Judged by Machines
In the mid-1990s, various US carriers raised an antitrust case against American Airlines and United Airlines. The complaint was that online search systems were biased against foreign and domestic carriers. Their case focused on the algorithms used in ticket reservation systems. These systems prioritized flights that used the same carrier for all the legs of a trip, so an agent searching for a ticket from Louisville, Kentucky, to London, going through New York, would see flights involving no change of carrier higher on the list than flights involving two carriers. Since in the 1990s, screens showed only four to five flights, and 90 percent of all bookings came from the first screen, small differences in ranking had big financial implications.

As this airline reservation example shows, algorithms are not always fair. In fact, algorithmic bias is now a prevalent topic of discussion as it concerns computer vision systems, university admission protocols, natural language processing, recommender systems, courtroom risk assessment tools, online advertisements, and finance. But much like human biases, algorithmic biases are not simply the result of maleficence. They can emerge from both practical and fundamental considerations.

On the practical side, both people and machines learn from data that are often biased and incomplete—the data we have instead of the data we wish we had.
Biased data can lead to biased learning and behavior. But even with the existence of perfect data, guaranteeing fairness may not be possible. Fairness is a concept that can be defined in multiple ways, so it is not always possible to satisfy multiple definitions of fairness simultaneously.\(^\text{11}\)

To illustrate this fundamental limitation, consider two populations: A and B. A and B could be people identifying with different genders, or belonging to different nationalities, age groups, or ethnicities. For the purpose of this exercise, the type of difference or its source doesn’t matter. What matters is that we want to achieve a fair outcome when it comes to our treatment of populations A and B.

But what constitutes a fair outcome? To keep things simple, consider two definitions.

The first definition is known as *statistical parity* or *demographic parity*. This means guaranteeing that outcomes affect equal proportions of A and B.\(^\text{12}\) The second definition is *equality of false rejections* or *equality of opportunity*.\(^\text{13}\) This means guaranteeing that the probability of being rejected if you are from population A or population B is the same.

In principle, satisfying both definitions is possible if we consider an outcome that doesn’t hinge on any particular selection criterion or merit. For instance, if we pass out free concert tickets at random, we would satisfy both statistical parity and equality of false rejections. But what if the fans of the band playing in the concert were not equally distributed among both populations, A and B? In that case, distributing tickets at random would be unfair for the group that included most of these fans. Fans in this group would get fewer tickets and be more likely to be rejected.

This simple example can help us motivate more complex—and relevant—cases. Instead of free concert tickets, consider giving out loans, admitting students to college, or giving someone a promotion at work. These are all cases that not only are more delicate, but also imply some degree of selection or merit. The case of the loan is more straightforward. In principle, loans should be allocated to those who are more likely to
repay them. Promotions and college admissions are trickier because they invoke the idea of merit, which may be harder to measure, even post hoc, than whether someone can repay a loan.

To illustrate how selection or merit interacts with our two notions of fairness, let populations A and B be of the same size but have a different probability of paying back a loan. To keep things simple, assume that 40 percent of the people in A would repay a loan, but only 20 percent of the people in B would.

We can achieve statistical parity by giving loans to 20 percent of the people in A and 20 percent of the people in B. This would be fair, in that the same fraction of both populations would get a loan, but would violate equality of opportunity, since we would be rejecting 20 percent of people in A who would repay their loans. But if we enforce equality of opportunity, we will end up giving more loans to people in group A, violating statistical parity.

All this goes to show that, even in simple examples, satisfying multiple definitions of fairness cannot be guaranteed. This is not because fairness cannot be defined, but because it allows multiple definitions. In this particular example, we used only two definitions, but we could have used many more. We may include a third definition, requiring equality in false acceptances (e.g., giving loans to people who will not pay them with the same probability in both groups).

Fairness is a complex concept that accepts multiple definitions that (in most cases) cannot be satisfied simultaneously. The world is unfair—not only because people and machines are biased—but because it affords multiple ways of defining a fair outcome.

