3 Bias

Humans are remarkable thinkers. We make hundreds of decisions a day in extraordinarily complex environments with very limited information. And despite the fact that we can't possibly calculate all the consequences of our actions, we are nonetheless surprisingly successful at navigating the modern world. How are we so successful? The simple answer is we cheat. Since the 1970s psychologists have been cataloguing the many rough and ready rules that our brains apply behind the scenes to help us get important decisions right most of the time in circumstances we often meet. Known as "heuristics and biases,"[1] these psychological mechanisms help us make all the decisions we don't really have sufficient information to make. There are few types of human reasoning that don't benefit from this sort of "fast and frugal" thinking.[2] Here are some examples.

The *availability* heuristic tells us that the chance of some phenomenon happening is to be estimated by adding up how often we run into it.[3] It works well if we're trying to estimate the chance of meeting a cat, but not if we want to know, right now, the chance of being murdered. That is because media organizations make a point of informing us about everyone who gets murdered, but they don't similarly inform us about each and every cat sighting, so the availability of information about murders is deceptively high.

Object permanence describes our brains' assumption that objects continue to exist even if we're not experiencing them. It's this built-in Occam's razor that tells you that the flu you woke up with this morning is the same one you went to bed with last night. It's a good principle of critical thinking, but, like all simple "hard-wired" rules, it will inevitably occasionally lead you astray, as when the rabbit the magician pulls out of a hat isn't the one they showed you thirty seconds ago.

When we make estimates of value (How much should I pay for that car?), the *anchoring* heuristic makes us rely too heavily on an initial estimate someone else has given us. We do this even when we don't know how the estimate was reached or indeed whether it's a genuine estimate at all.[4]

In judging whether someone will, for example, be a bad driver, it would seem sensible to start by thinking about the sorts of things that cause bad driving, such as inexperience, inattention, or intoxication. But that's not the way our brains typically work. Rather than thinking about causes, we estimate how similar this person is to the central-casting stereotype of a bad driver (a bias known as *representativeness*). We're afraid of sharks because of their fearsome reputation, not the actual likelihood of encountering one, let alone being harmed by one. This so-called generic reasoning is driven by a mix of the representativeness and availability biases.[5]

The idea that we're better at making decisions than we really are is a bias known as *overconfidence*.[6] The extent and exact nature of all these biases is the subject of ongoing debate, but everyone understands their root cause. Human beings are limited reasoners. We have fallible memories. We can't take account of large amounts of data, and we can't weigh up the effects of many different factors at once. Nowhere are these failings more apparent than when *usually* reliable heuristics and biases result in decisions that are prejudiced.

Prejudices are biased decisions in favor of or against particular things, people, or groups, but not all biased decisions are prejudiced. Refusing to allow young children to drive is a bias, but it's not a prejudice. So, what's the extra ingredient that turns a bias into a prejudice? Academics disagree about exactly what prejudice is. Some think it's irrationality caused by making poor estimates of probabilities.[7] Everyone you've met from the neighboring village seemed rude, so you conclude they are all rude even though you've only met a small proportion of the village. Others think prejudice is a moral failing caused by negligence in the way we reason about groups, particularly when our prejudices justify behavior that benefits us.[8] A third explanation sees prejudice as a side effect of the sort of nonprobabilistic generic reasoning that's built into our psychological make-up and that serves us perfectly well when we avoid scary-looking dogs and food that seems "off."[9] This latter view seems particularly compelling, as it predicts correctly that much prejudice will be unconscious and stubbornly resistant to counterevidence.

What all these views have in common is that prejudice and the discrimination it yields is caused by the human inability to reason objectively from limited information about the complex environments in which we live. The problem is made worse by the powerful effects of our emotions on the objectivity of our decision-making.[10] Negative emotions like fear make us particularly prone to prejudice. Much of humanity's success is due to the development of ideas and institutions that compensate for these cognitive shortcomings: philosophical and historical analysis, the development of law and legal institutions, the scientific method, modern mathematical, statistical and logical principles, and so on. For much of our existence, these generally more accurate and objective types of reasoning have been limited in scope and applied primarily by experts, but now they're being harnessed in widely available and easily usable devices that promise objective and accurate decision-making for everyone.

