FAIRNESS AND MACHINE LEARNING

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

Created: Wed 16 Jun 2021 01:46:08 PM PDT

Contents

1

About the book

This book gives a perspective on machine learning that treats fairness as a central concern rather than an afterthought. We'll review the practice of machine learning in a way that highlights ethical challenges. We'll then discuss approaches to mitigate these problems.

We've aimed to make the book as broadly accessible as we could, while preserving technical rigor and confronting difficult moral questions that arise in algorithmic decision making.

This book won't have an all-encompassing formal definition of fairness or a quick technical fix to society's concerns with automated decisions. Addressing issues of fairness requires carefully understanding the scope and limitations of machine learning tools. This book offers a critical take on current practice of machine learning as well as proposed technical fixes for achieving fairness. It doesn't offer any easy answers. Nonetheless, we hope you'll find the book enjoyable and useful in developing a deeper understanding of how to practice machine learning responsibly.

Why now?

Machine learning has made rapid headway into socio-technical systems ranging from video surveillance to automated resume screening. Simultaneously, there has been heightened public concern about the impact of digital technology on society.

These two trends have led to the rapid emergence of fairness, accountability, transparency in socio-technical systems as a research field. While exciting, this has led to a proliferation of terminology, rediscovery and simultaneous discovery, conflicts between disciplinary perspectives, and other types of confusion.

This book aims to move the conversation forward by synthesizing long-standing bodies of knowledge, such as causal inference, with recent work in the community, sprinkled with a few observations of our own.

How did the book come about?

In the fall semester of 2017, the three authors each taught courses on fairness and ethics in machine learning: Barocas at Cornell, Hardt at Berkeley, and Narayanan at Princeton. We each approached the topic from a different perspective. We also presented two tutorials: Barocas and Hardt at NIPS 2017, and Narayanan at FAT* 2018. This book emerged from the notes we created for these three courses, and is the result of an ongoing dialog between us.

Who is this book for?

We've written this book to be useful for multiple audiences. You might be a student or practitioner of machine learning facing ethical concerns in your daily work. You might also be an ethics scholar looking to apply your expertise to the study of emerging technologies. Or you might be a citizen concerned about how automated systems will shape society, and wanting a deeper understanding than you can get from press coverage.

We'll assume you're familiar with introductory computer science and algorithms. Knowing how to code isn't strictly necessary to read the book, but will let you get the most out of it. We'll also assume you're familiar with basic statistics and probability. Throughout the book, we'll include pointers to introductory material on these topics.

On the other hand, you don't need any knowledge of machine learning to read this book: we've included an appendix that introduces basic machine learning concepts. We've also provided a basic discussion of the philosophical and legal concepts underlying fairness.[1]

[1] These haven't yet been released.

What's in this book?

This book is intentionally narrow in scope: you can see an outline here. Most of the book is about fairness, but we include a chapter[2] that touches upon a few related concepts: privacy, interpretability, explainability, transparency, and accountability. We omit vast swaths of ethical concerns about machine learning and artificial intelligence, including labor displacement due to automation, adversarial machine learning, and AI safety.

[2] This chapter hasn't yet been released.

Similarly, we discuss fairness interventions in the narrow sense of fair decision-making. We acknowledge that interventions may take many other forms: setting better policies, reforming institutions, or upending the basic structures of society.

A narrow framing of machine learning ethics might be tempting

to technologists and businesses as a way to focus on technical interventions while sidestepping deeper questions about power and accountability. We caution against this temptation. For example, mitigating racial disparities in the accuracy of face recognition systems, while valuable, is no substitute for a debate about whether such systems should be deployed in public spaces and what sort of oversight we should put into place.

About the authors

Solon Barocas is an Assistant Professor in the Department of Information Science at Cornell University. His research explores ethical and policy issues in artificial intelligence, particularly fairness in machine learning, methods for bringing accountability to automated decision-making, and the privacy implications of inference. He was previously a Postdoctoral Researcher at Microsoft Research, where he worked with the Fairness, Accountability, Transparency, and Ethics in AI group, as well as a Postdoctoral Research Associate at the Center for Information Technology Policy at Princeton University. Barocas completed his doctorate at New York University, where he remains a visiting scholar at the Center for Urban Science + Progress.

Moritz Hardt is an Assistant Professor in the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley. Hardt investigates algorithms and machine learning with a focus on reliability, validity, and societal impact. After obtaining a PhD in Computer Science from Princeton University, he held positions at IBM Research Almaden, Google Research and Google Brain.

