6

Testing discrimination in practice

In previous chapters, we have seen statistical, causal, and normative fairness criteria. This chapter is about the complexities that arise when we want to apply them in practice.

A running theme of this book is that there is no single test for fairness. Rather, there are many quantitative criteria that can be used to diagnose potential unfairness or discrimination.[1] There's often a gap between moral notions of fairness and what is measurable by available experimental or observational methods. This does not mean that we can select and apply a fairness test based on convenience. Far from it: we need moral reasoning and domain-specific considerations to determine which test(s) are appropriate, how to apply them, determine whether the findings indicate wrongful discrimination, and whether an intervention is called for. We will see examples of such reasoning throughout this chapter. Conversely, if a system passes a fairness test, we should not interpret it as a certificate that the system is fair.[2]

In this chapter, our primary objects of study will be real systems rather than models of systems. We must bear in mind that there are many necessary assumptions in creating a model which may not hold in practice. For example, so-called automated decision making systems rarely operate without any human judgment. Or, we may assume that a machine learning system is trained on a sample drawn from the same population on which it makes decisions, which is also almost never true in practice. Further, decision making in real life is rarely a single decision point, but rather a cumulative series of small decisions. For example, hiring includes sourcing, screening, interviewing, selection, and evaluation, and those steps themselves include many components.[1]

An important source of difficulty for testing discrimination in practice is that researchers have a limited ability to observe — much less manipulate — many of the steps in a real-world system. In fact, we'll see that even the decision maker faces limitations in its ability to study the system.

Despite these limitations and difficulties, empirically testing fairness is vital. The studies that we'll discuss serve as an existence proof of discrimination and provide a lower bound of its prevalence. They enable tracking trends in discrimination over time. When the findings are sufficiently blatant, they justify the need

[1] We will use the terms unfairness and discrimination roughly synonymously. There is no overarching definition of either term, but we will make our discussion precise by referring to a specific criterion whenever possible. Linguistically, the term discrimination puts more emphasis on the agency of the decision maker.

[2] We'll use "system" as a shorthand for a decision-making system, such as hiring at a company. It may or may not involve any automation or machine learning.

for intervention regardless of any differences in interpretation. And when we do apply a fairness intervention, they help us measure its effectiveness. Finally, empirical research can also help uncover the mechanisms by which discrimination takes place, which enables more targeted and effective interventions. This requires carefully formulating and testing hypotheses using domain knowledge.

The first half of this chapter surveys classic tests for discrimination that were developed in the context of human-decision making systems. The underlying concepts are just as applicable to the study of fairness in automated systems. Much of the first half will build on the causality chapter and explain concrete techniques including experiments, difference-in-differences, and regression discontinuity. While these are standard tools in the causal inference toolkit, we'll learn about the specific ways in which they can be applied to fairness questions. Then we will turn to the application of the observational criteria from Chapter 3. The summary table at the end of the first half lists, for each test, the fairness criterion that it probes, the type of access to the system that is required, and other nuances and limitations. The second half of the chapter is about testing fairness in algorithmic decision making, focusing on issues specific to algorithmic systems.

Part 1: Traditional tests for discrimination

Audit studies

The audit study is a popular technique for diagnosing discrimination. It involves a study design called a field experiment. "Field" refers to the fact that it is an experiment on the actual decision making system of interest (in the "field", as opposed to a lab simulation of decision making). Experiments on real systems are hard to pull off. For example, we usually have to keep participants unaware that they are in an experiment. But field experiments allow us to study decision making as it actually happens rather than worrying that what we're discovering is an artifact of a lab setting. At the same time, the experiment, by carefully manipulating and controlling variables, allows us to observe a treatment effect, rather than merely observing a correlation.

How to interpret such a treatment effect is a more tricky question. In our view, most audit studies, including the ones we'll describe, are best seen as attempts to test blindness: whether a decision maker directly uses a sensitive attribute. Recall that this notion of discrimination is not necessarily a counterfactual in a valid causal model (Chapter 5). Even as tests of blindness, there is debate about precisely what it is that they measure, since the researcher can at best signal race, gender, or another sensitive attribute. This will become clear when we discuss specific studies.

Audit studies were pioneered by the US Department of Housing and Urban Development in the 1970s for the purpose of studying the adverse treatment faced by minority home buyers and renters.[2] They have since been successfully applied to many other domains.

In one landmark study, researchers recruited 38 testers to visit about 150 car dealerships to bargain for cars, and record the price they were offered at the end of bargaining.[3] Testers visited dealerships in pairs; testers in a pair differed in terms of race or gender. Both testers in a pair bargained for the same model of car, at the same dealership, usually within a few days of each other.

Pulling off an experiment such as this in a convincing way requires careful attention to detail; here we describe just a few of the many details in the paper. Most significantly, the researchers went to great lengths to minimize any differences between the testers that might correlate with race or gender. In particular, all testers were 28–32 years old, had 3–4 years of postsecondary education, and "were subjectively chosen to have average attractiveness". Further, to minimize the risk of testers' interaction with dealers being correlated with race or gender, every aspect of their verbal or nonverbal behavior was governed by a script. For example, all testers "wore similar 'yuppie' sportswear and drove to the dealership in similar rented cars." They also had to memorize responses to a long list of questions they were likely to encounter. All of this required extensive training and regular debriefs.

The paper's main finding was a large and statistically significant price penalty in the offers received by Black testers. For example, Black males received final offers that were about $1,100 more than White males, which represents a threefold difference in dealer profits based on data on dealer costs. The analysis in the paper has alternative target variables (initial offers instead of final offers; percentage markup instead of dollar offers), alternate model specifications (e.g. to account the two audits in each pair having correlated noise), and additional controls (e.g. bargaining strategy). Thus, there are a number of different estimates, but the core findings remain robust.[3]

A tempting interpretation of this study is that if two people were identical except for race, with one being White and the other being Black, then the offers they should expect to receive would differ by about $1,100. But what does it mean for two people to be identical except for race? Which attributes about them would be the same, and which would be different?

With the benefit of the discussion of ontological instability in Chapter 5, we can understand the authors' implicit framework for making these decisions. In our view, they treat race as a stable source node in a causal graph, attempt to hold constant all of its descendants, such as attire and behavior, in order to estimate the direct effect of race on the outcome. But what if one of the mechanisms of what we understand as "racial discrimination" is based on attire and behavior differences? The social construction of race suggests that this is plausible.[4]

Note that the authors did not attempt to eliminate differences in accent between testers. Why not? From a practical standpoint, accent is difficult to manipulate. But a more principled defense of the authors' choice is that accent is a part of how

[3] In an experiment such as this where the treatment is randomized, the addition or omission of control variables in a regression estimate of the treatment effect does not result in an incorrect estimate, but control variables can explain some of the noise in the observations and thus increase the precision of the treatment effect estimate, i.e., decrease the standard error of the coefficient.

we understand race; a part of what it means to *be* Black, White, etc., so that even if the testers could manipulate their accents, they shouldn't. Accent is subsumed into the "race" node in the causal graph.

To take an informed stance on questions such as this, we need a deep understanding of cultural context and history. They are the subject of vigorous debate in sociology and critical race theory. Our point is this: the design and interpretation of audit studies requires taking positions on contested social questions. It may be futile to search for a single "correct" way to test even the seemingly straightforward fairness notion of whether the decision maker treats similar individuals similarly regardless of race. Controlling for a plethora of attributes is one approach; another is to simply recruit Black testers and White testers, have them behave and bargain as would be their natural inclination, and measure the demographic disparity. Each approach tells us something valuable, and neither is "better".[4]

Another famous audit study tested discrimination in the labor market.[5] Instead of sending testers in person, the researchers sent in fictitious resumes in response to job ads. Their goal was to test if an applicant's race had an impact on the likelihood of an employer inviting them for an interview. They signaled race in the resumes by using White-sounding names (Emily, Greg) or Black-sounding names (Lakisha, Jamal). By creating pairs of resumes that were identical except for the name, they found that White names were 50% more likely to result in a callback than Black names. The magnitude of the effect was equivalent to an additional eight years of experience on a resume.

Despite the study's careful design, debates over interpretation have inevitably arisen, primarily due to the use of candidate names as a way to signal race to employers. Did employers even notice the names in all cases, and might the effect have been even stronger if they had? Or, can the observed disparities be better explained based on factors correlated with race, such as a preference for more common and familiar names, or an inference of higher socioeconomic status for the candidates with White-sounding names? (Of course, the alternative explanations don't make the observed behavior morally acceptable, but they are important to consider.) Although the authors provide evidence against these interpretations, debate has persisted. For a discussion of critiques of the validity of audit studies, see Pager's survey.[6]

In any event, like other audit studies, this experiment tests fairness as blindness. Even simple proxies for race, such as residential neighborhood, were held constant between matched pairs of resumes. Thus, the design likely underestimates the extent to which morally irrelevant characteristics affect callback rates in practice. This is just another way to say that attribute flipping does not generally produce counterfactuals that we care about, and it is unclear if the effect sizes measured have any meaningful interpretation that generalizes beyond the context of the experiment.

Rather, audit studies are valuable because they trigger a strong and valid moral

[4]In most other domains, say employment, testing demographic disparity would be less valuable, because there are relevant differences between candidates. Price discrimination is unusual in that there are no morally salient qualities of buyers that may justify it.

intuition.[7] They also serve a practical purpose: when designed well, they illuminate the mechanisms that produce disparities and help guide interventions. For example, the car bargaining study concluded that the preferences of owners of dealerships don't explain the observed discrimination, that the preferences of other customers may explain some of it, and strong evidence that dealers themselves (rather than owners or customers) are the primary source of the observed discrimination.

Resume-based audit studies, also known as correspondence studies, have been widely replicated. We briefly present some major findings, with the caveat that there may be publication biases. For example, studies finding no evidence of an effect are in general less likely to be published. Alternately, published null findings might reflect poor experiment design, or might simply indicate that discrimination is only expressed in certain contexts.

A 2016 survey lists 30 studies from 15 countries covering nearly all continents revealing pervasive discrimination against racial and ethnic minorities.[8] The method has also been used to study discrimination based on gender, sexual orientation, and physical appearance.[8] It has also been used outside the labor market, in retail and academia.[8] Finally, trends over time have been studied: a meta-analysis found no change in racial discrimination in hiring against African Americans from 1989 to 2015. There was some indication of declining discrimination against Latinx Americans, although the data on this question was sparse.[9]

Collectively, audit studies have helped nudge the academic and policy debate away from the naive view that discrimination is a concern of a bygone era. From a methodological perspective, our main takeaway from the discussion of audit studies is the complexity of defining and testing blindness.

Testing the impact of blinding

In some situations, it is not possible to test blindness by randomizing the decision maker's perception of race, gender, or other sensitive attribute. For example, suppose we want to test if there's gender bias in peer review in a particular research field. Submitting real papers with fictitious author identities may result in the reviewer attempting to look up the author and realizing the deception. A design in which the researcher changes author names to those of real people is even more problematic.