Yet, not all unfairness comes from mathematical impossibilities. In fact, unfairness also comes from algorithms and the data used to design them. While in principle, these sources of unfairness could be vexing, in practice they are also sources of unfairness that potentially could be corrected.
Consider the example of word embeddings. *Word embedding* is a natural language-processing technique used to translate words into mathematical representations. It is also a popular example of algorithmic bias. This is because word embeddings can perpetuate the racial and gender stereotypes found in its training data. In a word embedding, adding the vector for the word *Queen* and that for the word *Man* gives you the word *King*. This means that these vectors satisfy semantic relationships (e.g., “a King is a male Queen”). But not all the relationships learned by word embedding are as simple and uncontroversial. Word embeddings also encode relationships, such as “Man is to computer programmer as woman is to homemaker.” In fact, if the text used to train the embedding contains mentions of women performing stereotypical actions, such as cooking or cleaning, the embedding will codify, maintain, and sometimes even enhance these stereotypical associations.

Recent research, however, has focused not only on documenting these biases, but also on how to reduce them. For instance, word embedding bias can be reduced by expanding text with sentences that counterbalance biases, or by identifying and “subtracting” the dimensions where bias manifests itself more strongly.

Another example of data-driven algorithmic bias is facial recognition systems. People studying the accuracy of these algorithms have found them to be less accurate at identifying darker faces, especially those of black women. This has motivated the creation of data sets that are more comprehensive in terms of demographic attributes, poses, and image quality, as well as the rise of auditing efforts designed to check and report on the accuracy and biases of facial recognition systems.

Another discussion on algorithmic bias involves the use of pretrial “risk assessment” tools. These are algorithms used to predict the probability that a defendant will reoffend (*recidivism*) or fail to appear in court. Pretrial risk assessment tools have become popular in the US, but they also have been found to show biases. In 2016, investigators working for ProPublica published an article based on “risk scores assigned to
more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014.”

They used that data to “see how many were charged with new crimes over the next two years,” which was “the same benchmark used by the creators of the algorithm.” They found that the algorithm “was particularly likely to falsely flag black defendants as future criminals, . . . at almost twice the rate as white defendants.” They also found that disparities could not be explained by prior crimes.

Unfortunately, biases are not unique to algorithms. Humans have them too. Scholars in the social sciences, for instance, have long studied the biases affecting job applications by looking at the callback rates for résumés with ethnically differentiated names or photographs. Thus, neither humans nor machines can guarantee fairness.

Here, we compare people’s reactions to cases of bias attributed to humans or machines. We present them in the context of college admissions, police enforcement, salaries, counseling, and human resources; in scenarios where humans or algorithms are the source of bias or the ones helping reduce bias. As in the previous chapter, we base our study on scenarios and measure people’s reactions to them using the following questions (as appropriate):

- Were the [person/algorithm]’s actions **harmful**? (from “Not harmful at all” to “Extremely harmful”)
- Would you **hire** this [person/algorithm] for a similar position? (from “Definitely not” to “Definitely yes”)
- Were the [person/algorithm]’s actions **intentional**? (from “Not intentional at all” to “Extremely intentional”)
- Do you **like** the [person/algorithm]?
  (from “Strongly dislike” to “Strongly like”)
- How **morally** wrong or right were the [person/algorithm]’s actions? (from “Extremely wrong” to “Extremely right”)
• Do you agree that the [person/algorithm] should be promoted to a position with more responsibilities? (from “Strongly disagree” to “Strongly agree”)
• Do you agree that the [person/algorithm] should be replaced by a(n) [algorithm/person] (replace different)? (from “Strongly disagree” to “Strongly agree”)
• Do you agree that the [person/algorithm] should be replaced by another [person/algorithm] (replace same)? (from “Strongly disagree” to “Strongly agree”)
• Do you think the [person/algorithm] is responsible for the action)? (from “Not responsible at all” to “Extremely responsible”)
• Do you think the [person/algorithm] is responsible for the [discriminatory/fair] outcome)? (from “Not responsible at all” to “Extremely responsible”)
• If you were in a similar situation as the [person/algorithm], would you have done the same? (from “Definitely not” to “Definitely yes”)

In addition to considering situations where a machine or a human either acted unfairly or corrected an unfair act, we considered variations in the ethnicity of the person being discriminated against (Hispanic, African American, or Asian). Different ethnicities are associated with different core stereotypes, so we expect different judgments in discriminatory situations.

In total, we considered a total of twenty-four possible scenarios and forty-eight conditions. In the next section, we document the results obtained for scenarios involving human resource (HR) screenings, college admissions, salary increases, and policing.

The four groups of scenarios are listed next.