Arvind Narayanan is an Associate Professor of Computer Science at Princeton. He studies the risks associated with large datasets about people: anonymity, privacy, and bias. He leads the Princeton Web Transparency and Accountability Project to uncover how companies collect and use our personal information. His doctoral research showed the fundamental limits of de-identification. He co-created a Massive Open Online Course as well as a textbook on Bitcoin and cryptocurrency technologies. Narayanan is a recipient of the Presidential Early Career Award for Scientists and Engineers.

Thanks and acknowledgements

This book wouldn't have been possible without the profound contributions of our collaborators and the community at large.

We are grateful to our students for their active participation in pilot courses at Berkeley, Cornell, and Princeton. Thanks in particular

to Claudia Roberts for lecture notes of the Princeton course.

Special thanks to Katherine Yen for editorial and technical help with the book.

Moritz Hardt is indebted to Cynthia Dwork for introducing him to the topic of this book during a formative internship in 2010.

We benefitted from substantial discussions, feedback and comments from Rediet Abebe, Andrew Brunskill, Aylin Caliskan, Frances Ding, Michaela Hardt, Lily Hu, Ben Hutchinson, Lauren Kaplan, Niki Kilbertus, Kathy Kleiman, Issa Kohler-Hausmann, Eric Lawrence, Zachary Lipton, Lydia T. Liu, John Miller, Smitha Milli, Shira Mitchell, Robert Netzorg, Juan Carlos Perdomo, Claudia Roberts, Olga Russakovsky, Matthew J. Salganik, Carsten Schwemmer, Ludwig Schmidt, Matthew Sun, Angelina Wang, Annette Zimmermann, Tijana Zrnic.

2
Introduction

Our success, happiness, and wellbeing are never fully of our own making. Others' decisions can profoundly affect the course of our lives: whether to admit us to a particular school, offer us a job, or grant us a mortgage. Arbitrary, inconsistent, or faulty decision-making thus raises serious concerns because it risks limiting our ability to achieve the goals that we have set for ourselves and access the opportunities for which we are qualified.

So how do we ensure that these decisions are made the right way and for the right reasons? While there's much to value in fixed rules, applied consistently, *good* decisions take available evidence into account. We expect admissions, employment, and lending decisions to rest on factors that are relevant to the outcome of interest.

Identifying details that are relevant to a decision might happen informally and without much thought: employers might observe that people who study math seem to perform particularly well in the financial industry. But they could test these observations against historical evidence by examining the degree to which one's major correlates with success on the job. This is the traditional work of statistics—and it promises to provide a more reliable basis for decision-making by quantifying how much weight to assign certain details in our determinations.

Decades of research have compared the accuracy of statistical models to the judgments of humans, even experts with years of experience, and found that in many situations data-driven decisions trounce those based on intuition or expertise.[3] These results have been welcomed as a way to ensure that the high-stakes decisions that shape our life chances are both accurate and fair.

[3] Dawes, Faust, and Meehl, "Clinical Versus Actuarial Judgment," *Science* 243, no. 4899 (1989): 1668–74.

Machine learning promises to bring greater discipline to decision-making because it offers to uncover factors that are relevant to decision-making that humans might overlook, given the complexity or subtlety of the relationships in historical evidence. Rather than starting with some intuition about the relationship between certain

factors and an outcome of interest, machine learning lets us defer the question of relevance to the data themselves: which factors—among all that we have observed—bear a statistical relationship to the outcome.

Uncovering patterns in historical evidence can be even more powerful than this might seem to suggest. Recent breakthroughs in computer vision—specifically object recognition—reveal just how much pattern-discovery can achieve. In this domain, machine learning has helped to overcome a strange fact of human cognition: while we may be able to effortlessly identify objects in a scene, we are unable to specify the full set of rules that we rely upon to make these determinations. We cannot hand code a program that exhaustively enumerates all the relevant factors that allow us to recognize objects from every possible perspective or in all their potential visual configurations. Machine learning aims to solve this problem by abandoning the attempt to teach a computer through explicit instruction in favor of a process of learning by example. By exposing the computer to many examples of images containing pre-identified objects, we hope the computer will learn the patterns that reliably distinguish different objects from one another and from the environments in which they appear.

This can feel like a remarkable achievement, not only because computers can now execute complex tasks but also because the rules for deciding what appears in an image seem to emerge from the data themselves.

But there are serious risks in learning from examples. Learning is not a process of simply committing examples to memory. Instead, it involves generalizing from examples: honing in on those details that are characteristic of (say) cats in general, not just the specific cats that happen to appear in the examples. This is the process of induction: drawing general rules from specific examples—rules that effectively account for past cases, but also apply to future, as yet unseen cases, too. The hope is that we'll figure out how future cases are likely to be similar to past cases, even if they are not exactly the same.