Artificial Intelligence to the Rescue

Perhaps the first truly successful artificial intelligences were expert systems. These were designed to emulate the thinking of experts in specific domains, such as medical diagnostics. The subject of a great deal of work in the 1970s and 80s, expert systems never really took off for a number of reasons. They were extremely expensive to produce, and (most importantly) they could only emulate reasoning to the extent that it followed strictly definable rules. The great majority of human reasoning isn't rule-based in this strict, deterministic way. Instead, it rests on probabilities and likelihoods, risks and rewards. All the heuristics and biases we mentioned earlier evolved to help us reason about probability and value. Indeed, it's this sort of reasoning that's made humans such a successful species, so it should come as no surprise that the commercial success of AI has resulted from the development of systems capable of learning about probabilities.

The reason modern AI tools are thought to hold so much promise in minimizing bias is precisely because their power, speed, and accuracy mean they can dispense with the use of fast and frugal rules of thumb. AI can analyze large datasets. It can make decisions based on many more types of factors than humans can take into account. It isn't prone to errors in statistical reasoning of the kind humans routinely make and it doesn't use generic reasoning about stereotypes. Above all, it's relentlessly probabilistic.

It's surprising, then, that the most persistent objections to the use of AI in government, commerce, and daily life include allegations of unfairness and bias. Here are the main ones:

- it perpetuates or exacerbates existing inequalities;[11]
- it discriminates against minorities;[12]
- it hyper-scrutinizes the poor and disadvantaged;[13]
- its risk assessments in domains like justice and policing are fundamentally unfair;[14]
- it is pseudo-objective;[15]
- it evades existing protections against discriminatory reasoning based on race, gender, and other protected categories;[16]
- it obscures the often complex decisions made by developers about how to interpret and categories facts about people's lives;[17] and
- it fundamentally distorts the nature of commerce,[18] politics,[19] and everyday life.[20]

In this chapter, we'll explain and assess these allegations. Some, we will argue, are more pernicious than others. And some, due to faulty applications of probabilistic reasoning, are clearly fixable.

Taking Humans out of the Equation

Human decision-making is accurate enough in everyday circumstances, but for the reasons just discussed, the intuitive judgments of individuals aren't sufficiently reliable in high-stakes environments (Who gets a heart bypass? Who gets imprisoned?). Also, humans have a problem with partiality. Our beliefs are famously infected by our desires (colloquially known as *wishful thinking*), and evolution seems to have implanted within us a strong drive to put our own interests first, followed by those of close relatives,[21] and then those of our social groups.[22] We get around this problem in various ways in high-stakes circumstances. We vote as jurors and board members as a way of averaging out individual preferences. Social workers utilize structured decision-making when they use checkbox forms to assess risks to individuals. In many contexts, we require expert decision makers, such as judges, to set out their reasoning. But all these mechanisms are cumbersome and prone to error and deception. As we saw in chapter 2, humans

are adept at rationalizing their decisions in ways that make their reasoning look nobler and more sensible than it often is. Being forewarned about the dangers of unconscious bias is surprisingly ineffective at enhancing our objectivity. Even professionals using structured decision-making tools, like checkbox forms, are known to cheat when filling out such forms in order to achieve the result that their intuitions tell them is the right one.[23] So part of the promise of AI is that in such high-stakes circumstances it could take people out of the equation, replacing them with algorithms that are reliable and impartial. But removing human bias from AI isn't as simple as it sounds. Machine learning is extraordinarily powerful, but it's humans who decide how to build and train such systems, and both the building and the training processes are open to bias. What we called "process" transparency in chapter 2 is therefore of the utmost importance.

Computer scientists are fond of pointing out that bias is not inherently a bad thing. It's part of the way we think and a very important part of what makes modern machine-learning systems so successful. It's the very fact that people's biases in what they prefer to read are reflected in the statistical patterns of their previous purchases that makes Amazon reasonably successful at predicting what they'd like to read next. Exploiting such biases allows us to develop algorithms for social sorting, market segmentation, personalization, recommendations, the management of traffic flows, and much more. But there's much to decide in working out how to successfully exploit biases in the data.