There is a slightly different strategy that's more viable: an editor of a scholarly journal in the research field could conduct an experiment in which each paper received is randomly assigned to be reviewed in either a single-blind fashion (in which the author identities are known to the referees) or double-blind fashion (in which author identities are withheld from referees). Indeed, such experiments have been conducted,[10] but in general even this strategy can be impractical.

At any rate, suppose that a researcher has access to only observational data on journal review policies and statistics on published papers. Among ten journals in the research field, some introduced double-blind review, and did so in different years. The researcher observes that in each case, right after the switch, the fraction

of female-authored papers rose, whereas there was no change for the journals that stuck with single-blind review. Under certain assumptions, this enables estimating the impact of double-blind reviewing on the fraction of accepted papers that are female-authored. This hypothetical example illustrates the idea of a "natural experiment", so called because experiment-like conditions arise due to natural variation. Specifically, the study design in this case is called differences-in-differences. The first "difference" is between single-blind and double-blind reviewing, and the second "difference" is between journals (row 2 in the summary table).

Differences-in-differences is methodologically nuanced, and a full treatment is beyond our scope.[11] We briefly note some pitfalls. There may be unobserved confounders: perhaps the switch to double-blind reviewing at each journal happened as a result of a change in editorship, and the new editors also instituted policies that encouraged female authors to submit strong papers. There may also be spillover effects (which violates the Stable Unit Treatment Value Assumption): a change in policy at one journal can cause a change in the set of papers submitted to other journals. Outcomes are serially correlated (if there is a random fluctuation in the gender composition of the research field due to an entry or exodus of some researchers, the effect will last many years). This complicates the computation of the standard error of the estimate.[12] Finally, the effect of double blinding on the probability of acceptance of female-authored papers (rather than on the fraction of accepted papers that are female authored) is not identifiable using this technique without additional assumptions or controls.

Even though testing the impact of blinding sounds similar to testing blindness, there is a crucial conceptual and practical difference. Since we are not asking a question about the impact of race, gender, or another sensitive attribute, we avoid running into ontological instability. The researcher doesn't need to intervene on the observable features by constructing fictitious resumes or training testers to use a bargaining script. Instead, the natural variation in features is left unchanged; the study involves real decision subjects. The researcher only intervenes on the decision making procedure (or exploits natural variation) and evaluates the impact of that intervention on groups of candidates defined by the sensitive attribute A. Thus, A is not a node in a causal graph, but merely a way to split the units into groups for analysis. Questions of whether the decision maker actually inferred the sensitive attribute or merely a feature correlated with it are irrelevant to the interpretation of the study. Further, the effect sizes measured do have a meaning that generalizes to scenarios beyond the experiment. For example, a study tested the effect of "resume whitening", in which minority applicants deliberately conceal cues of their racial or ethnic identity in job application materials to improve their chances of getting a callback.[13] The effects reported in the study are meaningful to job seekers who engage in this practice.

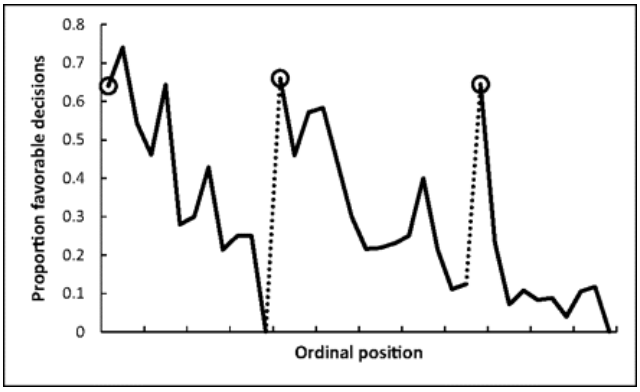

Figure 1: (from Danziger et al.): fraction of favorable rulings over the course of a day. The dotted lines indicate food breaks.

Revealing extraneous factors in decisions

Sometimes natural experiments can be used to show the arbitrariness of decision making rather than unfairness in the sense of non-blindness (row 3 in the summary table). Recall that arbitrariness is one type of unfairness that we are concerned about in this book (Chapter 2). Arbitrariness may refer to the lack of a uniform decision making procedure or to the incursion of irrelevant factors into the procedure.

For example, a study looked at decisions made by judges in Louisiana juvenile courts, including sentence lengths.[14] It found that in the week following an upset loss suffered by the Louisiana State University (LSU) football team, judges imposed sentences that were 7% longer on average. The impact was greater for Black defendants. The effect was driven entirely by judges who got their undergraduate degrees at LSU, suggesting that the effect is due to the emotional impact of the loss.[5]

Another well-known study on the supposed unreliability of judicial decisions is in fact a poster child for the danger of confounding variables in natural experiments. The study tested the relationship between the order in which parole cases are heard by judges and the outcomes of those cases.[15] It found that the percentage of favorable rulings started out at about 65% early in the day before gradually dropping to nearly zero right before the judges' food break, returned to ~65% after the break, with the same pattern repeated for the following food break! The authors suggested that judges' mental resources are depleted over the course of a session, leading to poorer decisions. It quickly became known as the "hungry judges" study and has been widely cited as an example of the fallibility of human decision makers.

The finding would be extraordinary if the order of cases was truly random.[6]

[5]For readers unfamiliar with the culture of college football in the United States, the paper helpfully notes that "Describing LSU football just as an event would be a huge understatement for the residents of the state of Louisiana."

[6]In fact, it would be so extraordinary that it has been argued that the study should be dismissed

The authors were well aware that the order wasn't random, and performed a few tests to see if it is associated with factors pertinent to the case (since those factors might also impact the probability of a favorable outcome in a legitimate way). They did not find such factors. But it turned out they didn't look hard enough. A follow-up investigation revealed multiple confounders and potential confounders, including the fact that prisoners without an attorney are presented last within each session, and tend to prevail at a much lower rate.[17] This invalidates the conclusion of the original study.

Testing the impact of decisions and interventions

An underappreciated aspect of fairness in decision making is the impact of the decision on the decision subject. In our prediction framework, the target variable (Y) is not impacted by the score or prediction (R). But this is not true in practice. Banks set interest rates for loans based on the predicted risk of default, but setting a higher interest rate makes a borrower more likely to default. The impact of the decision on the outcome is a question of causal inference.

There are other important questions we can ask about the impact of decisions. What is the utility or cost of a positive or negative decision to different decision subjects (and groups)? For example, admission to a college may have a different utility to different applicants based on the *other* colleges where they were or weren't admitted. Decisions may also have effects on people who are not decision subjects: for instance, incarceration impacts not just individuals but communities.[18] Measuring these costs allows us to be more scientific about setting decision thresholds and adjusting the tradeoff between false positives and negatives in decision systems.

One way to measure the impact of decisions is via experiments, but again, they can be infeasible for legal, ethical, and technical reasons. Instead, we highlight a natural experiment design for testing the impact of a decision — or a fairness intervention — on the candidates, called regression discontinuity (row 4 in the summary table).

Suppose we'd like to test if a merit-based scholarship program for first-generation college students has lasting beneficial effects — say, on how much they earn after college. We cannot simply compare the average salary of students who did and did not win the scholarship, as those two variables may be confounded by intrinsic ability or other factors. But suppose the scholarships were awarded based on test scores, with a cutoff of 85%. Then we can compare the salary of students with scores of 85% to 86% (and thus were awarded the scholarship) with those of students with scores of 84% to 85% (and thus were not awarded the scholarship). We may assume that within this narrow range of test scores, scholarships are awarded essentially randomly.[7] Thus we can estimate the impact of the scholarship as if we did a randomized controlled trial.

simply based on the fact that the effect size observed is far too large to be caused by psychological phenomena such as judges' attention. See[16]

[7]For example, if the variation (standard error) in test scores for students of identical ability is 5 percentage points, then the difference between 84% and 86% is of minimal significance.

We need to be careful, though. If we consider too narrow a band of test scores around the threshold, we may end up with insufficient data points for inference. If we consider a wider band of test scores, the students in this band may no longer be exchangeable units for the analysis.

Another pitfall arises because we assumed that the set of students who receive the scholarship is precisely those that are above the threshold. If this assumption fails, it immediately introduces the possibility of confounders. Perhaps the test score is not the only scholarship criterion, and income is used as a secondary criterion. Or, some students offered the scholarship may decline it because they already received another scholarship. Other students may not avail of the offer because the paperwork required to claim it is cumbersome. If it is possible to take the test multiple times, wealthier students may be more likely to do so until they meet the eligibility threshold.

Purely observational tests

The final category of quantitative tests for discrimination is purely observational. When we are not able to do experiments on the system of interest, nor have the conditions that enable quasi-experimental studies, there are still many questions we can answer with purely observational data.

One question that is often studied using observational data is whether the decision maker used the sensitive attribute; this can be seen as a loose analog of audit studies. This type of analysis is often used in the legal analysis of disparate treatment, although there is a deep and long-standing legal debate on whether and when explicit consideration of the sensitive attribute is necessarily unlawful.[19]

The most common way to do this is to use regression analysis to see if attributes other than the protected attributes can collectively "explain" the observed decisions[20] (row 5 in the summary table). If they don't, then the decision maker must have used the sensitive attribute. However, this is a brittle test. As discussed in Chapter 3, given a sufficiently rich dataset, the sensitive attribute can be reconstructed using the other attributes. It is no surprise that attempts to apply this test in a legal context can turn into dueling expert reports, as seen in the SFFA vs. Harvard case discussed in Chapter 5.

We can of course try to go deeper with observational data and regression analysis. To illustrate, consider the gender pay gap. A study might reveal that there is a gap between genders in wage per hour worked for equivalent positions in a company. A rebuttal might claim that the gap disappears after controlling for college GPA and performance review scores. Such studies can be seen as tests for *conditional demographic parity* (row 6 in the summary table).[8]

It can be hard to make sense of competing claims based on regression analysis. Which variables should we control for, and why? There are two ways in which we can put these observational claims on a more rigorous footing. The first is to use a

[8]Testing conditional demographic parity using regression requires strong assumptions about the functional form of the relationship between the independent variables and the target variable.

causal framework to make our claims more precise. In this case, causal modeling might alert us to unresolved questions: why do performance review scores differ by gender? What about the gender composition of different roles and levels of seniority? Exploring these questions may reveal unfair practices. Of course, in this instance the questions we raised are intuitively obvious, but other cases may be more intricate.

The second way to go deeper is to apply our normative understanding of fairness to determine which paths from gender to wage are morally problematic. If the pay gap is caused by the (well-known) gender differences in negotiating for pay raises, does the employer bear the moral responsibility to mitigate it? This is, of course, a normative and not a technical question.

Outcome-based tests

So far in this chapter we've presented many scenarios — screening job candidates, peer review, parole hearings — that have one thing in common: while they all aim to predict some outcome (job performance, paper quality, recidivism), the researcher does not have access to data on the true outcomes.

Lacking ground truth, the focus shifts to the observable characteristics at decision time, such as job qualifications. A persistent source of difficulty in these settings is for the researcher to construct two sets of samples that differ only in the sensitive attribute and not in any of the relevant characteristics. This is often an untestable assumption. Even in an experimental setting such as a resume audit study, there is substantial room for different interpretations: did employers infer race from names, or socioeconomic status? And in observational studies, the findings might turn out to be invalid because of unobserved confounders (such as in the hungry judges study).