This means that reliably generalizing from historical examples to future cases requires that we provide the computer with *good* examples: a sufficiently large number of examples to uncover subtle patterns; a sufficiently diverse set of examples to showcase the many different types of appearances that objects might take; a sufficiently well-annotated set of examples to furnish machine learning with reliable ground truth; and so on. Thus, evidence-based decision-making is only as reliable as the evidence on which it is based, and high quality examples are critically important to machine learning. The fact that machine learning is "evidence-based" by no means

ensures that it will lead to accurate, reliable, or fair decisions.

This is especially true when using machine learning to model human behavior and characteristics. Our historical examples of the relevant outcomes will almost always reflect historical prejudices against certain social groups, prevailing cultural stereotypes, and existing demographic inequalities. And finding patterns in these data will often mean replicating these very same dynamics.

We write this book as machine learning begins to play a role in especially consequential decision-making. In the criminal justice system, defendants are assigned statistical scores that are intended to predict the risk of committing future crimes, and these scores inform decisions about bail, sentencing, and parole. In the commercial sphere, firms use machine learning to analyze and filter resumes of job applicants. And statistical methods are of course the bread and butter of lending, credit, and insurance underwriting.

At the same time, machine learning powers everyday applications that might seem frivolous in comparison but collectively have a powerful effect on shaping our culture: search engines, news recommenders, and ad targeting algorithms influence our information diet and our worldviews; chatbots and social recommendation engines mediate our interactions with the world.

This book is an attempt to survey the risks in these and many other applications of machine learning, and to provide a critical review of an emerging set of proposed solutions. It will show how even well-intentioned applications of machine learning might give rise to objectionable results. And it will introduce formal methods for characterizing these problems and assess various computational methods for addressing them.

Demographic disparities

Amazon uses a data-driven system to determine the neighborhoods in which to offer free same-day delivery.[4] A 2016 study found stark disparities in the demographic makeup of these neighborhoods: in many U.S. cities, white residents were more than twice as likely as Black residents to live in one of the qualifying neighborhoods.[5]

In Chapter 2 we'll see how to make our intuition about demographic disparities mathematically precise, and we'll see that there are many possible ways of measuring these inequalities. The pervasiveness of such disparities in machine learning applications is a key concern of this book.

When we observe disparities, it doesn't imply that the designer of the system intended for such inequalities to arise. Looking beyond intent, it's important to understand when observed disparities can

[4] We don't know the details of how Amazon's system works, and in particular we don't know to what extent it uses machine learning. The same is true of many other systems reported on in the press. Nonetheless, we'll use these as motivating examples when a machine learning system for the task at hand would plausibly show the same behavior.

[5] Ingold and Soper, "Amazon Doesn't Consider the Race of Its Customers. Should It?" (https://www.bloomberg.com/graphics/2016-amazon-same-day/, 2016).

be considered to be discrimination. In turn, two key questions to ask are whether the disparities are justified and whether they are harmful. These questions rarely have simple answers, but the extensive literature on discrimination in philosophy and sociology can help us reason about them.

To understand why the racial disparities in Amazon's system might be harmful, we must keep in mind the history of racial prejudice in the United States, its relationship to geographic segregation and disparities, and the perpetuation of those inequalities over time. Amazon argued that its system was justified because it was designed based on efficiency and cost considerations and that race wasn't an explicit factor. Nonetheless, it has the effect of providing different opportunities to consumers at racially disparate rates. The concern is that this might contribute to the perpetuation of long-lasting cycles of inequality. If, instead, the system had been found to be partial to ZIP codes ending in an odd digit, it would not have triggered a similar outcry.

The term *bias* is often used to refer to demographic disparities in algorithmic systems that are objectionable for societal reasons. We'll avoid using this sense of the word bias in this book, since it means different things to different people. There's a more traditional use of the term bias in statistics and machine learning. Suppose that Amazon's estimates of delivery dates/times were consistently too early by a few hours. This would be a case of *statistical bias*. A statistical estimator is said to be biased if its expected or average value differs from the true value that it aims to estimate. Statistical bias is a fundamental concept in statistics, and there is a rich set of established techniques for analyzing and avoiding it.

There are many other measures that quantify desirable statistical properties of a predictor or an estimator, such as precision, recall, and calibration. These are similarly well understood; none of them require any knowledge of social groups and are relatively straightforward to measure. The attention to demographic criteria in statistics and machine learning is a relatively new direction. This reflects a change in how we conceptualize machine learning systems and the responsibilities of those building them. Is our goal to faithfully reflect the data? Or do we have an obligation to question the data, and to design our systems to conform to some notion of equitable behavior, regardless of whether or not that's supported by the data currently available to us? These perspectives are often in tension, and the difference between them will become clearer when we delve into stages of machine learning.