In unsupervised machine learning, developers must decide what problems they want to solve. If we're developing a system for assessing job applicants, we need to understand and prioritize the human resources challenges facing the company that will use the system. Is there a widespread mismatch between skill sets and job descriptions? Is staff turnover too high? Does the workplace need more diversity? Having settled on a problem or problems to solve, developers identify the patterns detected in the dataset relevant to those problems. Whether we're successful will ultimately depend on which data we use to train the system. How are the data collected and organized? Is the dataset diverse enough to give us accurate information about the variety of people who'll use the algorithm or on whom it'll be used? Can a system devised to detect crime in Chicago be successfully deployed in New York? What about Mumbai?

In supervised machine learning, we train the algorithm by telling it when it has the right answer. To do so, we must start by identifying success—simple

enough if we're training a facial recognition system, but what if we're developing a dating app? To judge success, we'd need to understand what users wanted from a dating app. Are people looking for a life partner? If so, is there an exploitable correlation between the information people provide to dating apps and the longevity of relationships? No doubt there'll be significant disagreement about which dates are "good" from this perspective, because different people look for different things from life partners. But maybe there's *some* correlation in the data. If not, should we set our predictive sights lower—maybe the predictors of repeat dates instead of life partnerships—and hence decide not to serve the interests of those looking for long-term relationships? How much will the owners of the tool tell users about whose interests they decide to serve?

All these human choices about how algorithms are developed are further influenced by the choices of users regarding how they interact with AI and how they interpret what it tells them. Although AI continues to get better at prediction and detection tasks, researchers and journalists have identified a surprising variety of ways in which it can perpetuate disadvantage and cause harm to minorities and individuals who are far away from the population average (known by data scientists as "statistical outliers").

Predicting the Future by Aggregating the Past

"It's tough to make predictions, especially about the future." So said baseball great Yogi Berra, and he was right. We can only make accurate predictions about the future by aggregating what we know of the past. That information is inevitably incomplete and of variable quality. When we try to emulate human predictions in artificial intelligence, it often means that the algorithms we produce are fueled by aggregating the same intuitive and sometimes prejudiced human decision-making we're trying to improve on. Aggregating human decisions can be successful when we rely on the "wisdom of the crowd," but crowds aren't always wise. The wisdom of the crowd is most accurate when a diverse group of people are prone to make random errors and averaging out their judgments can effectively minimize those errors. This was famously demonstrated in 1908 by Frances Galton, who averaged out the guesses of all the contestants trying to guess the weight of a prize ox at the Plymouth County Fair. But for this to work, the guesses have to be independent of each other—members of a "wise" crowd

can't influence other members.[24] You don't get an accurate estimate if guessers are free to copy the guesses of those they think are especially smart or knowledgeable. The wisdom of the crowd also fails when the errors of individuals aren't random.[25] We might all be guessing independently of one another and yet be prone to a shared bias. If, for example, humans typically estimate dark-colored animals to be heavier than light-colored ones, averaging out the guesses at the County Fair will just give us a biased average.

These two facts about the way the wisdom of crowds fails tell us something important about the dangers of developing predictive risk models fueled by data that consist of human intuitions and estimates. Normal human decision-making is replete with cognitive biases that sometimes result in systematic prejudice. This is exactly when crowds fail to be wise. So aggregating the intuitions of large numbers of individuals can produce algorithms that mirror the systematic human biases we're trying to avoid. Of course, machine learning systems are doing something much more complex than averaging. Nevertheless, fed biased data, they'll produce biased results. Nowhere are these risks more apparent than in the use of predictive risk models in policing.

Predictive policing employs artificial intelligence to identify likely targets for police intervention, to prevent crime, and to solve past crimes. Its most famous incarnation is PredPol, which began in 2006 as a collaboration between the Los Angeles Police Department and a group of criminologists and mathematicians at the University of California, San Diego. The company was incorporated in 2012 and is now used by more than sixty police departments around the US,[26] and also in the United Kingdom. PredPol makes predictions about the locations of future crime that help cash-strapped police forces allocate their resources. The predictions appear in real time as high-risk crime "hot-spots" displayed as red boxes in a Google Maps window. Each box covers an area of 150 square meters.

