. In several cases *The New York Times* investigated in Florida, there was no additional evidence beyond the AI image match itself, which violates Florida's own rules. If a human in the loop is to operate as a check on the AI, it needs to do more than rubber stamp the software output and rush to the arrest of the AI fingered person. Law enforcement should develop a strong governance approach to minimize misuse and maximize effective checks on AI.

Beyond the lack of governance within law enforcement (or governance frameworks that are not followed), there is no clear governance at the state or federal level. The FBI system, which runs about twice as many queries per year as Florida's system, uses state driver's license photos from 21 states that, according to the ACLU, do not explicitly allow the use of data in this way. State legislatures and the federal government should govern the use of AI—this is generally called "regulation" and as of this writing, neither Congress nor most state legislatures have passed relevant laws for AI use. When there are regulations, adhering to the regulations is referred to as *compliance*. Compliance with the law, when there is one, is the bare minimum for AI governance. This section focuses on self-governance within an organization which goes one step further and draws from Singapore's Model Governance Framework for AI. Here's a summary of Singapore's model framework:

Guiding Principles:

1. Decisions made by AI should be EXPLAINABLE, TRANSPARENT & FAIR
2. AI systems should be HUMAN-CENTRIC

From principles to practice, here are the key points in the governance framework:

1. Internal Governance Structures and Measures
   - Clear roles and responsibilities in your organization
   - Standard Operating Procedures (SOPs) to monitor and manage risks
   - Staff training
2. Determining the Level of Human Involvement in AI-augmented Decision-making
   - Appropriate degree of human involvement
   - Minimize the risk of harm to individuals
3. Operations Management
   - Minimize bias in data and in the model
   - Take a risk-based approach to measures such as explainability, robustness and regular tuning
4. Stakeholder Interaction and Communication
   - Make AI policies known to users
   - Allow users to provide feedback, if possible
   - Make communications easy to understand

Links to Singapore's model framework, which includes a range of resources for evaluating risk, and other governance considerations here are included in the Chapter 13 AI Worksheet. This framework was developed in conjunction with a diverse set of ten companies including Amazon, DBS Bank, Google, Meta, Microsoft, Singapore Airlines, Singtel Group, and others. It is a solid place to start.

IBM's AI Fairness 360 is another resource to support governance. There are other frameworks, and I encourage selecting one as the model governance framework and engaging with colleagues to add industry specific governance practices.

Transparency: Part of AI governance should be transparency in reporting when and how AI is used. Without transparent reporting of when AI is used, it is impossible to know the overall efficacy or failure rate. In a June 2019 House Oversight Committee interview, the FBI confirmed that it does not track how many times face recognition has led to a conviction. Additionally, the agency does not track how many times the use of face recognition leads to arrests, including arrests of individuals ultimately acquitted.

There are currently no standards for reporting the use of AI in police work. Some departments obscure the use of AI by referring to "investigative techniques identified…" instead of calling out the use of AI. Most software companies have a clear disclaimer such as: "This document is not a positive identification. It is an investigative lead only and is not probable cause for arrest." But without training officers on the weaknesses of AI, it is possible for officers to overestimate AI's accuracy. A lack of appreciation for AI's hidden weaknesses among law enforcement rounds out the problem set.

In addition to providing transparency of AI's successes, failures, and uses, one should include reporting on AI cost and comparison to the next best alternative, as explained in Chapter 6. A transparent annual AI report should be generated by organizations using AI in domains this book would classify as *risky* and *very risky* to show how AI is governed to minimize risk.

Accountability: The first part of accountability is to be clear on what bias one is trying to remove. List any characteristics that should be safeguarded from bias—race, age, sex, and disability are protected classes for hiring, for example. It would be extremely problematic for an AI to entrench bias in a range of applications including law enforcement, lending, healthcare, education, and many other categories. What are the consequences for individuals or vendors that do not follow governance procedures? Is it a fireable offense? Are their clawback provisions in an AI vendor contract for failure to adhere to the governance framework? In other words, what is the accountability in the AI governance system?

RACI is a framework where people in an organization are classified in relation to decisions they make. There is a person that is Responsible for the decision. There is a person that is Accountable (often the supervisor, but it can be the same person that is responsible for the decision). There are decisions that require Consulting a specific person or Informing a specific person. For day-to-day operations of AI, what is the RACI to ensure fair and equitable use of AI? What is the RACI to ensure the governance framework is implemented and standard operating procedures are followed? Are others consulted before using AI? When AI is used, is someone informed so there is a record of the use? In terms of AI governance overall, who is ultimately responsible and accountable? For example, in a law enforcement agency, how is the decision to pursue a lead generated by AI facial recognition made? What is the review process to ensure a human is in the loop? There should be a clearly defined RACI that spells out the process of using AI in as responsible a way as possible.

We are fans of the Japanese quote from quality management: “A defect is a treasure.” It expresses the sentiment that we can learn from mistakes and improve the system. Does the accountable person have the authority to dig in to understand the root cause of the mistakes? Does the accountable person have the authority to change the AI system or the governance procedures to address root causes?

Another aspect of accountability is the responsibility to the person the decision affects. For example, in the law enforcement scenario, what is the responsibility to the person being arrested to ensure the AI is correct? What is the recompense for a false arrest, especially when the rules governing the use of AI are not followed? Governance systems work best when paired with accountability.

Next Steps

Bias in AI is a common occurrence and we can think of bias as part of what makes Today's AI imprecise. Today's AI can produce problematic and unfair results in some instances. There are countermeasures but they are not perfect. It is important for those in decision-making roles to carefully consider the implications of AI being wrong and biased and either avoid the use of AI in risky applications or implement countermeasures and governance. For example, in law enforcement, it seems to us that facial recognition can be a very important way of generating clues—as long as the officers know it provides a clue and is not equivalent to finding a smoking gun. Because AI is far from perfect, it is important that law enforcement has good governance, starting with setting the confidence interval and continuing all the way through the process of making an effort to find exculpatory evidence, such as GPS data that might eliminate a suspect from consideration and save a person the anguish of a false arrest. Departments footing the bills for the software should require that the vendor has on-going testing and a feedback loop to report both the hits and misses. The department should conduct their own independent testing periodically and update training accordingly.

If you don't have a governance framework for AI in situations this book would classify as *risky* and AI is being considered (or is currently being applied), make it a priority in the next 90 days to get a governance and accountability framework in place and to start training accordingly.