The machine learning loop

Let's study the pipeline of machine learning and understand how demographic disparities propagate through it. This approach lets us glimpse into the black box of machine learning and will prepare us for the more detailed analyses in later chapters. Studying the stages of machine learning is crucial if we want to intervene to minimize disparities.

The figure below shows the stages of a typical system that produces outputs using machine learning. Like any such diagram, it is a simplification, but it is useful for our purposes.

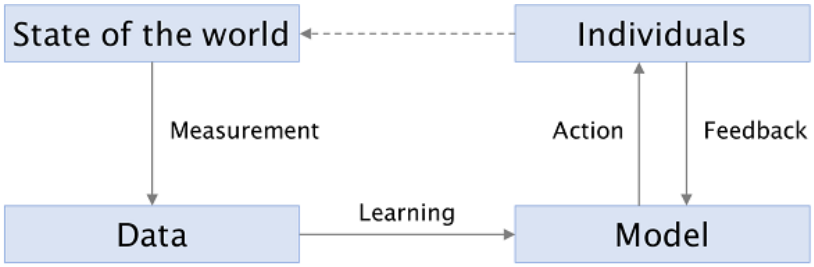


Figure 1: The machine learning loop

The first stage is measurement, which is the process by which the state of the world is reduced to a set of rows, columns, and values in a dataset. It's a messy process, because the real world is messy. The term measurement is misleading, evoking an image of a dispassionate scientist recording what she observes, whereas we'll see that it requires subjective human decisions.

The 'learning' in machine learning refers to the next stage, which is to turn that data into a model. A model summarizes the patterns in the training data; it makes generalizations. A model could be trained using supervised learning via an algorithm such as Support Vector Machines, or using unsupervised learning via an algorithm such as k-means clustering. It could take many forms: a hyperplane or a set of regions in n-dimensional space, or a set of distributions. It is typically represented as a set of weights or parameters.

The next stage is the action we take based on the model's *predictions*, which are applications of the model to new, unseen inputs. 'Prediction' is another misleading term—while it does sometimes involve trying to predict the future ("is this patient at high risk for cancer?"), usually it doesn't. It can take the form of classification (determine whether a piece of email is spam), regression (assigning risk scores to defendants), or information retrieval (finding documents that best match a search query).

The corresponding actions in these three applications might be:

depositing the email in the user's inbox or spam folder, deciding whether to set bail for the defendant's pre-trial release, and displaying the retrieved search results to the user. They may differ greatly in their significance to the individual, but they have in common that the collective responses of individuals to these decisions alter the state of the world—that is, the underlying patterns that the system aims to model.

Some machine learning systems record feedback from users (how users react to actions) and use them to refine the model. For example, search engines track what users click on as an implicit signal of relevance or quality. Feedback can also occur unintentionally, or even adversarially; these are more problematic, as we'll explore later in this chapter.

The state of society

In this book, we're concerned with applications of machine learning that involve data about *people*. In these applications, the available training data will likely encode the demographic disparities that exist in our society. For example, the figure shows the gender breakdown of a sample of occupations in the United States, based on data released by the Bureau of Labor Statistics for the year 2017.[6]

Unsurprisingly, many occupations have stark gender imbalances. If we're building a machine learning system that screens job candidates, we should be keenly aware that this is the baseline we're starting from. It doesn't necessarily mean that the outputs of our system will be inaccurate or discriminatory, but throughout this chapter we'll see how it complicates things.

Why do these disparities exist? There are many potentially contributing factors, including a history of explicit discrimination, implicit attitudes and stereotypes about gender, and differences in the distribution of certain characteristics by gender. We'll see that even in the absence of explicit discrimination, stereotypes can be self-fulfilling and persist for a long time in society. As we integrate machine learning into decision-making, we should be careful to ensure that ML doesn't become a part of this feedback loop.

What about applications that aren't about people? Consider "Street Bump," a project by the city of Boston to crowdsource data on potholes. The smartphone app automatically detects pot holes using data from the smartphone's sensors and sends the data to the city. Infrastructure seems like a comfortably boring application of data-driven decision-making, far removed from the ethical quandaries we've been discussing.