This is a growing industry, with PredPol now facing competition from companies such as Compustat and Hunchlab, who are racing to incorporate any information they think will help police do more with less. Hunchlab now includes Risk Terrain Analysis that incorporates features such as ATMs and convenience stores known to be locations for small-scale crime.[27] On the face of it, this seems like a sensible and admirably evidence-driven approach to crime prevention. It doesn't focus on individuals, and it doesn't know about ethnicity. It just knows crime stats and widely accepted

criminological results (e.g., the likelihood of your house being burgled is strongly influenced by the occurrence of recent burglaries nearby).

But the apparent objectivity of these tools is deceptive. They suffer from the same problem that has always plagued policing. Police must make judgments about how best to prevent crime, but they only have patchy information about the incidence of crime, relying as they do on statistics about reports of crime, arrests, and convictions. Many types of crime are, by their nature, difficult to detect. You certainly know if you've been mugged, but you may well not know if you've been defrauded. Other types of crime, such as domestic violence, are persistently underreported. And of course, many reported crimes don't lead to arrests and subsequent convictions. This means that there's often a danger that the crime data informing predictive policing tools is influenced by the intuitive judgments of individual officers about where to go, whom to talk to, which leads to follow up on, and so on. It would be a Herculean task to assess the objectivity of these intuitive judgments. So, for all that predictive policing sounds scientific and the crime maps produced by companies like PredPol look objective and data-driven, we really have to accept that we cannot tell how objective the output of such tools really is. It's perhaps not surprising that civil rights groups have not been convinced by PredPol's claims that its use of only three data points (crime type, crime location, and crime date/time) eliminates the possibility for privacy or civil rights violations. A joint statement by the American Civil Liberties Union and fourteen other civil rights and related organizations focused directly on this problem:

> Predictive policing tools threaten to provide a misleading and undeserved imprimatur of impartiality for an institution that desperately needs fundamental change. Systems that are engineered to support the status quo have no place in American policing. The data driving predictive enforcement activities—such as the location and timing of previously reported crimes, or patterns of community- and officer-initiated 911 calls—is profoundly limited and biased.[28]

It's tempting to think that this problem of tools relying on data contaminated by human intuitions will gradually dissipate over time as police come to rely more on statistically accurate algorithms and less on the intuitions of officers, but the actual effect of developing a predictive risk model based on a systematically biased dataset can be to bake in bias rather than to allow it to gradually dissipate. Kristian Lum and William Isaac show that predictive policing models are likely to predict crime in areas already believed

to be crime hotspots by police.[29] These areas will then be subject to more police scrutiny, leading to observation of criminal behavior that confirms the prior beliefs of officers about where crime is most common. These newly observed crimes are fed back into the algorithm, generating increasingly biased predictions: "This creates a feedback loop where the model becomes increasingly confident that the locations most likely to experience further criminal activity are exactly the locations they had previously believed to be high in crime: selection bias meets confirmation bias."[30]

These problems aren't inevitable and we certainly don't mean to suggest that predictive policing strategies are useless. We want high-stakes decisions to be data-driven and predictive risk models are likely to be more accurate and more reliable at making predictions from those data. The problem, as always in policing, is getting better data. Essentially the same problem strikes algorithms designed to help make decisions about the allocation of bank loans, medical procedures, citizenship, college admissions, jobs, and much else besides. So it's a good sign that significant effort is being put into tools and techniques for detecting bias in existing datasets. As of 2019, major tech companies including Google, Facebook, and Microsoft, have all announced their intention to develop tools for bias detection, although it's notable that these are all "in-house." External audits of biases in the artificial intelligence developed by those companies would probably be more effective. That said, bias detection is only the start, particularly in domains like policing. Many countries already know that their arrest and incarceration statistics are heavily biased toward ethnic minorities. The big question is, can we redesign predictive policing models, including rules about how they're used and how their source data are collected, so that we're at least not further disadvantaging poor "crime-ridden" communities?

At this point, the most important thing governments and citizens can do is to acknowledge and publicize the dangers of the algorithms that suffer from the "bias in/bias out" problem. It'll then be in the interests of the makers and owners of such algorithms to modify their products to help minimize this type of bias. It must become common knowledge that AI predictions are only as good as the data used to drive them. The onus is at least partly on us to be savvy consumers and voters.