But if outcome data are available, then we can do at least one test of fairness without needing any of the observable features (other than the sensitive attribute): specifically, we can test for sufficiency, which requires that the true outcome be conditionally independent of the sensitive attribute given the prediction ($Y \perp A|R$). For example, in the context of lending, if the bank's decisions satisfy sufficiency, then among applicants in any narrow interval of predicted probability of default (R), we should find the same rate of default (Y) for applicants of any group (A).

Typically, the decision maker (the bank) can test for sufficiency, but an external researcher cannot, since the researcher only gets to observe $\widehat{Y}$ and not R (i.e., whether or not the loan was approved). Such a researcher can test predictive parity rather than sufficiency. Predictive parity requires that the rate of default (Y) for favorably classified applicants ($\widehat{Y} = 1$) of any group (A) be the same. This observational test is called the *outcome test* (row 7 in the summary table).

Here is a tempting argument based on the outcome test: if one group (say women) who receive loans have a *lower* rate of default than another (men), it suggests that the bank applies a *higher* bar for loan qualification for women. Indeed, this type of argument was the original motivation behind the outcome test. But it is a logical fallacy; sufficiency does not imply predictive parity (or vice versa).

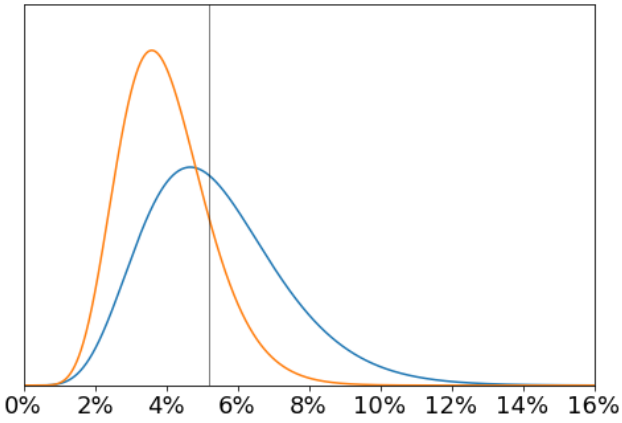

Figure 2: Hypothetical probability density of loan default for two groups, women (orange) and men (blue).

To see why, consider a thought experiment involving the Bayes optimal predictor. In the hypothetical figure below, applicants to the left of the vertical line qualify for the loan. Since the area under the curve to the left of the line is concentrated further to the right for men than for women, men who receive loans are more likely to default than women. Thus, the outcome test would reveal that predictive parity is violated, whereas it is clear from the construction that sufficiency is satisfied, and the bank applies the same bar to all groups.

This phenomenon is called *infra-marginality*, i.e., the measurement is aggregated over samples that are far from the decision threshold (margin). If we are indeed interested in testing sufficiency (equivalently, whether the bank applied the same threshold to all groups), rather than predictive parity, this is a problem. To address it, we can somehow try to narrow our attention to samples that are close to the threshold. This is not possible with $(\widehat{Y}, A, Y)$ alone: without knowing R, we don't know which instances are close to the threshold. However, if we also had access to some set of features X' (which need not coincide with the set of features X observed by the decision maker), it becomes possible to test for violations of sufficiency. The *threshold test* is a way to do this (row 8 in the summary table). A full description is beyond our scope.[21] One limitation is that it requires a model of the joint distribution of (X', A, Y) whose parameters can be inferred from the data, whereas the outcome test is model-free.

While we described infra-marginality as a limitation of the outcome test, it can also be seen as a benefit. When using a marginal test, we treat the distribution of applicant characteristics as a given, and miss the opportunity to ask: *why* are some individuals so far from the margin? Ideally, we can use causal inference to answer this question, but when the data at hand don't allow this, non-marginal tests might be a useful starting point for diagnosing unfairness that originates "upstream" of the decision maker. Similarly, error rate disparity, to which we will now turn, while crude by comparison to more sophisticated tests for discrimination, attempts to

capture some of our moral intuitions for why certain disparities are problematic.

Separation and selective labels

Recall that separation is defined as $R \perp A|Y$. At first glance, it seems that there is a simple observational test analogous to our test for sufficiency ($Y \perp A|R$). However, this is not straightforward, even for the decision maker, because outcome labels can be observed only for some of the applicants (i.e. the ones who received favorable decisions). Trying to test separation using this sample suffers from selection bias. This is an instance of what is called the *selective labels problem*. The issue also affects the computation of false positive and false negative rate parity, which are binary versions of separation.

More generally, the selective labels problem is the issue of selection bias in evaluating decision making systems due to the fact that the very selection process we wish to study determines the sample of instances that are observed. It is not specific to the issue of testing separation or error rates: it affects the measurement of other fundamental metrics such as accuracy as well. It is a serious and often overlooked issue that has been the subject of recent study.[22]

One way to get around this barrier is for the decision maker to employ an experiment in which some sample of decision subjects receive positive decisions regardless of the prediction (row 9 in the summary table). However, such experiments raise ethical concerns and are rarely done in practice. In machine learning, some experimentation is necessary in settings where there does not exist offline data for training the classifier, which must instead simultaneously learn and make decisions.[23]

One scenario where it is straightforward to test separation is when the "prediction" is not actually a prediction of a future event, but rather when machine learning is used for automating human judgment, such as harassment detection in online comments. In these applications, it is indeed possible and important to test error rate parity.

Summary of traditional tests and methods

Table 1: Summary of traditional tests and methods, highlighting the relationship to fairness, the observational and experimental access required by the researcher, and limitations.

	Test / study design	Fairness notion / application	Access	Notes / limitation
1	Audit study	Blindness	A-exp $:=$, $X :=$, R	Difficult to interpre
2	Natural experiment especially diff-in-diff	Impact of blinding	A-exp $\sim$, R	Confounding; SUTVA violations; ot
3	Natural experiment	Arbitrariness	$W \sim$, R	Unobserved confoun

	Test / study design	Fairness notion / application	Access	Notes / limitation
4	Natural experiment especially regr. disc.	Impact of decision	R, Y or Y'	Sample size; confounc other technical difficu
5	Regression analysis	Blindness	X, A, R	Unreliable due to pro
6	Regression analysis	Cond. demographic parity	X, A, R	Weak moral justificat
7	Outcome test	Predictive parity	$A, Y \mid \widehat{Y} = 1$	Infra-marginality
8	Threshold test	Sufficiency	$X', A, Y \mid \widehat{Y} = 1$	Model-specific
9	Experiment	Separation/error rate parity	$A, R, \widehat{Y} := , Y$	Often unethical or impr
10	Observational test	Demographic parity	A, R	See Chapter 3
11	Mediation analysis	"Relevant" mechanism	X, A, R	See Chapter 5

Legend:

- := indicates intervention on some variable (that is, $X :=$ does not represent a new random variable but is simply an annotation describing how X is used in the test)
- $\sim$ natural variation in some variable exploited by the researcher
- A-exp exposure of a signal of the sensitive attribute to the decision maker
- W a feature that is considered irrelevant to the decision
- X' a set of features which may not coincide with those observed by the decision maker |
- Y' an outcome that may or may not be the one that is the target of prediction |

Taste-based and statistical discrimination

We have reviewed several methods of detecting discrimination but we have not addressed the question of why discrimination happens. A long-standing way to try to answer this question from an economic perspective is to classify discrimination as *taste-based* or *statistical*. A taste-based discriminator is motivated by an irrational animus or prejudice for a group. As a result, they are willing to make sub-optimal decisions by passing up opportunities to select candidates from that group, even though they will incur a financial penalty for doing so. This is the classic model of discrimination in labor markets.[24]

A statistical discriminator, in contrast, aims to make optimal predictions about the target variable using all available information, including the protected attribute.[25,26] In the simplest model of statistical discrimination, two conditions hold: first, the distribution of the target variable differs by group. The usual example is of gender discrimination in the workplace, involving an employer who believes that women are more likely to take time off due to pregnancy (resulting in lower job performance). The second condition is that the observable characteristics do not allow a perfect prediction of the target variable, which is essentially always the case in practice. Under these two conditions, the optimal prediction will differ by group even when the relevant characteristics are identical. In this example, the employer would be less likely to hire a woman than an equally qualified man. There's a

nuance here: from a moral perspective we would say that the employer above discriminates against all female candidates. But under the definition of statistical discrimination, the employer only discriminates against the female candiates who would not have taken time off if hired (and in fact discriminates in favor of the female candidates who would take time off if hired).

While some authors put much weight understanding discrimination based on the taste-based vs. statistical categorization, we will de-emphasize it in this book. Several reasons motivate our choice. First, since we are interested in extracting lessons for statistical decision making systems, the distinction is not that helpful: such systems will not exhibit taste-based discrimination unless prejudice is explicitly programmed into them (while that is certainly a possibility, it is not a primary concern of this book).

Second, there are practical difficulties in distinguishing between taste-based and statistical discrimination. Often, what might seem to be a "taste" for discrimination is simply the result of an imperfect understanding of the decision-maker's information and beliefs. For example, at first sight the findings of the car bargaining study may look like a clear-cut case of taste-based discrimination. But maybe the dealer knows that different customers have different access to competing offers and therefore have different willingness to pay for the same item. Then, the dealer uses race as a proxy for this amount (correctly or not). In fact, the paper provides tentative evidence towards this interpretation. The reverse is also possible: if the researcher does not know the full set of features observed by the decision maker, taste-based discrimination might be mischaracterized as statistical discrimination.

Third, many of the fairness questions of interest to us, such as structural discrimination, don't map to either of these criteria (as they only consider causes that are relatively proximate to the decision point). We will discuss structural discrimination in Chapter 8.

Finally, the distinction is also not especially valuable from a normative perspective. Recall that our moral understanding of fairness emphasizes the effects on the decision subjects and does not put much weight on the mental state of the decision maker. It's also worth nothing that this dichotomy is associated with the policy position that fairness interventions are unnecessary — firms that practice taste-based discrimination will go out of business; as for statistical discrimination, either it is argued to be justified or futile to proscribe, because firms will find workarounds. [9] Of course, that's not necessarily a reason to avoid discussing taste-based and statistical discrimination, as the policy position in no way follows from the technical definitions and models themselves; it's just a relevant caveat for the reader who might encounter these dubious arguments in other sources.

Although we de-emphasize this distinction, we consider it critical to study the sources and mechanisms of discrimination. This helps us design effective and well-targeted interventions. For example, several studies (including the car bargaining study) test whether the source of discrimination lies in the owner, employees, or customers.