* Certainly, we would have liked to consider more conditions, such as additional ethnicities and nonbinary gender identities, but that would have increased the number of scenarios and independent groups that we had to recruit to an unwieldy number. We leave the exercise of extending this analysis to more conditions for the future.
Human Resource Screenings

A company replaces their HR manager with a new [manager/algorithm] tasked with screening candidates for job interviews.

Unfair treatment
An audit reveals that the new [manager/algorithm] never selects [Hispanic/African American/Asian] candidates even when they have the same qualifications as other candidates.

Fair treatment
An audit reveals that the new [manager/algorithm] produces a fairer process for [Hispanic/African American/Asian] candidates, who were discriminated against by the previous system.

College Admissions

To improve their admissions process, a university hires a new [recruiter/algorithm] to evaluate the grades, test scores, and recommendation letters of applicants.
JUDGED BY MACHINES

Unfair treatment
An audit reveals that the new [recruiter/algorithm] is biased against [Hispanic/African American/Asian] applicants.

Fair treatment
An audit reveals that the new [recruiter/algorithm] is fairer to [Hispanic/African American/Asian] applicants, who were discriminated against by the previous system.

Salary Increases

A financial company hires a new [manager/algorithm] to decide the yearly salary increases of its employees.

Unfair treatment
An audit reveals that the new [manager/algorithm] consistently gives lower raises to [Hispanic/African American/Asian] employees, even when they are equal to other employees.

Fair treatment:
An audit reveals that the new [manager/algorithm] is fairer to [Hispanic/African American/Asian] employees, who were being discriminated against by the previous process.
CHAPTER 3

Policing

An audit reveals that the new squad has been detaining a disproportionately large percentage of innocent [Hispanics/African Americans/Asians].

An audit reveals that the new squad is fairer to innocent [Hispanic/African American/Asian], who had been detained in large numbers by the previous law enforcement procedures.

Figure 3.1 shows people’s reactions to the scenarios in which discrimination was observed. In all cases, humans are seen as more intentional, and also as more responsible for actions and outcomes. But beyond these obvious effects, we do observe some interesting, albeit relatively weak, patterns.

First, we find that—unlike most previous cases—moral judgments are not favorable to humans. In fact, we find that human actions are judged worse than machine actions (i.e., less moral), and are seen as more harmful in several scenarios, such as the college admissions and salary scenarios for African Americans and Hispanics. This provides additional evidence supporting the idea that reactions to machine actions are not
simply the result of a generalized bias against machines since these biases change with a scenario’s context and moral dimensions.

We also find small but interesting differences among the various ethnic groups depending on the particular scenario. The college admissions scenario elicits the strongest differences in judgment, especially for African Americans and Hispanics (who suffer more discrimination than Asians in contexts related to intellectual traits because of differences in stereotypes). Here, human actions are judged as relatively less moral and more harmful than the actions of machines. We also find that biases against African Americans and Hispanics result in slightly stronger differences in judgment between humans and machines compared to Asians. This suggests that differences in judgment between human and machine actions are slightly modulated by the ethnic group of the victim and the situation described in the scenario (e.g., college admissions). These differences aligns with our expectations for a US sample.

Moreover, people also think that the human should be replaced with another person. What is paradoxical, however, is that even though humans are seen as more intentional and more responsible than machines, people still prefer not to replace them with machines (as has been the case in all previous scenarios), adding further evidence in support of the idea of algorithm aversion.30

Figure 3.2 shows people’s reactions to scenarios in which discrimination was corrected. In general, we find a tendency for people to be more willing to promote humans, meaning that humans may receive more credit when they are involved in actions that correct unfair treatment. For the most part, however, we don’t find strong differences in judgment, except for the policing scenario, where the actions of humans are judged as much better than those of machines across several dimensions.
Figure 3.1

Participant reactions to four discrimination scenarios:
*Human Resource Screenings* (S19,S20,S21), *College Admissions* (S25,S26,S27), *Salary Increases* (S31,S32,S33) and *Policing* (S37,S38,S39).
College Admissions

Unfair treatment
CHAPTER 3

Salary Increases

_Unfair treatment_

---

Graph showing statistics for various categories:

- **S31**
  - Harm: p-value 1.2e-02
  - Hire: p-value 4.6e-01
  - Intentional: p-value 0
  - Like: p-value 4.2e-01
  - Moral: p-value 1.0e-03
  - Promote: p-value 3.3e-01
  - Replace Different: p-value 0
  - Replace Same: p-value 6.3e-06
  - Responsible for Action: p-value 1.6e-13
  - Responsible for Outcome: p-value 0
  - Similar Situation: p-value 6.2e-02