To better understand how to tackle algorithmic bias, we're going to need to know more about the ways human biases can creep into the development of AI. There are many ways to categorize algorithmic biases, but we'll

divide them into three groups: biases in how we *make* AI; biases in how we *train* it; and biases in how we *use* it in particular contexts.

Built-in Bias

Human bias is a mix of hardwired and learned biases, some of which are sensible (such as "you should wash your hands before eating"), and others of which are plainly false (such as "atheists have no morals"). Artificial intelligence likewise suffers from both built-in and learned biases, but the mechanisms that produce AI's built-in biases are different from the evolutionary ones that produce the psychological heuristics and biases of human reasoners.

One group of mechanisms stems from decisions about how practical problems are to be solved in AI. These decisions often incorporate programmers' sometimes-biased expectations about how the world works. Imagine you've been tasked with designing a machine learning system for landlords who want to find good tenants. It's a perfectly sensible question to ask, but where should you go looking for the data that will answer it? There are many variables you might choose to use in training your system—age, income, sex, current postcode, high school attended, solvency, character, alcohol consumption? Leaving aside variables that are often misreported (like alcohol consumption) or legally prohibited as discriminatory grounds of reasoning (like sex or age), the choices you make are likely to depend at least to some degree on your own beliefs about which things influence the behavior of tenants. Such beliefs will produce bias in the algorithm's output, particularly if developers omit variables which are actually predictive of being a good tenant, and so harm individuals who would otherwise make good tenants but won't be identified as such.

The same problem will appear again when decisions have to be made about the way data is to be collected and labeled. These decisions often won't be visible to the people using the algorithms. Some of the information will be deemed commercially sensitive. Some will just be forgotten. The failure to document potential sources of bias can be particularly problematic when an AI designed for one purpose gets coopted in the service of another—as when a credit score is used to asses someone's suitability as an employee. The danger inherent in adapting AI from one context to another has recently been dubbed the "portability trap."[31] It's a trap because it has

the potential to degrade both the accuracy and fairness of the repurposed algorithms.

Consider also a system like TurnItIn. It's one of many anti-plagiarism systems used by universities. Its makers say that it trawls 9.5 billion web pages (including common research sources such as online course notes and reference works like Wikipedia). It also maintains a database of essays previously submitted through TurnItIn that, according to its marketing material, grows by more than fifty thousand essays per day. Student-submitted essays are then compared with this information to detect plagiarism. Of course, there will always be some similarities if a student's work is compared to the essays of large numbers of other students writing on common academic topics. To get around this problem, its makers chose to compare relatively long strings of characters. Lucas Introna, a professor of organization, technology and ethics at Lancaster University, claims that TurnItIn is biased.[32]

TurnItIn is designed to detect copying but all essays contain something like copying. Paraphrasing is the process of putting other people's ideas into your own words, demonstrating to the marker that you understand the ideas in question. It turns out that there's a difference in the paraphrasing of native and nonnative speakers of a language. People learning a new language write using familiar and sometimes lengthy fragments of text to ensure they're getting the vocabulary and structure of expressions correct.[33] This means that the paraphrasing of nonnative speakers of a language will often contain longer fragments of the original. Both groups are paraphrasing, not cheating, but the nonnative speakers get persistently higher plagiarism scores. So a system designed in part to minimize biases from professors unconsciously influenced by gender and ethnicity seems to inadvertently produce a new form of bias because of the way it handles data.

There's also a long history of built-in biases deliberately designed for commercial gain. One of the greatest successes in the history of AI is the development of recommender systems that can quickly and efficiently find consumers the cheapest hotel, the most direct flight, or the books and music that best suit their tastes. The design of these algorithms has become extremely important to merchants—and not just online merchants. If the design of such a system meant your restaurant never came up in a search, your business would definitely take a hit. The problem gets worse the more recommender systems become entrenched and effectively compulsory in certain industries. It can set up a dangerous conflict of interest if the same

company that owns the recommender system also owns some of the products or services it's recommending.