And yet! Kate Crawford points out that the data reflect the pat-

[6] The percentage of women in a sample of occupations in the United States. The area of the bubble represents the number of workers.

terns of smartphone ownership, which are higher in wealthier parts of the city compared to lower-income areas and areas with large elderly populations.[7]

The lesson here is that it's rare for machine learning applications to not be about people. In the case of Street Bump, the data is collected by people, and hence reflects demographic disparities; besides, the reason we're interested in improving infrastructure in the first place is its effect on people's lives.

To drive home the point that most machine learning applications involve people, we analyzed Kaggle, a well-known platform for data science competitions. We focused on the top 30 competitions sorted by prize amount. In 14 of these competitions, we observed that the task is to make decisions about individuals. In most of these cases, there exist societal stereotypes or disparities that may be perpetuated by the application of machine learning. For example, the Automated Essay Scoring[8] task seeks algorithms that attempt to match the scores of human graders of student essays. Students' linguistic choices are signifiers of social group membership, and human graders are known to sometimes have prejudices based on such factors.[9] Thus, because human graders must provide the original labels, automated grading systems risk enshrining any such biases that are captured in the training data.

In a further 5 of the 30 competitions, the task did not call for making decisions about people, but decisions made using the model would nevertheless directly impact people. For example, one competition sponsored by real-estate company Zillow calls for improving the company's "Zestimate" algorithm for predicting home sale prices. Any system that predicts a home's future sale price (and publicizes these predictions) is likely to create a self-fulfilling feedback loop in which homes predicted to have lower sale prices deter future buyers, suppressing demand and lowering the final sale price.

In 9 of the 30 competitions, we did not find an obvious, direct impact on people, such as a competition on predicting ocean health (of course, even such competitions have indirect impacts on people, due to actions that we might take on the basis of the knowledge gained). In two cases, we didn't have enough information to make a determination.

To summarize, human society is full of demographic disparities, and training data will likely reflect these. We'll now turn to the process by which training data is constructed, and see that things are even trickier.

[7] Crawford, "The Hidden Biases in Big Data," *Harvard Business Review* 1 (2013).

[8] Kaggle, "The Hewlett Foundation: Automated Essay Scoring" (https://www.kaggle.com/c/asap-aes, 2012).

[9] Hanna and Linden, "Discrimination in Grading," *American Economic Journal: Economic Policy* 4, no. 4 (2012): 146–68; Sprietsma, "Discrimination in Grading: Experimental Evidence from Primary School Teachers," *Empirical Economics* 45, no. 1 (2013): 523–38.

The trouble with measurement

The term measurement suggests a straightforward process, calling to mind a camera objectively recording a scene. In fact, measurement is fraught with subjective decisions and technical difficulties.

Consider a seemingly straightforward task: measuring the demographic diversity of college campuses. A recent New York Times article aimed to do just this, and was titled "Even With Affirmative Action, Blacks and Hispanics Are More Underrepresented at Top Colleges Than 35 Years Ago."[10] The authors argue that the gap between enrolled Black and Hispanic freshmen and the Black and Hispanic college-age population has grown over the past 35 years. To support their claim, they present demographic information for more than 100 American universities and colleges from the year 1980 to 2015, and show how the percentages of Black, Hispanic, Asian, White, and multiracial students have changed over the years. Interestingly, the multiracial category was only recently introduced in 2008, but the comparisons in the article ignore the introduction of this new category. How many students who might have checked the "White" or "Black" box checked the "multiracial" box instead? How might this have affected the percentages of "White" and "Black" students at these universities? Furthermore, individuals' and society's conception of race changes over time. Would a person with Black and Latino parents be more inclined to self-identify as Black in 2015 than in the 1980s? The point is that even a seemingly straightforward question about trends in demographic diversity is impossible to answer without making some assumptions, and illustrates the difficulties of measurement in a world that resists falling neatly into a set of checkboxes. Race is not a stable category; how we measure race often changes how we conceive of it, and changing conceptions of race may force us to alter what we measure.

To be clear, this situation is typical: measuring almost any attribute about people is similarly subjective and challenging. If anything, things are more chaotic when machine learning researchers have to create categories, as is often the case.

One area where machine learning practitioners often have to define new categories is in defining the target variable.[11] This is the outcome that we're trying to predict – will the defendant recidivate if released on bail? Will the candidate be a good employee if hired? And so on.

Biases in the training set's target variable are especially critical, because they are guaranteed to bias the predictions (not necessarily so with other attributes). But the target variable is arguably the hardest from a measurement standpoint, because it is often a construct that

[10] Ashkenas, Park, and Pearce, "Even with Affirmative Action, Blacks and Hispanics Are More Underrepresented at Top Colleges Than 35 Years Ago" (https://www.nytimes.com/interactive/2017/08/24/us/affirmative-action.html, 2017).