[9] For example, laws restricting employers from asking about applicants' criminal history resulted in employers using race as a proxy for it. See.[27]

An example of a study that can be difficult to interpret without understanding the mechanism is a 2015 resume-based audit study that revealed a 2:1 faculty preference for women for STEM tenure-track positions.[28] Consider the range of possible explanations: animus against men; a desire to compensate for past disadvantage suffered by women in STEM fields; a preference for a more diverse faculty (assuming that the faculties in question are currently male dominated); a response to financial incentives for diversification frequently provided by universities to STEM departments; and an assumption by decision makers that due to prior descrimination, a female candidate with an equivalent CV to a male candidate is of greater intrinsic ability.[10]

To summarize, rather than a one-size-fits-all approach to understanding mechanisms such as taste-based vs statistical discrimination, more useful is a nuanced and domain-specific approach where we formulate hypotheses in part by studying decision making processes and organizations, especially in a qualitative way. Let us now turn to those studies.

Studies of decision making processes and organizations

One way to study decision making processes is through surveys of decision makers or organizations. Sometimes such studies reveal blatant discrimination, such as strong racial preferences by employers.[29] Over the decades, however, such overt attitudes have become less common, or at least less likely to be expressed.[30] Discrimination tends to operate in more subtle, indirect, and covert ways.

Ethnographic studies excel at helping us understand covert discrimination. Ethnography is one of the main research methods in the social sciences and is based on the idea of the researcher being embedded among the research subjects for an extended period of time as they go about their daily activities. It is a set of qualitative methods that are complementary to and symbiotic with quantitative ones. Ethnography allows us to ask questions that are deeper than quantitative methods permit and to produce richly detailed accounts of culture. It also helps formulate hypotheses that can be tested quantitatively.

A good illustration is the book *Pedigree* which examines hiring practices in a set of elite consulting, banking, and law firms.[31] These firms together constitute the majority of the highest-paying and most desirable entry-level jobs for college graduates. The author used two standard ethnographic research methods. The first is a set of 120 interviews in which she presented as a graduate student interested in internship opportunities. The second method is called participant observation: she worked in an unpaid Human Resources position at one of the firms for 9 months, after obtaining consent to use her observations for research. There are several benefits to the researcher becoming a participant in the culture: it provides a greater level of access, allows the researcher to ask more nuanced questions, and

[10]Note that if this assumption is correct, then a preference for female candidates is both accuracy maximizing (as a predictor of career success) and required under some notions of fairness, such as counterfactual fairness.

makes it more likely that the research subjects would behave as they would when not being observed.

Several insights from the book are relevant to us. First, the hiring process has about nine stages, including outreach, recruitment events, screening, multiple rounds of interviews and deliberations, and "sell" events. This highlights why any quantitative study that focuses on a single slice of the process (say, evaluation of resumes) is limited in scope. Second, the process bears little resemblance to the ideal of predicting job performance based on a standardized set of attributes, albeit noisy ones, that we described in Chapter 1. Interviewers pay a surprising amount of attention to attributes that should be irrelevant or minimally relevant, such as leisure activities, but which instead serve as markers of class. Applicants from privileged backgrounds are more likely to be viewed favorably, both because they are able to spare more time for such activities, and because they have the insider knowledge that these seemingly irrelevant attributes matter in recruitment. The signals that firms do use as predictors of job performance, such as admission to elite universities — the *pedigree* in the book's title — are also highly correlated with socioeconomic status. The authors argue that these hiring practices help explain why elite status is perpetuated in society along hereditary lines. In our view, the careful use of statistical methods in hiring, despite their limits, may mitigate the strong social class based preferences exposed in the book.

Another book, *Inside Graduate Admissions*, focuses on education rather than labor market.[32] It resulted from the author's observations of decision making by graduate admissions committees in nine academic disciplines over two years. A striking theme that pervades this book is the tension between formalized and holistic decision making. For instance, committees arguably over-rely on GRE scores despite stating that they consider their predictive power to be limited. As it turns out, one reason for the preference for GRE scores and other quantitative criteria is that they avoid the difficulties of subjective interpretation associated with signals such as reference letters. This is considered valuable because it *minimizes tensions between faculty members* in the admissions process. On the other hand, decision makers are implicitly aware (and occasionally explicitly articulate) that if admissions criteria are too formal, then some groups of applicants — notably, applicants from China — would be successful at a far greater rate, and this is considered undesirable. This motivates a more holistic set of criteria, which often include idiosyncratic factors such as an applicant's hobby being considered "cool" by a faculty member. The author argues that admissions committees use a facially neutral set of criteria, characterized by an almost complete absence of explicit, substantive discussion of applicants' race, gender, or socioeconomic status, but which nonetheless perpetuates inequities. For example, there is a reluctance to take on students from underrepresented backgrounds whose profiles suggest that they would benefit from more intensive mentoring.

This concludes the first part of the chapter. Now let us turn to algorithmic systems. The background we've built up so far will prove useful. In fact, the traditional tests of discrimination are just as applicable to algorithmic systems. But we will also encounter many novel issues.

Part 2: Testing discrimination in algorithmic systems

An early example of discrimination in an algorithmic system is from the 1950s. In the United States, applicants for medical residency programs provide a ranked list of their preferred hospital programs to a centralized system, and hospitals likewise rank applicants. A matching algorithm takes these preferences as input and produces an assignment of applicants to hospitals that optimizes mutual desirability.[11]

Early versions of the system discriminated against couples who wished to stay geographically close, because couples could not accurately express their joint preferences: for example, each partner might prefer a hospital over all others but only if the other partner also matched to the same hospital.[33,34] This is a non-comparative notion of discrimination: the system does injustice to an applicant (or a couple) when it does not allow them to express their preferences, regardless of how other applicants are treated. Note that none of the tests for fairness that we have discussed are capable of detecting this instance of discrimination, as it arises because of dependencies between pairs of units, which is not something we have modeled.

There was a crude attempt in the residency matching system to capture joint preferences, involving designating one partner in each couple as the "leading member"; the algorithm would match the leading member without constraints and then match the other member to a proximate hospital if possible. Given the prevailing gender norms at that time, it is likely that this method had a further discriminatory impact on women in heterosexual couples.

Despite these early examples, it is the 2010s that testing unfairness in real-world algorithmic systems has become a pressing concern and a distinct area of research.[12] This work has much in common with the social science research that we reviewed, but the targets of research have expanded considerably. In the rest of this chapter, we will review and attempt to systematize the research methods in several areas of algorithmic decision making: various applications of natural-language processing and computer vision; ad targeting platforms; search and information retrieval tools; and online markets (ride hailing, vacation rentals, etc). Much of this research has focused on drawing attention to the discriminatory effects of specific, widely-used tools and platforms at specific points in time. While that is a valuable goal, we will aim to highlight broader, generalizable themes in our review. We will close the chapter by identifying common principles and methods behind this body of research.

[11] Specifically, it satisfies the requirement that if applicant A is *not* matched to hospital H, then either A matched to a hospital that he ranked higher than H, or H matched to a set of applicants all of whom it ranked higher than A.

[12] A 2014 paper issued a call to action towards this type of research. Most of the studies that we cite postdate that piece.[35]

Fairness considerations in applications of natural language processing

One of the most central tasks in NLP is language identification: determining the language that a given text is written in. It is a precursor to virtually any other NLP operation on the text such as translation to the user's preferred language on social media platforms. It is considered a more-or-less solved problem, with relatively simple models based on n-grams of characters achieving high accuracies on standard benchmarks, even for short texts that are a few words long.

However, a 2016 study showed that a widely used tool, langid.py, which incorporates a pre-trained model, had substantially more false negatives for tweets written in African-American English (AAE) compared to those written in more common dialectal forms: 13.2% of AAE tweets were classified as non-English compared to 7.6% of "White-aligned" English tweets. AAE is a set of English dialects commonly spoken by Black people in the United States (of course, there is no implication that all Black people in the United States primarily speak AAE or even speak it at all)[13]. The authors' construction of the AAE and White-aligned corpora themselves involved machine learning as well as validation based on linguistic expertise; we will defer a full discussion to the Measurement chapter. The observed error rate disparity is likely a classic case of underrepresentation in the training data.

Unlike the audit studies of car sales or labor markets discussed earlier, here it is not necessary (or justifiable) to control for any features of the texts, such as the level of formality. While it may certainly be possible to "explain" disparate error rates based on such features, that is irrelevant to the questions of interest in this context, such as whether NLP tools will perform less well for one group of users compared to another.

NLP tools range in their application from aids to online interaction to components of decisions with major career consequences. In particular, NLP is used in predictive tools for screening of resumes in the hiring process. There is some evidence of potential discriminatory impacts of such tools, both from employers themselves[37] and from applicants,[38] but it is limited to anecdotes. There is also evidence from the lab experiments on the task of predicting occupation from online biographies.[39]

We briefly survey other findings. Automated essay grading software tends to assign systematically lower scores to some demographic groups[40] compared to human graders, who may themselves provide biased ratings.[41] Hate speech detection models use markers of dialect as predictors of toxicity, according to a lab study,[42] resulting in discrimination against minority speakers. Many sentiment analysis tools assign systematically different scores to text based on race-aligned or gender-aligned names of people mentioned in the text.[43] Speech-to-text systems perform worse for speakers with certain accents.[44] In all these cases, the author or speaker of the text is potentially harmed. In other NLP systems, i.e., those involving

[13] For a treatise on AAE, see.[36] The linguistic study of AAE highlights the complexity and internal consistency of its grammar, vocabulary, and other distinctive features, and refutes the basis of prejudiced views of AAE as inferior to standard English.

natural language generation or translation, there is a different type of fairness concern, namely the generation of text reflecting cultural prejudices resulting in representational harm to a group of people.[45] The table below summarizes this discussion.

There is a line of research on cultural stereotypes reflected in word embeddings. Word embeddings are representations of linguistic units; they do not correspond to any linguistic or decision-making task. As such, lacking any notion of ground truth or harms to people, it is not meaningful to ask fairness questions about word embeddings without reference to specific downstream tasks in which they might be used. More generally, it is meaningless to ascribe fairness as an attribute of models as opposed to actions, outputs, or decision processes.

Table 2: **Four types of NLP tasks and the types of unfairness that can result.** Note that the traditional tests discussed in Part 1 operate in the context of predicting outcomes (row 3 in this table).

Type of task	Examples	Sources of disparity	Ha
Perception	Language id speech-to-text	Underrep. in training corpus	Degrade
Automating judgment	Toxicity detection essay grading	Human labels, underrep. in training corpus	Adverse
Predicting outcomes	Resume filtering	Various, including human labels	Adverse
Sequence prediction	Language generation translation	Cultural stereotypes, historical prejudices	Repres

Demographic disparities and questionable applications of computer vision

Like NLP, computer vision technology has made major headway in the 2010s due to the availability of large-scale training corpora and improvements in hardware for training neural networks. Today, many types of classifiers are used in commercial products to analyze images and videos of people. Unsurprisingly, they often exhibit disparities in performance based on gender, race, skin tone, and other attributes, as well as deeper ethical problems.

A prominent demonstration of error rate disparity comes from an analysis of three commercial tools designed to classify a person's gender as female or male based on an image, developed by Microsoft, IBM, and Face++ respectively.[46] The study found that all three classifiers perform better on male faces than female faces (8.1% – 20.6% difference in error rate). Further, all perform better on lighter faces than darker faces (11.8% – 19.2% difference in error rate), and worst on darker female faces (20.8% – 34.7% error rate). Finally, since all classifiers treat gender as binary, the error rate for people of nonbinary gender can be considered to be 100%.