This problem was first documented in the 1960s after the launch of the SABRE airline reservation and scheduling system jointly developed by IBM and American Airlines.[34] It was a huge advance over call center operators armed with seating charts and drawing pins, but it soon became apparent that users wanted a system that could compare the services offered by a range of airlines. A descendent of the resulting recommender engine is still in use, driving services such as Expedia and Travelocity. It wasn't lost on American Airlines that their new system was, in effect, advertising the wares of their competitors. So they set about investigating ways in which search results could be presented so that users would more often select American Airlines. So although the system would be driven by information from many airlines, it would systematically bias the purchasing habits of users toward American Airlines. Staff called this strategy *screen science*.[35]

American Airline's screen science didn't go unnoticed. Travel agents soon spotted that SABRE's top recommendation was often worse than those further down the page. Eventually the president of American Airlines, Robert L. Crandall, was called to testify before Congress. Astonishingly, Crandall was completely unrepentant, testifying that "the preferential display of our flights, and the corresponding increase in our market share, is the competitive raison d'être for having created the [SABRE] system in the first place."[36] Crandall's justification has been christened "Crandall's complaint," namely, "Why would you build and operate an expensive algorithm if you can't bias it in your favor?"

Looking back, Crandall's complaint seems rather quaint. There are many ways recommender engines can be monetized. They don't need to produce biased results in order to be financially viable. That said, screen science hasn't gone away. There continue to be allegations that recommender engines are biased toward the products of their makers. Ben Edelman collated all the studies in which Google was found to promote its own products via prominent placements in such results. These include Google Blog Search, Google Book Search, Google Flight Search, Google Health, Google Hotel Finder, Google Images, Google Maps, Google News, Google Places, Google+, Google Scholar, Google Shopping, and Google Video.[37]

Deliberate bias doesn't only influence what you are offered by recommender engines. It can also influence what you're charged for the services

recommended to you. Search personalization has made it easier for companies to engage in *dynamic pricing*. In 2012, an investigation by the Wall Street Journal found that the recommender system employed by a travel company called Orbiz appeared to be recommending more expensive accommodation to Mac users than to Windows users.[38]

Learning Falsehoods

In 2016, Microsoft launched an artificially intelligent chatbot that could converse with Twitter users. The bot called "Tay" (which stands for "thinking about you") was designed to mimic the language patterns of a nineteen-year-old American girl. It was a sophisticated learning algorithm capable of humor and of seeming to have beliefs about people and ideas.[39] Initially, the experiment went well, but after a few hours, Tay's tweets became increasingly offensive. Some of this invective was aimed at individuals, including President Obama, consisting of false and inflammatory claims about events and ethnicities. In the sixteen hours and ninety-three thousand tweets before Microsoft shut it down, Tay had called for a race war, defended Hitler, and claimed that Jews had caused the 9/11 attacks.[40]

Microsoft's explanation of what went wrong was limited to "a coordinated attack by a subset of people exploited a vulnerability in Tay. Although we had prepared for many types of abuses of the system, we had made a critical oversight for this specific attack."[41] There is little doubt that Tay was attacked by trolls who were deliberately feeding it false and offensive information, but perhaps the more important problem is that Tay wasn't in a position to know that the trolls *were* trolls. It was designed to learn, but the vulnerability to which Microsoft refers left it without the capacity to assess the quality of the information it was given. Certainly it lacked the cognitive complexity, social environment, and educational background of a real teenager. In short, Tay was a learner but not a *thinker*.

As we mentioned in the prologue, we may one day invent a general AI, capable of understanding what it says and possessing a level of complexity that would allow it to assess the truth of the data it received, but we're not there yet, and may not be for a long time to come. Until then we're stuck with narrow AI that is completely dependent on human beings supplying it with accurate and relevant data. There are several ways in which this can go awry.

Clearly Tay wasn't the only one being lied to on the internet in 2016. From people lying on dating sites to lying about hotels on TripAdvisor, there's no shortage of individuals happy to submit false information into recommender engines. Although progress is being made in using machine learning to detect lying online, at least the more savvy users of sites relying on aggregated user contributions retain some skepticism about the views they express.