[11] Barocas and Selbst, "Big Data's Disparate Impact," *UCLA Law Review*, 2016.

is made up for the purposes of the problem at hand rather than one that is widely understood and measured. For example, "creditworthiness" is a construct that was created in the context of the problem of how to successfully extend credit to consumers;[12] it is not an intrinsic property that people either possess or lack.

If our target variable is the idea of a "good employee," we might use performance review scores to quantify it. This means that our data inherits any biases present in managers' evaluations of their reports. Another example: the use of computer vision to automatically rank people's physical attractiveness.[13] The training data consists of human evaluation of attractiveness, and, unsurprisingly, all these classifiers showed a preference for lighter skin.

In some cases we might be able to get closer to a more objective definition for a target variable, at least in principle. For example, in criminal risk assessment, the training data is not judges' decisions on who should get bail, but rather based on who actually went on to commit a crime. But there's at least one big caveat—we can't really measure who committed a crime, so we use arrests as a proxy. This replaces the biases of judges with the biases of policing. On the other hand, if our target variable is whether the defendant appears or fails to appear in court for trial, we would be able to measure it directly with perfect accuracy. That said, we may still have concerns about a system that treats defendants differently based on predicted probability of appearance, given that some reasons for failing to appear are less objectionable than others (trying to hold down a job that would not allow for time off versus trying to avoid prosecution).

In hiring, instead of relying on performance reviews for (say) a sales job, we might rely on the number of sales closed. But is that an objective measurement or is it subject to the biases of the potential customers (who might respond more positively to certain salespeople than others) and workplace conditions (which might be a hostile environment for some, but not others)?

In some applications, researchers repurpose an existing scheme of classification to define the target variable rather than creating one from scratch. For example, an object recognition system can be created by training a classifier on ImageNet, a database of images organized in a hierarchy of concepts.[14] ImageNet's hierarchy comes from Wordnet, a database of words, categories, and the relationships among them.[15] Wordnet's authors in turn imported the word lists from a number of older sources, such as thesauri. As a result, WordNet (and ImageNet) categories contain numerous outmoded words and associations, such as occupations that no longer exist and stereotyped gender associations. Thus, ImageNet-trained object recognition systems assume a categorization of the world that is mismatched

[12] Barocas and Selbst.

[13] Plaugic, "FaceApp's Creator Apologizes for the App's Skin-Lightening 'Hot' Filter" (The Verge. https://www.theverge.com/2017/4/25/15419522/faceapp-hot-filter-racist-apology, 2017); Manthorpe, "The Beauty.AI Robot Beauty Contest Is Back" (Wired UK. https://www.wired.co.uk/article/robot-beauty-contest-beauty-ai, 2017).

[14] Deng et al., "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. CVPR*, 2009.

[15] Miller, "WordNet: A Lexical Database for English," *Communications of the ACM* 38, no. 11 (1995): 39–41.

with the world in which they operate.

We think of technology changing rapidly and society being slow to adapt, but at least in this instance, the categorization scheme at the heart of much of today's machine learning technology has been frozen in time while social norms have changed rapidly.

Our favorite example of measurement bias has to do with cameras, which we referenced at the beginning of the section as the exemplar of dispassionate observation and recording. But are they?

The visual world has an essentially infinite bandwidth compared to what can be captured by cameras, whether film or digital, which means that photography technology involves a series of choices about what is relevant and what isn't, and transformations of the captured data based on those choices. Both film and digital cameras have historically been more adept at photographing lighter-skinned individuals.[16] One reason is the default settings such as color balance which were optimized for lighter skin tones. Another, deeper reason is the limited "dynamic range" of cameras, which makes it hard to capture brighter and darker tones in the same image. This started changing in the 1970s, in part due to complaints from furniture companies and chocolate companies about the difficulty of photographically capturing the details of furniture and chocolate respectively! Another impetus came from the increasing diversity of television subjects at this time.

When we go from individual images to datasets of images, we introduce another layer of potential biases. Consider the image datasets that are used to train today's computer vision systems for tasks such as object recognition. If these datasets were representative samples of an underlying visual world, we might expect that a computer vision system trained on one such dataset would do well on another dataset. But in reality, we observe a big drop in accuracy when we train and test on different datasets.[17] This shows that these datasets are biased relative to each other in a statistical sense, and is a good starting point for investigating whether these biases include cultural stereotypes.