If we treat the classifier's target variable as gender and the sensitive attribute as skin tone, we can decompose the observed disparities into two separate issues: first, female faces are classified as male more often than male faces are classified as female. This can be addressed relatively easily by recalibrating the classification threshold without changing the training process. The second and deeper issue is that darker faces are misclassified more often than lighter faces.

Image classification tools have found it particularly challenging to achieve geographic equity due to the skew in training datasets. A 2019 study evaluated five popular object recognition services on images of household objects from 54 countries.[47] It found significant accuracy disparities between countries, with images from lower-income countries being less accurately classified. The authors point out that household objects such as dish soap or spice containers tend to look very different in different countries. These issues are exacerbated when images of people are being classified. A 2017 analysis found that models trained on ImageNet and Open Images, two prominent datasets for object recognition, performed dramatically worse at recognizing images of bridegrooms from countries such as Pakistan and India compared to those from North American and European countries (the former were often classified as chain mail, a type of armor).[48]

Several other types of unfairness are known through anecdotal evidence in image classification and face recognition systems. At least two different image classification systems are known to have applied demeaning and insulting labels to photos of people.[49,50] Face recognition systems have been anecdotally reported to exhibit the cross-race effect wherein they are more likely to confuse faces of two people who are from a racial group that is underrepresented in the training data.[51] This possibility was shown in a simple linear model of face recognition as early as 1991.[52] Many commercial products have had difficulty detecting faces of darker-skinned people.[53,54] Similar results are known from lab studies of publicly available object detection models.[55]

More broadly, computer vision techniques seem to be particularly prone to use in ways that are fundamentally ethically questionable regardless of accuracy. Consider gender classification: while Microsoft, IBM, and Face++ have worked to mitigate the accuracy disparities discussed above, a more important question is why build a gender classification tool in the first place. By far the most common application appears to be displaying targeted advertisements based on inferred gender (and many other inferred characteristics, including age, race, and current mood) in public spaces, such as billboards, stores, or screens in the back seats of taxis. We won't recap the objections to targeted advertising here, but it is an extensively discussed topic, and the practice is strongly opposed by the public, at least in the United States.[56]

Morally dubious computer vision technology goes well beyond this example, and includes apps that "beautify" images of users' faces, i.e., edit them to better conform to mainstream notions of attractiveness; emotion recognition, which has been alleged to be a pseudoscience; and the analysis of video footage for cues such as body language for screening job applicants.[57]

Search and recommendation systems: three types of harms

Search engines, social media platforms, and recommendation systems have different goals and underlying algorithms, but they do have many things in common from a fairness perspective. They are not decision systems and don't provide or deny people opportunities, at least not directly. Instead, there are (at least) three types of disparities and attendant harms that may arise in these systems. First, they may serve the informational needs of some consumers (searchers or users) better than others. Second, they may create inequities among *producers* (content creators) by privileging certain content over others. Third, they may create representational harms by amplifying and perpetuating cultural stereotypes. There are a plethora of other ethical concerns about information platforms, such as the potential to contribute to the political polarization of society. However, we will limit our attention to harms that can be considered to be forms of discrimination.

Unfairness to consumers. An illustration of unfairness to consumers comes from a study of collaborative filtering recommender systems that used theoretical and simulation methods (rather than a field study of a deployed system).[58] Collaborative filtering is an approach to recommendations that is based on the explicit or implicit feedback (e.g. ratings and consumption, respectively) provided by other users of the system. The intuition behind it is seen in the "users who liked this item also liked..." feature on many services. The study found that such systems can underperform for minority groups in the sense of being worse at recommending content that those users would like. A related but distinct reason for underperformance occurs when users from one group are less observable, e.g., less likely to provide ratings. The underlying assumption is that different groups have different preferences, so that what the system learns about one group doesn't generalize to other groups.

In general, this type of unfairness is hard to study in real systems (not just by external researchers but also by system operators themselves). The main difficulty is accurately measuring the target variable. The relevant target construct from a fairness perspective is users' satisfaction with the results or how well the results served the users' needs. Metrics such as clicks and ratings serve as crude proxies for the target, and are themselves subject to demographic measurement biases. Companies do expend significant resources on A/B testing or other experimental methods for optimizing search and recommendation systems, and frequently measure demographic differences as well. But to reiterate, such tests almost always emphasize metrics of interest to the firm rather than benefit or payoff for the user.

A rare attempt to transcend this limitation comes from an (internal) audit study of the Bing search engine.[59] The authors devised methods to disentangle user satisfaction from other demographic-specific variation by controlling for the effects of demographic factors on behavioral metrics. They combined it with a method for inferring latent differences directly instead of estimating user satisfaction for each demographic group and then comparing these estimates. This method infers which impression, among a randomly selected pair of impressions, led to greater user satisfaction. They did this using proxies for satisfaction such as reformulation

rate. Reformulating a search query is a strong indicator of dissatisfaction with the results. Based on these methods, they found no gender differences in satisfaction but mild age differences.

Unfairness to producers. In 2019, a group of content creators sued YouTube alleging that YouTube's algorithms as well as human moderators suppressed the reach of LGBT-focused videos and the ability to earn ad revenue from them. This is a distinct type of issue from that discussed above, as the claim is about a harm to producers rather than consumers (although, of course, YouTube viewers interested in LGBT content are also presumably harmed). There are many other ongoing allegations and controversies that fall into this category: partisan bias in search results and social media platforms, search engines favoring results from their own properties over competitors, fact-checking of online political ads, and inadequate (or, conversely, over-aggressive) policing of purported copyright violations. It is difficult to meaningfully discuss and address these issues through the lens of fairness and discrimination rather than a broader perspective of power and accountability. The core issue is that when information platforms have control over public discourse, they become the arbiters of conflicts between competing interests and viewpoints. From a legal perspective, these issues fall primarily under antitrust law and telecommunication regulation rather than antidiscrimination law. [14]

Representational harms. The book *Algorithms of Oppression* drew attention to the ways in which search engines reinforce harmful racial, gender, and intersectional stereotypes.[64] There have also been quantitative studies of some aspects of these harms. In keeping with our quantitative focus, let's discuss a study that measured how well the gender skew in Google image search results for 45 occupations (*author*, *bartender*, *construction worker* ...) corresponded to the real-world gender skew of the respective occupations.[65] This can be seen as test for calibration.[15] The study found weak evidence for stereotype exaggeration, that is, imbalances in occupational statistics are exaggerated in image search results. However, the deviations were minor.

Consider a thought experiment: suppose the study had found no evidence of miscalibration. Is the resulting system fair? It would be simplistic to answer in the affirmative for at least two reasons. First, the study tested calibration between image search results and occupational statistics *in the United States*. Gender stereotypes of occupations as well as occupational statistics differ substantially between countries and cultures. Second, accurately reflecting real-world statistics may still constitute a representational harm when those statistics are skewed and themselves reflect a history of prejudice. Such a system contributes to the lack of visible role models for underrepresented groups. To what extent information platforms should bear responsibility for minimizing these imbalances, and what types of interventions are justified, remain matters of debate.

[14] For in-depth treatments of the history and politics of information platforms, see:[60,61,62,63]

[15] Specifically, instances are occupations and the fraction of women in the search results is viewed as a predictor of the fraction of women in the occupation in the real world.

Understanding unfairness in ad targeting

Ads have long been targeted in relatively crude ways. For example, a health magazine might have ads for beauty products, exploiting a coarse correlation. In contrast to previous methods, online targeting offers several key advantages to advertisers: granular data collection about individuals, the ability to reach niche audiences (in theory, the audience size can be one, since ad content can be programmatically generated and customized with user attributes as inputs), and the ability to measure conversion (conversion is when someone who views the ad clicks on it, and then takes another action such as a purchase). To date, ad targeting has been one of the most commercially impactful applications of machine learning.

The complexity of modern ad targeting results in many avenues for disparities in the demographics of ad views, which we will study. But it is not obvious how to connect these disparities to fairness. After all, many types of demographic targeting such as clothing ads by gender are considered innocuous.

There are two frameworks for understanding potential harms from ad targeting. The first framework sees ads as unlocking opportunities for their recipients, because they provide information that the viewer might not have. This is why targeting employment or housing ads based on protected categories may be unfair and unlawful. The domains where targeting is legally prohibited broadly correspond to those which impact civil rights, and reflect the complex histories of discrimination in those domains, as discussed in Chapter 6.

The second framework views ads as tools of persuasion rather than information dissemination. In this framework, harms arise from ads being manipulative — that is, exerting covert influence instead of making forthright appeals — or exploiting stereotypes.[66] Users are harmed by being targeted with ads that provide them negative utility, as opposed to the first framework, in which the harm comes from missing out on ads with positive utility. The two frameworks don't necessarily contradict each other. Rather, individual ads or ad campaigns can be seen as either primarily informational or primarily persuasive, and accordingly, one or the other framework might be appropriate for analysis.[16]

There is a vast literature on how race and gender are portrayed in ads that we consider to fall under the persuasion framework.[17] However, this line of inquiry has yet to turn its attention to online targeted advertising, which has the potential for accentuating the harms of manipulation and stereotyping by targeting specific people and groups. Thus, the empirical research that we will highlight falls under the informational framework.

There are roughly three mechanisms by which the same targeted ad may reach one group more often than another. The most obvious is the use of explicit targeting criteria by advertisers: either the sensitive attribute itself or a proxy for it (such as ZIP code as a proxy for race). For example, Facebook allows thousands of targeting categories, including categories that are automatically constructed by the system

[16] The economic analysis of advertising includes a third category, complementary, that's related to persuasive or manipulative category.[67]

[17] See, e.g.[68]

based on users' free-form text descriptions of their interests. These categories were found to include "Jew haters" and many other antisemitic terms.[69] The company has had difficulty eliminating even direct proxies for sensitive categories, resulting in repeated exposés.

The second disparity-producing mechanism is optimization of click rate (or another measure of effectiveness), which is one of the core goals of algorithmic targeting. Unlike the first category, this does not require explicit intent by the advertiser or the platform. The algorithmic system may predict a user's probability of engaging with an ad based on her past behavior, her expressed interests, and other factors (including, potentially, explicitly expressed sensitive attributes).

The third mechanism is market effects: delivering an ad to different users may cost the advertiser different amounts. For example, some researchers have observed that women cost more to advertise to than men and hypothesized that this is because women clicked on ads more often, leading to a higher measure of effectiveness.[70,71] Thus if the advertiser simply specifies a total budget and leaves the delivery up to the platform (which is a common practice), then the audience composition will vary depending on the budget: smaller budgets will result in the less expensive group being overrepresented.

In terms of methods to detect these disparities, researchers and journalists have used broadly two approaches: interact with the system either as a user or as an advertiser. Tschantz et al. created simulated users that had the "gender" attribute in Google's Ad settings page set to female or male, and found that Google showed the simulated male users ads from a certain career coaching agency that promised large salaries more frequently than the simulated female users.[72] While this type of study establishes that employment ads through Google's ad system are not blind to gender (as expressed in the ad settings page), it cannot uncover the mechanism, i.e., distinguish between explicit targeting by the advertiser and platform effects of various kinds.