- **S32**
  - p-value: 2.2e-06

- **S33**
  - p-value: 4.0e-01

---

Legend:

- **Human**
- **Machine**
Policing

Unfair treatment
Human Resource Screenings

Fair treatment

**Figure 3.2**

Participant reactions to four corrected discrimination scenarios: *Human Resource Screenings* (S22,S23,S24), *College Admissions* (S28,S29,S30), *Salary Increases* (S34,S35,S36) and *Policing* (S40,S41,S42).
College Admissions

*Fair treatment*

<table>
<thead>
<tr>
<th>S28</th>
<th>S29</th>
<th>S30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harm</td>
<td>p-value: 9.3e-03</td>
<td>p-value: 5.4e-02</td>
</tr>
<tr>
<td>Hire</td>
<td>p-value: 6.0e-02</td>
<td>p-value: 2.0e-01</td>
</tr>
<tr>
<td>Intentional</td>
<td>p-value: 5</td>
<td>p-value: 7.0e-15</td>
</tr>
<tr>
<td>Like</td>
<td>p-value: 2.2e-01</td>
<td>p-value: 4.4e-02</td>
</tr>
<tr>
<td>Moral</td>
<td>p-value: 4.4e-01</td>
<td>p-value: 1.4e-01</td>
</tr>
<tr>
<td>Promote</td>
<td>p-value: 1.8e-02</td>
<td>p-value: 2.8e-01</td>
</tr>
<tr>
<td>Replace Different</td>
<td>p-value: 9.1e-16</td>
<td>p-value: 7.0e-16</td>
</tr>
<tr>
<td>Replace Same</td>
<td>p-value: 1.5e-01</td>
<td>p-value: 4.2e-01</td>
</tr>
<tr>
<td>Responsible for Action</td>
<td>p-value: 4.3e-09</td>
<td>p-value: 2.8e-06</td>
</tr>
<tr>
<td>Responsible for Outcome</td>
<td>p-value: 8.3e-11</td>
<td>p-value: 9.0e-07</td>
</tr>
<tr>
<td>Similar Situation</td>
<td>p-value: 2.1e-01</td>
<td>p-value: 6.0e-03</td>
</tr>
</tbody>
</table>
Salary Increases

*Fair treatment*

[Graph showing data with labels and p-values]
Policing

*Fair treatment*

**S40**
- Harm: p-value: 3.2e-01
- Hire: p-value: 2.5e-05
- Intentional: p-value: 6.6e-10
- Like: p-value: 9.0e-03
- Moral: p-value: 8.7e-03
- Promote: p-value: 3.8e-03
- Replace Different: p-value: 0
- Replace Same: p-value: 3.4e-02
- Responsible for Action: p-value: 8.5e-04
- Responsible for Outcome: p-value: 6.7e-07
- Similar Situation: p-value: 1.2e-02

**S41**
- p-value: 2.7e-01
- p-value: 3.0e-08
- p-value: 9.8e-13
- p-value: 1.9e-03
- p-value: 2.2e-04
- p-value: 4.6e-12
- p-value: 0
- p-value: 1.6e-04
- p-value: 5.6e-06

**S42**
- p-value: 5.0e-01
- p-value: 7.3e-05
- p-value: 7.5e-13
- p-value: 7.4e-03
- p-value: 1.6e-04
- p-value: 5.6e-06
- p-value: 0
- p-value: 5.9e-02
- p-value: 5.2e-02
- p-value: 5.6e-03
While biases can be problematic, the psychologists who have long studied them would be hard pressed to classify them as simple cognitive flaws. Instead, biases exist as rules of thumb or heuristics that evolved to make fast decisions in environments with limited information.\textsuperscript{31}

An example of these heuristics is the idea that people may perceive groups using two dominant characteristics: warmth and competence.\textsuperscript{32} This model predicts that groups high in warmth and low in competence (e.g., disabled people, babies, and the elderly) elicit sympathy, whereas groups low in warmth and high in competence elicit envy or jealousy.