More concerning are cases in which a training set contains data that accurately represent one part of a population, but don't represent the population as a whole—the phenomenon we called "selection" (or "sampling") bias earlier. In a famous recent instance of this problem, Asian users of Nikon cameras complained that the camera's software was incorrectly suggesting that they were blinking.[42] Users have similarly complained that tools relying on speech recognition technology, such as Amazon's Alexa, are inaccurate at interpreting some accents. They're particularly poor at recognizing the speech of Hispanic and Chinese Americans.[43]

These issues can have serious consequences in high-stakes environments. The feeds of CCTV cameras are now routinely assessed using facial recognition systems, which have been shown to be particularly sensitive to the racial and gender diversity of the data on which they're trained.[44] When these algorithms are trained on criminal databases for use in law enforcement, it's not surprising that they show biases in their abilities to recognize the faces of different groups of people. They identify men more frequently than women; older people more frequently than younger; and Asians, African Americans, and other races more often than whites.[45] As a result, false accusations or unnecessary questioning is more likely to land on women and members of racial minorities.

Are we being too hard on the makers of facial recognition algorithms? After all, something like this problem occurs when humans recognize faces. The "other-race" effect refers to the fact that humans are better at recognizing the faces of people of their own ethnicity.[46] So the problem isn't that AI makes mistakes—people make mistakes too. The problem is that the people using these systems might be inclined to think that they're bias-free. And even if they're aware of issues like selection bias, users may still not be in the best position to identify the strength and direction of bias in the tools they use.

But the news isn't all bad. Well trained, these systems can be even better than humans. There is much we can do to increase the diversity and

representativeness of the databases on which such algorithms are trained. And although there are barriers to the collection and processing of quality data owing to anti-discrimination and intellectual property law (e.g., copyright in images), they aren't insurmountable.[47] Such problems can also be addressed when people purchasing or using such systems make a point of asking what the developers have done to ensure that they are fair.

Bias in the Context of Use

A 2015 study detected that Google's ad-targeting system was displaying higher-paying jobs to male applicants on job search websites.[48] It may be that this algorithm had built-in biases or that something could've been done to address selection bias in the dataset on which it was trained. But putting those issues aside, there is likely something else going on—something at once more subtle but also (in a way) obvious. All countries suffer from gender pay gaps. What the system may have *correctly* been detecting is the fact that, on average, women are paid less than men. What it couldn't detect is that that fact about the world is unjust. The fact that women don't do as much high-paid work has nothing to do with their inability to perform such tasks. Rather, it reflects a history of gender stereotypes, discrimination, and structural inequality that society is only now beginning to address systematically. Such cases are instructive for those who develop and deploy predictive algorithms. They need to be attuned to the ways in which a "neutral" technology, applied in particular circumstances, can perpetuate and even legitimize existing patterns of injustice. As a society changes, its use of algorithms needs to be sensitive to those changes.

Sadly, problems like these are difficult to address. We can, of course, remove variables such as gender and ethnicity from the datasets on which we train such algorithms, but it's much more difficult to remove the *effect* of such factors in those datasets. The makers of PredPol deliberately avoided using ethnicity as a variable in their predictions about the locations of future crime, but PredPol is all about geography, and geography is strongly correlated with ethnicity. Data scientists would say that geography is a *proxy variable* for ethnicity, so the result of using such systems is often to focus police scrutiny on minority communities.

Can AI Be Fair?

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is described by its makers, the company Northpointe (now Equivant), as a "risk and needs assessment instrument," used by criminal justice agencies across the US "to inform decisions regarding the placement, supervision, and case management of offenders."[49] It consists of two primary predictive risk models designed to predict the likelihood of recidivism amongst prisoners in general and specifically amongst violent offenders. The details of its design are a trade secret but prisoners encounter it as a lengthy questionnaire that is used to generate a risk score between one and ten. That score is used in many contexts within the justice system, including decisions about where they're imprisoned and when they're released.