Interacting with ad platforms as an advertiser has proved to be a more fruitful approach so far, especially to analyze Facebook's advertising system. This is because Facebook exposes vastly more details about its advertising system to advertisers than to users. In fact, it allows advertisers to learn more information it has inferred or purchased about a user than it will allow the user himself to access.[73] The existence of anti-semitic auto-generated targeting categories, mentioned above, was uncovered using the advertiser interface. Ad delivery on Facebook has been found to introduce demographic disparities due to both market effects and effectiveness optimization effects.[70] To reiterate, this means that even if the advertiser does not explicitly target an ad by (say) gender, there may be a systematic gender skew in the ad's audience. The optimization effects are enabled by Facebook's analysis of the contents of ads. Interestingly, this includes image analysis, which researchers revealed using the clever technique of serving ads with transparent content that is invisible to humans but nonetheless had an effect on ad delivery.[70]

Fairness considerations in the design of online marketplaces

Online platforms for ride hailing, short-term housing, and freelance (gig) work have risen to prominence in the 2010s: notable examples are Uber, Lyft, Airbnb, and TaskRabbit. They are important targets for the study of fairness because they directly impact people's livelihoods and opportunities. We will set aside some types of markets from our discussion. Online dating apps share some similarities with these markets, but they require an entirely separate analysis because the norms governing romance are different from those governing commerce and employment.[74] Then there are marketplaces for goods such as Amazon and eBay. In these markets the characteristics of the participants are less salient than the attributes of the product, so discrimination is less of a concern.[18]

Unlike the domains studied so far, machine learning is not a core component of the algorithms in online marketplaces. (Nonetheless, we consider it in scope because of our broad interest in decision making and fairness, rather than just machine learning.) Therefore fairness concerns are less about training data or algorithms; the far more serious issue is discrimination by buyers and sellers. For example, one study found that Uber drivers turned off the app in areas where they did not want to pick up passengers.[76]

Methods to detect discrimination in online marketplaces are fairly similar to traditional settings such as housing and employment; a combination of audit studies and observational methods have been used. A notable example is a field experiment targeting Airbnb.[77] The authors created fake guest accounts whose names signaled race (African-American or White) and gender (female or male), but were otherwise identical. Twenty different names were used: five in each combination of race and gender. They then contacted the hosts of 6,400 listings in five cities through these accounts to inquire about availability. They found a 50% probability of acceptance of inquiries from guests with White-sounding names, compared to 42% for guests with African-American-sounding names. The effect was persistent regardless of the host's race, gender, and experience on the platform, as well as listing type (high or low priced; entire property or shared), and diversity of the neighborhood. Note that the accounts did not have profile pictures; if inference of race by hosts happens in part based on appearance, a study design that varied the accounts' profile pictures might find a greater effect.

Compared to traditional settings, some types of observational data are readily available on online platforms, which can be useful to the researcher. In the above study, the public availability of reviews of listed properties proved useful. It was not essential to the design of the study, but allowed an interesting validity check. When the analysis was restricted to the 29% hosts in the sample who had received at least one review from an African-American guest, the racial disparity in responses declined sharply. If the study's findings were a result of a quirk of the experimental design, rather than actual racial discrimination by Airbnb hosts, it would be difficult to explain why the effect would disappear for this subset of

[18] This is not to say that discrimination is nonexistent. See, e.g.,.[75]

hosts. This supports the study's external validity.

In addition to discrimination by participants, another fairness issue that many online marketplaces must contend with is geographic differences in effectiveness. One study of TaskRabbit and Uber found that neighborhoods with high population density and high-income neighborhoods receive the largest benefits from the sharing economy.[78] Due to the pervasive correlation between poverty and race/ethnicity, these also translate to racial disparities.

Of course, geographic and structural disparities in these markets are not caused by online platforms, and no doubt exist in offline analogs such as word-of-mouth gig work. In fact, the magnitude of racial discrimination is much larger in scenarios such as hailing taxis on the street[79] compared to technologically mediated interactions. However, in comparison to markets regulated by antidiscrimination law, such as hotels, discrimination in online markets is more severe. In any case, the formalized nature of online platforms makes audits easier. As well, the centralized nature of these platforms is a powerful opportunity for fairness interventions.

There are many ways in which platforms can use design to minimize users' ability to discriminate (such as by withholding information about counterparties) and the impetus to discriminate (such as by making participant characteristics less salient compared to product characteristics in the interface).[80] There is no way for platforms to take a neutral stance towards discrimination by participants: even choices made without explicit regard for discrimination can affect how vulnerable users are to bias.

As a concrete example, the authors of the Airbnb study recommend that the platform withhold guest information from hosts prior to booking. (Note that ride hailing services do withhold customer information. Carpooling services, on the other hand, allow users to view names when selecting matches; unsurprisingly, this enables discrimination against ethnic minorities.[81]) The authors of the study on geographic inequalities suggest, among other interventions, that ride hailing services provide a "geographic reputation" score to drivers to combat the fact that drivers often incorrectly perceive neighborhoods to be more dangerous than they are.

Mechanisms of discrimination

We've looked at a number of studies on detecting unfairness in algorithmic systems. Let's take stock.

In the introductory chapter we discussed, at a high-level, different ways in which unfairness could arise in machine learning systems. Here, we see that the specific sources and mechanisms of unfairness can be intricate and domain-specific. Researchers need an understanding of the domain to effectively formulate and test hypotheses about sources and mechanisms of unfairness.

For example, consider the study of gender classification systems discussed above. It is easy to guess that unrepresentative training datasets contributed to the observed accuracy disparities, but unrepresentative in what way? A follow-

up paper considered this question.[82] It analyzed several state-of-the-art gender classifiers (in a lab setting, as opposed to field tests of commercial APIs in the original paper) and argued that underrepresentation of darker skin tones in the training data is *not* a reason for the observed disparity. Instead, one mechanism suggested by the authors is based on the fact that many training datasets of human faces comprise photos of celebrities. [19] They found that photos of female celebrities have more prominent makeup compared to photos of women in general. This led to classifiers using makeup as a proxy for gender in a way that didn't generalize to the rest of the population.

Slightly different hypotheses can produce vastly different conclusions, especially in the presence of complex interactions between content producers, consumers, and platforms. For example, one study tested claims of partisan bias by search engines, as well as related claims that search engines return results that reinforce searchers' existing views (the "filter bubble" hypothesis).[84] The researchers recruited participants with different political views, collected Google search results on a political topic in both standard and incognito windows from those participants' computers, and found that standard (personalized) search results were no more partisan than incognito (non-personalized) ones, seemingly finding evidence against the claim that online search reinforces users' existing beliefs.

This finding is consistent with the fact that Google doesn't personalize search results except based on searcher location and immediate (10-minute) history of searches. This is known based on Google's own admission[85] and prior research.[86]

However, a more plausible hypothesis for the filter bubble effect in search comes from a qualitative study.[87] Simplified somewhat for our purposes, it goes as follows: when an event with political significance unfolds, key influencers (politicians, partisan news outlets, interest groups, political message boards) quickly craft their own narratives of the event. Those narratives selectively reach their respective partisan audiences through partisan information networks. Those people then turn to search engines to learn more or to "verify the facts". Crucially, however, they use different search terms to refer to the same event, reflecting the different narratives to which they have been exposed.[20] The results for these different search terms are often starkly different, because the producers of news and commentary selectively and strategically cater to partisans using these same narratives. Thus, searchers' beliefs are reinforced. Note that this filter-bubble-producing mechanism operates effectively even though the search algorithm itself is arguably neutral.[21]

A final example to reinforce the fact that disparity-producing mechanisms can

[19] This overrepresentation is because photos of celebrities are easier to gather publicly, and celebrities are thought to have weakened privacy rights due to the competing public interest in their activities. However, for a counterpoint, see.[83]

[20] For example, in 2017, US president Donald Trump called for the National Football League to fire players who engaged in a much-publicized political protest during games. Opposing narratives of this event were that NFL viewership had declined due to fans protesting players' actions, or that it had increased despite the protests. Search terms reflecting these views might be "NFL ratings down" versus "NFL ratings up".

[21] But see[88] ("Data Void Type #4: Fragmented Concepts") for an argument that search engines' decision not to collapse related concepts contributes to this fragmentation.

be subtle and that domain expertise is required to formulate the right hypothesis: an investigation by journalists found that *staples.com* showed discounted prices to individuals in some ZIP codes; these ZIP codes were, on average, wealthier.[89] However, the actual pricing rule that explained most of the variation, as they reported, was that if there was a competitor's physical store located within 20 miles or so of the customer's inferred location, then the customer would see a discount! Presumably this strategy is intended to infer the customer's reservation price or willingness to pay. Incidentally, this is a similar kind of "statistical discrimination" as seen in the car sales discrimination study discussed at the beginning of this chapter.

Fairness criteria in algorithmic audits

While the mechanisms of unfairness are different in algorithmic systems, the applicable fairness criteria are the same for algorithmic decision making as other kinds of decision making. That said, some fairness notions are more often relevant, and others less so, in algorithmic decision making compared to human decision making. We offer a few selected observations on this point.

Fairness as blindness is seen less often in audit studies of algorithmic systems; such systems are generally designed to be blind to sensitive attributes. Besides fairness concerns often arise precisely from the fact that blindness is generally not an effective fairness intervention in machine learning. Two exceptions are ad targeting and online marketplaces (where the non-blind decisions are in fact being made by users and not the platform).

Unfairness as arbitrariness. There are roughly two senses in which decision making could be considered arbitrary and hence unfair. The first is when decisions are made on a whim rather than a uniform procedure. Since automated decision making results in procedural uniformity, this type of concern is generally not salient.

The second sense of arbitrariness applies even when there is a uniform procedure, if that procedure relies on a consideration of factors that are thought to be irrelevant, either statistically or morally. Since machine learning excels at finding correlations, it commonly identifies factors that seem puzzling or blatantly unacceptable. For example, in aptitude tests such as the Graduate Record Examination, essays are graded automatically. Although e-rater and other tools used for this purpose are subject to validation checks, and are found to perform similarly to human raters on samples of actual essays, they are able to be fooled into giving perfect scores to machine-generated gibberish. Recall that there is no straightforward criterion that allows us to assess if a feature is morally valid (Chapter 2), and this question must be debated on a case-by-case basis.

More serious issues arise when classifiers are not even subjected to proper validity checks. For example, there are a number of companies that claim to predict candidates' suitability for jobs based on personality tests or body language and other characteristics in videos.[57] There is no peer-reviewed evidence that job

performance is predictable using these factors, and no basis for such a belief. Thus, even if these systems don't produce demographic disparities, they are unfair in the sense of being arbitrary: candidates receiving an adverse decision lack due process to understand the basis for the decision, contest it, or determine how to improve their chances of success.