It's not all bad news: machine learning can in fact help mitigate measurement biases. Returning to the issue of dynamic range in cameras, computational techniques, including machine learning, are making it possible to improve the representation of tones in images.[18] Another example comes from medicine: diagnoses and treatments are sometimes personalized by race. But it turns out that race is used as a crude proxy for ancestry and genetics, and sometimes environmental and behavioral factors.[19] If we can measure these genetic and lifestyle factors and incorporate them—instead of race—into statistical models of disease and drug response, we can increase

[16] Roth, "Looking at Shirley, the Ultimate Norm: Colour Balance, Image Technologies, and Cognitive Equity," *Canadian Journal of Communication* 34, no. 1 (2009): 111.

[17] Torralba and Efros, "Unbiased Look at Dataset Bias," in *Proc. CVPR* (IEEE, 2011), 1521–28.

[18] Liu, Zhang, and Zhang, "Learning-Based Perceptual Image Quality Improvement for Video Conferencing," in *Multimedia and Expo, 2007 IEEE International Conference on* (IEEE, 2007), 1035–38; Kaufman, Lischinski, and Werman, "Content-Aware Automatic Photo Enhancement," in *Computer Graphics Forum*, vol. 31, 8 (Wiley Online Library, 2012), 2528–40; Kalantari and Ramamoorthi, "Deep High Dynamic Range Imaging of Dynamic Scenes," *ACM Trans. Graph* 36, no. 4 (2017): 144.

[19] Bonham, Callier, and Royal, "Will Precision Medicine Move Us Beyond Race?" *The New England Journal of Medicine* 374, no. 21 (2016): 2003; Wilson et al., "Population Genetic Structure of Variable Drug Response," *Nature Genetics* 29, no. 3 (2001): 265.

the accuracy of diagnoses and treatments while mitigating racial biases.

To summarize, measurement involves defining your variables of interest, the process for interacting with the real world and turning your observations into numbers, and then actually collecting the data. Usually machine learning practitioners don't think about these steps, because someone else has already done those things. And yet it is crucial to understand the provenance of the data. Even if someone else has collected the data for you, it's almost always too messy for your algorithms to handle, hence the dreaded "data cleaning" step. But the messiness of the real world isn't just an annoyance to be dealt with by cleaning, it is instead a manifestation of the limitations of data-driven techniques.

From data to models

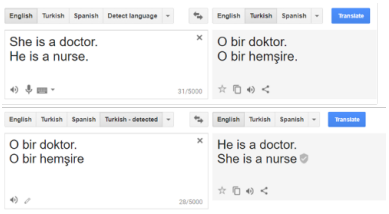
We've seen that training data reflects the disparities, distortions, and biases from the real world and the measurement process. This leads to an obvious question: when we learn a model from such data, are these disparities preserved, mitigated, or exacerbated?

Predictive models trained with supervised learning methods are often good at calibration: ensuring that the model's prediction subsumes all features in the data for the purpose of predicting the outcome. By contrast, human intuition is notoriously poor at accounting for priors, and this is a major reason that statistical predictions perform better in a wide variety of settings. But calibration also means that by default, we should expect our models to faithfully reflect disparities found in the input data.

Here's another way to think about it. Some patterns in the training data (smoking is associated with cancer) represent knowledge that we wish to mine using machine learning, while other patterns (girls like pink and boys like blue) represent stereotypes that we might wish to avoid learning. But learning algorithms have no general way to distinguish between these two types of patterns, because they are the result of social norms and moral judgments. Absent specific intervention, machine learning will extract stereotypes, including incorrect and harmful ones, in the same way that it extracts knowledge.

A telling example of this comes from machine translation. The screenshot on the right shows the result of translating sentences from English to Turkish and back.[20] The same stereotyped translations result for many pairs of languages and other occupation words in all translation engines we've tested. It's easy to see why. Turkish has gender neutral pronouns, and when translating such a pronoun to

[20] Translating from English to Turkish, then back to English injects gender stereotypes.**



English, the system picks the sentence that best matches the statistics of the training set (which is typically a large, minimally curated corpus of historical text and text found on the web).

When we build a statistical model of language from such text, we should expect the gender associations of occupation words to roughly mirror real-world labor statistics. In addition, because of the male-as-norm bias[21] (the use of male pronouns when the gender is unknown) we should expect translations to favor male pronouns. It turns out that when we repeat the experiment with dozens of occupation words, these two factors—labor statistics and the male-as-norm bias—together almost perfectly predict which pronoun will be returned.[22]

[21] Danesi, *Dictionary of Media and Communications* (Routledge, 2014).

[22] Caliskan, Bryson, and Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases," *Science* 356, no. 6334 (2017): 183–86.