In 2016, ProPublica, an independent journalism organization, conducted a two-year study of more than ten thousand criminal defendants in Broward County, Florida. It compared their predicted recidivism rates with the rate that actually occurred. The study found that "black defendants were far more likely than white defendants to be incorrectly judged to be at higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly tagged as low risk."[50]

Northpointe strongly denied the allegation, claiming that ProPublica made various technical errors (which ProPublica denied). Northpointe also argued that ProPublica had not taken into account the fact that recidivism is higher amongst the black prison population. ProPublica was unconvinced. Black Americans are more likely to be poor which diminishes their educational opportunities. They're also more likely to live in high-crime neighborhoods with fewer job opportunities. If these facts are even part of the explanation for their higher recidivism rates, it seems unjust that they be further penalized by an algorithm that decreases their chance of parole. It soon became clear that the debate between Northpointe and ProPublica was fundamentally a disagreement about how an algorithm such as COMPAS could be made fair.

It's true that we can develop algorithms that are, in some sense, "fairer." The challenge, however, is that there are many conflicting algorithmic interpretations of fairness, and it isn't possible to satisfy all of them. Three common definitions of fairness are discussed in the machine learning community.

- We can ensure that protected attributes, such as ethnicity and gender (and proxies for those attributes), aren't explicitly used to make decisions. This is known as *anti-classification*.
- We can ensure that common measures of predictive performance (e.g., false positive and false negative rates) are equal across groups defined by the protected attributes. This is known as *classification parity*.
- We can ensure that, irrespective of protected attributes, a risk estimate means the same thing. A high recidivism risk score, for example, should indicate the same likelihood of reoffending regardless of the ethnicity, gender, or other protected attribute of the assessed person. This is known as *calibration*.

It's not hard to show using a bit of high-school algebra that when the incidence (or "base rate") of a phenomenon like recidivism differs across distinct populations (e.g., ethnic groups), you simply can't simultaneously satisfy all such criteria; some of them are mutually exclusive.[51] So Although Northpointe could defend COMPAS by showing it to be well calibrated (risk scores mean the same regardless of group membership),[52] ProPublica was nonetheless able to claim that COMPAS is biased because it doesn't satisfy classification parity (specifically, *error rate balance*—where the rate of false positives and false negatives is equal across groups). As we saw, ProPublica found that black people assessed with COMPAS were more likely than white people to be incorrectly classified as high risk (so the rate of false positives differed across the two groups), whereas white people were more likely than black people to be incorrectly classified as low risk (so the rate of false negatives differed across the two groups). Fact is, you can't satisfy both calibration and error rate balance if the base rates differ—and, of course, for a variety of reasons (including historical patterns of prejudicial policing), recidivism base rates *do* differ between black and white populations in the United States.

There are also trade-offs between fairness and accuracy. For example, in the COMPAS case, anti-classification would require that the data used didn't contain any proxies for race. As noted above, race is strongly correlated with many important facts about people including income, educational achievement, and geographic location. So removing such variables would be likely to make the algorithm much less accurate and therefore much less useful.

Recent work in data science shows that conflict between these various notions of fairness isn't limited to the COMPAS case and isn't always a

result of design failures by those developing such algorithms. These different characterizations of fairness are logically incompatible. As scholars have noted, "there is a mathematical limit to how fair any algorithm—or human decision maker—can ever be."[53] Nor are these conflicts purely technical, either: they're as much political. Choosing between different standards of fairness calls for informed discussion and open, democratic debate.

Summing Up

This chapter might seem unfair to those developing AI, listing as it does a surprising variety of ways in which AI can be biased, unfair, and harmful, but we're not Luddites, and don't mean to portray AI itself as the problem. Many of the issues raised in this chapter affect human reasoning just as much as they affect AI. The incompatibility of anti-classification, classification parity, and calibration is first and foremost a fact about *fairness*, not AI. Human intuitions about fair decision-making are at odds with each other, and this tension is reflected in the statistical rules that formalize those intuitions. Likewise, algorithmic bias is parasitic on human bias, and, if anything, the effect of human bias is in some ways worse than algorithmic bias. The sort of generic reasoning that leads to human prejudice isn't only harmful; it's irrational, often unconscious, and insensitive to counter-evidence. By contrast, many of the biases from which AI suffers can be detected and, at least in principle, remedied. That said, there are few easy fixes. Algorithms can't be completely fair, and some ways of enhancing their fairness (such as anti-classification) make them less accurate.

So, the ball is in our court. As consumers and voters, we'll need to thrash out the sort of fairness we want in the many domains in which AI is used. This, of course, means we'll have to decide how much *unfairness* we're prepared to put up with.