Observational fairness criteria including demographic parity, error rate parity, and calibration have received much attention in algorithmic fairness studies. Convenience has probably played a big role in this choice: these metrics are easy to gather and straightforward to report without necessarily connecting them to moral notions of fairness. We reiterate our caution about the overuse of parity-based notions; parity should rarely be made a goal by itself. At a minimum, it is important to understand the sources and mechanisms that produce disparities as well as the harms that result from them before deciding on appropriate interventions.

Representational harms. Traditionally, allocative and representational harms were studied in separate literatures, reflecting the fact that they are mostly seen in separate spheres of life (for instance, housing discrimination versus stereotypes in advertisements). Many algorithmic systems, on the other hand, are capable of generating both types of harms. A failure of face recognition for darker-skinned people is demeaning, but it could also prevent someone from being able to access a digital device or enter a building that uses biometric security.

Information flow, fairness, privacy

A notion called information flow is seen frequently in algorithmic audits. This criterion requires that sensitive information about subjects not flow from one information system to another, or from one part of a system to another. For example, a health website may promise that user activity, such as searches and clicks, are not shared with third parties such as insurance companies (since that may lead to potentially discriminatory effects on insurance premiums). It can be seen as a generalization of blindness: whereas blindness is about not acting on available sensitive information, restraining information flow ensures that the sensitive information is not available to act upon in the first place.

There is a powerful test for testing violations of information flow constraints, which we will call the adversarial test.[72] It does not directly detect information flow, but rather decisions that are made on the basis of that information. It is powerful because it does not require specifying a target variable, which minimizes the domain knowledge required of the researcher. To illustrate, let's revisit the example of the health website. The adversarial test operates as follows:

1. Create two groups of simulated users (A and B), i.e., bots, that are identical except for the fact that users in group A, but not group B, browse the sensitive website in question.
2. Have both groups of users browse *other* websites that are thought to serve ads from insurance companies, or personalize content based on users' interests, or somehow tailor content to users based on health information. This is

the key point: the researcher does not need to hypothesize a mechanism by which potentially unfair outcomes result — e.g. which websites (or third parties) might receive sensitive data, whether the personalization might take the form of ads, prices, or some other aspect of content.

3. Record the contents of the web pages seen by all users in the previous step.
4. Train a binary classifier to distinguish between web pages encountered by users in group A and those encountered by users in group B. Use cross-validation to measure its accuracy.
5. If the information flow constraint is satisfied (i.e., the health website did not share any user information with any third parties), then the websites browsed in step 2 are blind to user activities in step 1; thus the two groups of users look identical, and there is no way to systematically distinguish the content seen by group A from that seen by group B. The classifier's test accuracy should not significantly exceed $\frac{1}{2}$. The permutation test can be used to quantify the probability that the classifier's observed accuracy (or better) could have arisen by chance if there is in fact no systematic difference between the two groups.[90]

There are additional nuances relating to proper randomization and controls, for which we refer the reader to the study.[72] Note that if the adversarial test fails to detect an effect, it does not mean that the information flow constraint is satisfied. Also note that the adversarial test is not capable of measuring an effect size. Such a measurement would be meaningless anyway, since the goal is to detect information flow, and any effect on observable behavior of the system is merely a proxy for it.

This view of information flow as a generalization of blindness reveals an important connection between privacy and fairness. Many studies based on this principle can be seen as either privacy or fairness investigations. For example, a study found that Facebook solicits phone numbers from users with the stated purpose of improving account security, but uses those numbers for ad targeting.[91] This is an example of undisclosed information flow from one part of the system to another. Another study used ad retargeting — in which actions taken on one website, such as searching for a product, result in ads for that product on another website — to infer the exchange of user data between advertising companies.[92] Neither study used the adversarial test.

Comparison of research methods

For auditing user fairness on online platforms, there are two main approaches: creating fake profiles and recruiting real users as testers. Each has its pros and cons. Both approaches have the advantage, compared to traditional audit studies, of allowing a potentially greater scale due to the ease of creating fake accounts or recruiting testers online (e.g. through crowd-sourcing).

Scaling is especially relevant for testing geographic differences, given the global reach of many online platforms. It is generally possible to simulate geographically dispersed users by manipulating testing devices to report faked locations. For

example, the above-mentioned investigation of regional price differences on *staples.com* actually included a measurement from each of the 42,000 ZIP codes in the United States.[93] They accomplished this by observing that the website stored the user's inferred location in a cookie, and proceeding to programmatically change the value stored in the cookie to each possible value.

That said, practical obstacles commonly arise in the fake-profile approach. In one study, the number of test units was practically limited by the requirement for each account to have a distinct credit card associated with it.[94] Another issue is bot detection. For example, the Airbnb study was limited to five cities, even though the researchers originally planned to test more, because the platform's bot-detection algorithms kicked in during the course of the study to detect and shut down the anomalous pattern of activity. It's easy to imagine an even worse outcome where accounts detected as bots are somehow treated differently by the platform (e.g. messages from those accounts are more likely to be hidden from intended recipients), compromising the validity of the study.

As this example illustrates, the relationship between audit researchers and the platforms being audited is often adversarial. Platforms' efforts to hinder researchers can be technical but also legal. Many platforms, notably Facebook, prohibit both fake-account creation and automated interaction in their Terms of Service. The ethics of Terms-of-Service violation in audit studies is a matter of ongoing debate, paralleling some of the ethical discussions during the formative period of traditional audit studies. In addition to ethical questions, researchers incur a legal risk when they violate Terms of Service. In fact, under laws such as the US Computer Fraud and Abuse Act, it is possible that they may face criminal as opposed to just civil penalties.

Compared to the fake-profile approach, recruiting real users allows less control over profiles, but is better able to capture the natural variation in attributes and behavior between demographic groups. Thus, neither design is always preferable, and they are attuned to different fairness notions. When testers are recruited via crowd-sourcing, the result is generally a convenience sample (i.e. the sample is biased towards people who are easy to contact), resulting in a non-probability (non-representative) sample. It is generally infeasible to train such a group of testers to carry out an experimental protocol; instead, such studies typically handle the interaction between testers and the platform via software tools (e.g. browser extensions) created by the researcher and installed by the tester. For more on the difficulties of research using non-probability samples, see the book Bit by Bit.[95]

Due to the serious limitations of both approaches, lab studies of algorithmic systems are commonly seen. The reason that lab studies have value at all is that since automated systems are fully specified using code, the researcher can hope to simulate them relatively faithfully. Of course, there are limitations: the researcher typically doesn't have access to training data, user interaction data, or configuration settings. But simulation is a valuable way for developers of algorithmic systems to test their *own* systems, and this is a common approach in the industry. Companies often go so far as to make de-identified user interaction data publicly available so that external researchers can conduct lab studies to develop and test algorithms.

The Netflix Prize is a prominent example of such a data release.[96] So far, these efforts have almost always been about improving the accuracy rather than the fairness of algorithmic systems.

Lab studies are especially useful for getting a handle on questions that cannot be studied by other empirical methods, notably the *dynamics* of algorithmic systems, i.e., their evolution over time. One prominent result from this type of study is the quantification of feedback loops in predictive policing.[97,98] Another insight is the increasing homogeneity of users' consumption patterns over time in recommender systems.[99]

Observational studies and observational fairness criteria continue to be important. Such studies are typically carried out by algorithm developers or decision makers, often in collaboration with external researchers.[100,101] It is relatively rare for observational data to be made publicly available. A rare exception, the COMPAS dataset, involved a Freedom of Information Act request.

Finally, it is worth reiterating that quantitative studies are narrow in what they can conceptualize and measure. Qualitative and ethnographic studies of decision makers thus provide an invaluable complementary perspective. To illustrate, we'll discuss one study that reports on six months of ethnographic fieldwork in a corporate data science team.[102] The team worked on a project in the domain of car financing that aimed to "improve the quality" of leads (leads are potential car buyers in need of financing who might be converted to actual buyers through marketing). Given such an amorphous high-level goal, formulating a concrete and tractable data science problem is a necessary and nontrivial task — a task that is further complicated by the limitations of the data available. The paper documents how there is substantial latitude in problem formulation, and spotlights the iterative process that was used, resulting in the use of a series of proxies for lead quality. The authors show that different proxies have different fairness implications: one proxy would maximize people's lending opportunities and another would alleviate dealers' existing biases, both potentially valuable fairness goals. However, the data scientists were not aware of the normative implications of their decisions and did not explicitly deliberate them.

Looking ahead

In this chapter, we covered traditional tests for discrimination as well as fairness studies of various algorithmic systems. Together, these methods constitute a powerful toolbox for interrogating a single decision system at a single point in time. But there are other types of fairness questions we can ask: what is the cumulative effect of the discrimination faced by a person over the course of a lifetime? What structural aspects of society result in unfairness? We cannot answer such a question by looking at individual systems. The next chapter is all about broadening our view of discrimination and then using that broader perspective to study a range of possible fairness interventions.

Bibliography

[1] Miranda Bogen and Aaron Rieke. Help wanted: an examination of hiring algorithms, equity, and bias. Technical report, Technical report, Upturn, 2018.

[2] Ronald E Wienk, Clifford E. Reid, John C. Simonson, and Frederick J. Eggers. Measuring racial discrimination in american housing markets: The housing market practices survey. 1979.

[3] Ian Ayres and Peter Siegelman. Race and gender discrimination in bargaining for a new car. *The American Economic Review*, pages 304–321, 1995.

[4] Jonathan B Freeman, Andrew M Penner, Aliya Saperstein, Matthias Scheutz, and Nalini Ambady. Looking the part: Social status cues shape race perception. *PloS one*, 6(9):e25107, 2011.

[5] Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013, 2004.

[6] Devah Pager. The use of field experiments for studies of employment discrimination: Contributions, critiques, and directions for the future. *The Annals of the American Academy of Political and Social Science*, 609(1):104–133, 2007.

[7] Issa Kohler-Hausmann. Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.*, 113:1163, 2018.

[8] Marianne Bertrand and Esther Duflo. Field experiments on discrimination. In *Handbook of economic field experiments*, volume 1, pages 309–393. Elsevier, 2017.

[9] Lincoln Quillian, Devah Pager, Ole Hexel, and Arnfinn H Midtbøen. Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences*, 114(41):10870–10875, 2017.

[10] Rebecca M Blank. The effects of double-blind versus single-blind reviewing: Experimental evidence from the american economic review. *The American Economic Review*, pages 1041–1067, 1991.

[11] Jorn-Steffen Pischke. Empirical methods in applied economics: Lecture notes. 2005.

[12] Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *The Quarterly journal of economics*, 119(1):249–275, 2004.

[13] Sonia K Kang, Katherine A DeCelles, András Tilcsik, and Sora Jun. Whitened resumes: Race and self-presentation in the labor market. *Administrative Science Quarterly*, 61(3):469–502, 2016.

[14] Ozkan Eren and Naci Mocan. Emotional judges and unlucky juveniles. *American Economic Journal: Applied Economics*, 10(3):171–205, 2018.

[15] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892, 2011.

[16] Daniel Lakens. Impossibly hungry judges. https://daniellakens.blogspot.com/2017/07/impossibly-hungry-judges.html, 2017.