Heuristics and stereotypes are certainly incorrect ways to judge individuals. Humans can have overgeneralized beliefs regarding members of a social group (stereotypes) and exhibit biased attitudes toward those groups (prejudice). Prejudice may then lead to unfair treatment or discrimination. But because heuristics work as a way to facilitate decision-making in information-deprived environments (or environments with excess information),\textsuperscript{33} it is not surprising that we find them in both humans and machines.

Similar to humans, cognitive machines are inferential and base their inferences on abstract forms of categorization (which is called \textit{stereotyping} in humans). In order to make predictions, machines often group and classify data using explicit and abstract features. Consider the idea of a principal component—a vector that accounts for most of the variance in a data set. Principal components are a common tool in machine learning, and they are similar to the idea of a stereotype, like classifying people using the vectors of warmth and competence. Unlike warmth and competence, however, principal components usually involve abstract features that are derived directly from data and can be difficult to interpret. This adds obscurity to algorithms and has led some people to advocate for increased transparency and interpretability as ways to mitigate algorithmic bias.
In fact, the use of explicit versus abstract features has been at the core of a nuanced discussion on the bias and fairness of algorithms lately. To avoid biases based on gender or race, scholars have proposed a variety of methods, from simply removing explicit demographic characteristics from a data set to predicting outcomes using only variables that are orthogonal to demographic characteristics. Yet recent research has shown that methods that tend to circumvent the possibility of bias may actually backfire because reaching fair outcomes is better served by using the most accurate predictors, even if these include explicit demographic information.34

In a recent paper on algorithmic fairness, Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan develop this idea by comparing an efficient and an equitable.35 These planners were tasked with admitting college applicants. The efficient planner was interested only in maximizing performance, measured by the grade point average (GPA) of the students admitted to college, while the equitable planner was interested in both performance and the racial composition of the admitted class.

To illustrate this idea, the scholars compared three methods: admissions that were blind to demographic variables (e.g., race was removed from the sample); admissions that included variables that were orthogonalized with respect to racial variables; and admissions that used racial variables explicitly. They report that the most equitable and efficient outcomes were reached using the model that explicitly included demographic variables.

To understand this distinction, consider students from two races: P (privileged) and U (underprivileged), who are applying to college. Because of their privileges, students in race P score higher in many of the variables that are predictive of future academic success, such as standardized test scores. Should we blind algorithms to race, then? Or is there a better solution?
Imagine a student from race U that obtains the same score as a student from race P on a standardized test. The student from race U was able to reach the same outcome as the student from race P in the absence of P’s privileges. Yet a model lacking an explicit racial variable will be unable to adjust for the lack of privilege affecting the scores of students from race U. A model that is blind to race will rate both students equally, and hence hurt the less privileged student. Instead, what the proposed theory suggests is to use the most accurate possible model (including racial variables, when relevant), and then setting different thresholds to achieve the desired level of equity (using, for instance, some of the definitions of fairness introduced earlier in this chapter).

This example illustrates the importance of separating the goals of equity and predictive accuracy. Even though it may be tempting to modify data to eliminate any trace of demographic characteristics, the best way to achieve efficient and equitable outcomes may be to treat prediction and equity as two separate parts of the same problem.

In the US, discriminatory treatment is not only frowned upon, but also illegal. Title VII of the 1964 Civil Rights Act is “a federal law that prohibits employers from discriminating against employees on the basis of sex, race, color, national origin, and religion.” The Supreme Court affirmed Title VII unanimously in 1971 in Griggs v. Duke Power Company, a class action suit claiming that Duke’s policies discriminated against African American employees. The court ruled that, independent of intent, discriminatory outcomes for protected classes violated Title VII.

In our data, we find important differences in the level of intent and responsibility assigned to discriminatory actions performed by humans and machines. However, in agreement with the Griggs decision, we find only small differences in moral judgment, suggesting that—in unfair cases—it is the outcome rather than the intention that is judged.

† It generally applies to employers with fifteen or more employees, including federal, state, and local governments.
The removal of intent from the legal judgment of bias has important implications for those working on the fairness of algorithms. It means that, even in the absence of intent, those creating the algorithms may be liable for biased outcomes.

This outcome-based approach to policing discrimination is opening a new market for an algorithm certification industry and discipline: a community focused on auditing the bias of algorithms and certifying them when they are not biased.

In the next chapter, we shift our gaze away from algorithmic bias and focus on another uncomfortable aspect of our digital reality: privacy. This will help us expand our understanding to another dimension of the way in which humans judge machines.