[17] Keren Weinshall-Margel and John Shapard. Overlooked factors in the analysis of parole decisions. *Proceedings of the National Academy of Sciences*, 108(42):E833–E833, 2011.

[18] Aziz Z Huq. Racial equity in algorithmic criminal justice. *Duke LJ*, 68:1043, 2018.

[19] Helen Norton. The supreme court's post-racial turn towards a zero-sum understanding of equality. *Wm. & Mary L. Rev.*, 52:197, 2010.

[20] Ian Ayres. Three tests for measuring unjustified disparate impacts in organ transplantation: The problem of" included variable" bias. *Perspectives in biology and medicine*, 48(1):68–S87, 2005.

[21] Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

[22] Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 275–284. ACM, 2017.

[23] Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz, and Hanna Wallach. Exploring or exploiting? social and ethical implications of autonomous experimentation in ai. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2016.

[24] Gary S. Becker. *The Economics of Discrimination*. University of Chicago Press, 1957.

[25] Edmund S. Phelps. The statistical theory of racism and sexism. *The American Economic Review*, 62(4):659–661, 1972.

[26] Kenneth Arrow et al. The theory of discrimination. *Discrimination in labor markets, Achenfelter, A. Ress (eds.)*, 1973.

[27] Amanda Agan and Sonja Starr. Ban the box, criminal records, and racial discrimination: A field experiment. *The Quarterly Journal of Economics*, 133(1):191–235, 2017.

[28] Wendy M Williams and Stephen J Ceci. National hiring experiments reveal 2: 1 faculty preference for women on stem tenure track. *Proceedings of the National Academy of Sciences*, 112(17):5360–5365, 2015.

[29] Kathryn M Neckerman and Joleen Kirschenman. Hiring strategies, racial bias, and inner-city workers. *Social problems*, 38(4):433–447, 1991.

[30] Devah Pager and Hana Shepherd. The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annu. Rev. Sociol*, 34:181–209, 2008.

[31] Lauren A Rivera. *Pedigree: How elite students get elite jobs*. Princeton University Press, 2016.

[32] Julie R Posselt. *Inside graduate admissions*. Harvard University Press, 2016.

[33] Alvin E Roth. The origins, history, and design of the resident match. *Jama*, 289(7):909–912, 2003.

[34] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347, 1996.

[35] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *ICA Pre-Conference on Data and Discrimination*, 2014.

[36] Lisa J Green. *African American English: a linguistic introduction*. Cambridge University Press, 2002.

[37] Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*, 2018.

[38] Stephen Buranyi. How to persuade a robot that you should get the job, 2018.

[39] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128. ACM, 2019.

[40] Chaitanya Ramineni and David Williamson. Understanding Mean Score Differences Between the e-rater Automated Scoring Engine and Humans for Demographically Based Groups in the GRE General Test. *ETS Research Report Series*, 2018(1):1–31, 2018.

[41] Evelin Amorim, Marcia Cançado, and Adriano Veloso. Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 229–237, 2018.

[42] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, 2019.

[43] Svetlana Kiritchenko and Saif M Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*, 2018.

[44] Rachael Tatman. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain, April 2017. Association for Computational Linguistics.

[45] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, and Jasmine Wang. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*, 2019.

[46] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proc. 1st Conference on Fairness, Accountability and Transparency (FAccT)*, pages 77–91, 2018.

[47] Terrance de Vries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019.

[48] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D. Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. In *NIPS 2017 workshop: Machine Learning for the Developing World*, 2017.

[49] Tom Simonite. When it comes to gorillas, google photos remains blind. *Wired, January*, 13, 2018.

[50] Alex Hern. Flickr faces complaints over 'offensive'auto-tagging for photos. *The Guardian*, 20, 2015.

[51] Paris Martineau. Cities examine proper—and improper—uses of facial recognition | wired. https://www.wired.com/story/cities-examine-proper-improper-facial-recognition/, 2019.

[52] Alice J O'Toole, Kenneth Deffenbacher, Hervé Abdi, and James C Bartlett. Simulating the 'other-race effect'as a problem in perceptual learning. *Connection Science*, 3(2):163–178, 1991.

[53] Adam Frucci. Hp face-tracking webcams don't recognize black people. https://gizmodo.com/hp-face-tracking-webcams-dont-recognize-black-people-5431190, 2009.

[54] Jane McEntegart. Kinect may have issues with dark-skinned users | tom's guide. https://www.tomsguide.com/us/Microsoft-Kinect-Dark-Skin-Facial-Recognition,news-8638.html, 2010.

[55] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. *arXiv preprint arXiv:1902.11097*, 2019.

[56] Joseph Turow, Jennifer King, Chris Jay Hoofnagle, Amy Bleakley, and Michael Hennessy. Americans reject tailored advertising and three activities that enable it. *Available at SSRN 1478214*, 2009.

[57] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. Mitigating bias in algorithmic employment screening: Evaluating claims and practices. *arXiv preprint arXiv:1906.09208*, 2019.

[58] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems*, pages 2921–2930, 2017.

[59] Rishabh Mehrotra, Ashton Anderson, Fernando Diaz, Amit Sharma, Hanna Wallach, and Emine Yilmaz. Auditing search engines for differential satisfaction across demographics. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 626–633, 2017.

[60] Tim Wu. *The master switch: The rise and fall of information empires*. Vintage, 2010.

[61] Tarleton Gillespie. The politics of 'platforms'. *New media & society*, 12(3):347–364, 2010.

[62] Tarleton Gillespie. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[63] Kate Klonick. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.*, 131:1598, 2017.

[64] Safiya Umoja Noble. *Algorithms of Oppression: How search engines reinforce racism*. NYU Press, 2018.

[65] Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proc. 33rd Conference on Human Factors in Computing Systems*, pages 3819–3828. ACM, 2015.

[66] Daniel Susser, Beate Roessler, and Helen Nissenbaum. Online manipulation: Hidden influences in a digital world. *Available at SSRN 3306006*, 2018.

[67] Kyle Bagwell. The economic analysis of advertising. *Handbook of industrial organization*, 3:1701–1844, 2007.

[68] Scott Coltrane and Melinda Messineo. The perpetuation of subtle prejudice: Race and gender imagery in 1990s television advertising. *Sex roles*, 42(5-6):363–389, 2000.

[69] Julia Angwin, Madeleine Varner, and Ariana Tobin. Facebook enabled advertisers to reach "jew haters". ProPublica. https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters, 2017.

[70] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. Discrimination through optimization: How facebook's ad delivery can lead to biased outcomes. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):199, 2019.

[71] Anja Lambrecht and Catherine Tucker. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*, 2019.

[72] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proc. Privacy Enhancing Technologies (PET)*, 2015(1):92–112, 2015.

[73] Athanasios Andreou, Oana Goga, Krishna Gummadi, Patrick Loiseau, and Alan Mislove. Adanalyst. https://adanalyst.mpi-sws.org/, 2017.

[74] Jevan A Hutson, Jessie G Taft, Solon Barocas, and Karen Levy. Debiasing desire: addressing bias & discrimination on intimate platforms. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):73, 2018.

[75] Ian Ayres, Mahzarin Banaji, and Christine Jolls. Race effects on ebay. *The RAND Journal of Economics*, 46(4):891–917, 2015.

[76] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1603–1612. ACM, 2015.

[77] Benjamin Edelman, Michael Luca, and Dan Svirsky. Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, 9(2):1–22, 2017.

[78] Jacob Thebault-Spieker, Loren Terveen, and Brent Hecht. Toward a geographic understanding of the sharing economy: Systemic biases in uberx and taskrabbit. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(3):1–40, 2017.

[79] Yanbo Ge, Christopher R Knittel, Don MacKenzie, and Stephen Zoepf. Racial and gender discrimination in transportation network companies. Technical report, National Bureau of Economic Research, 2016.

[80] Karen Levy and Solon Barocas. Designing against discrimination in online markets. *Berkeley Tech. LJ*, 32:1183, 2017.

[81] Jasper Dag Tjaden, Carsten Schwemmer, and Menusch Khadjavi. Ride with me—ethnic discrimination, social markets, and the sharing economy. *European Sociological Review*, 34(4):418–432, 2018.

[82] Vidya Muthukumar, Tejaswini Pedapati, Nalini Ratha, Prasanna Sattigeri, Chai-Wah Wu, Brian Kingsbury, Abhishek Kumar, Samuel Thomas, Aleksandra Mojsilovic, and Kush R Varshney. Understanding unequal gender classification accuracy from face images. *arXiv preprint arXiv:1812.00099*, 2018.

[83] Adam Harvey and Jules LaPlace. Megapixels: Origins, ethics, and privacy implications of publicly available face recognition image datasets, 2019.

[84] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):148, 2018.

[85] Jillian D'Onfro. Google tests changes to its search algorithm; how search works. https://www.cnbc.com/2018/09/17/google-tests-changes-to-its-search-algorithm-how-search-works.html, 2019.

[86] Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. Measuring personalization of web search. In *Proceedings of the 22nd international conference on World Wide Web*, pages 527–538. ACM, 2013.

[87] Francesca Tripodi. Searching for alternative facts: Analyzing scriptural inference in conservative news practices. *Data & Society*, 2018.

[88] M Golebiewski and D Boyd. Data voids: where missing data can easily be exploited. *Data & Society*, 29, 2018.

[89] Jennifer Valentino-Devries, Jeremy Singer-Vine, and Ashkan Soltani. Websites vary prices, deals based on users' information. *Wall Street Journal*, 10:60–68, 2012.

[90] Markus Ojala and Gemma C Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11(Jun):1833–1863, 2010.

[91] Giridhari Venkatadri, Elena Lucherini, Piotr Sapiezynski, and Alan Mislove. Investigating sources of pii used in facebook's targeted advertising. *Proceedings on Privacy Enhancing Technologies*, 2019(1):227–244, 2019.

[92] Muhammad Ahmad Bashir, Sajjad Arshad, William Robertson, and Christo Wilson. Tracing information flows between ad exchanges using retargeted ads. In *USENIX Security Symposium 16*, pages 481–496, 2016.

[93] Jeremy Singer-Vine, Jennifer Valentino-DeVries, and Ashkan Soltani. How the journal tested prices and deals online. Wall Street Journal. http://blogs.wsj.com/digits/2012/12/23/how-the-journal-tested-prices-and-deals-online, 2012.

[94] Le Chen, Alan Mislove, and Christo Wilson. Peeking beneath the hood of uber. In *Proceedings of the 2015 Internet Measurement Conference*, pages 495–508. ACM, 2015.

[95] Matthew Salganik. *Bit by bit: Social research in the digital age*. Princeton University Press, 2019.

[96] James Bennett and Stan Lanning. The netflix prize. In *Proceedings of KDD cup and workshop*, volume 2007, page 35. New York, NY, USA., 2007.

[97] Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016.

[98] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847*, 2017.

[99] Allison JB Chaney, Brandon M Stewart, and Barbara E Engelhardt. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 224–232. ACM, 2018.

[100] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[101] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148, 2018.

[102] Samir Passi and Solon Barocas. Problem formulation and fairness. In *Proc. 2nd Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 39–48, 2019.