Here's a tempting response to the observation that models reflect data biases. Suppose we're building a model for scoring resumes for a programming job. What if we simply withhold gender from the data? Surely the resulting model can't be gender biased? Unfortunately, it's not that simple, because of the problem of proxies[23] or redundant encodings,[24] as we'll discuss in the next chapter. There are any number of other attributes in the data that might correlate with gender. In our culture, the age at which someone starts programming is well known to be correlated with gender. This illustrates another problem with proxies: they may be genuinely relevant to the decision at hand. How long someone has been programming is a factor that gives us valuable information about their suitability for a programming job, but it also reflects the reality of gender stereotyping.

[23] Barocas and Selbst, "Big Data's Disparate Impact."

[24] Hardt, "How Big Data Is Unfair" (https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de, 2014).

Finally, it's also possible for the learning step to introduce demographic disparities that aren't in the training data. The most common reason for this is the sample size disparity. If we construct our training set by sampling uniformly from the training data, then by definition we'll have fewer data points about minorities. Of course, machine learning works better when there's more data, so it will work less well for members of minority groups, assuming that members of the majority and minority groups are systematically different in terms of the prediction task.[25]

[25] Hardt.

Worse, in many settings minority groups are underrepresented relative to population statistics. For example, minority groups are underrepresented in the tech industry. Different groups might also adopt technology at different rates, which might skew datasets assembled form social media. If training sets are drawn from these unrepresentative contexts, there will be even fewer training points from minority individuals. For example, many products that incorporate face-detection technology have been reported to have trouble with non-Caucasian faces, and it's easy to guess why.[26]

[26] Hardt.

When we develop machine-learning models, we typically only

test their overall accuracy; so a "5% error" statistic might hide the fact that a model performs terribly for a minority group. Reporting accuracy rates by group will help alert us to problems like the above example. In the next chapter, we'll look at metrics that quantify the error-rate disparity between groups.

There's one application of machine learning where we find especially high error rates for minority groups: anomaly detection. This is the idea of detecting behavior that deviates from the norm as evidence of abuse against a system. A good example is the *Nymwars* controversy, where Google, Facebook, and other tech companies aimed to block users who used uncommon (hence, presumably fake) names.

Further, suppose that in some cultures, most people receive names from a small set of names, whereas in other cultures, names might be more diverse, and it might be common for names to be unique. For users in the latter culture, a popular name would be more likely to be fake. In other words, the same feature that constitutes evidence towards a prediction in one group might constitute evidence against the prediction for another group.[27]

[27] Hardt.

If we're not careful, learning algorithms will generalize based on the majority culture, leading to a high error rate for minority groups. This is because of the desire to avoid overfitting, that is, picking up patterns that arise due to random noise rather than true differences. One way to avoid this is to explicitly model the differences between groups, although there are both technical and ethical challenges associated with this, as we'll show in later chapters.

The pitfalls of action

Any real machine-learning system seeks to make some change in the world. To understand its effects, then, we have to consider it in the context of the larger socio-technical system in which it is embedded.

In Chapter 2, we'll see that if a model is calibrated—it faithfully captures the patterns in the underlying data—predictions made using that model will inevitably have disparate error rates for different groups, if those groups have different *base rates*, that is, rates of positive or negative outcomes. In other words, understanding the properties of a prediction requires understanding not just the model, but also the population differences between the groups on which the predictions are applied.

Further, population characteristics can shift over time; this is a well-known machine learning phenomenon known as drift. If sub-populations change differently over time, that can introduce disparities. An additional wrinkle: whether or not disparities are

objectionable may differ between cultures, and may change over time as social norms evolve.

When people are subject to automated decisions, their perception of those decisions depends not only on the outcomes but also the process of decision-making. An ethical decision-making process might require, among other things, the ability to explain a prediction or decision, which might not be feasible with black-box models.

A major limitation of machine learning is that it only reveals correlations, but we often use its predictions as if they reveal causation. This is a persistent source of problems. For example, an early machine learning system in healthcare famously learned the seemingly nonsensical rule that patients with asthma had lower risk of developing pneumonia. This was a true pattern in the data, but the likely reason was that asthmatic patients were more likely to receive in-patient care.[28] So it's not valid to use the prediction to decide whether or not to admit a patient. We'll discuss causality in Chapter 4.

[28] Caruana et al., "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission," in *Proc. 21St ACM SIGKDD*, 2015, 1721–30.

Another way to view this example is that the prediction affects the outcome (because of the actions taken on the basis of the prediction), and thus invalidates itself. The same principle is also seen in the use of machine learning for predicting traffic congestion: if sufficiently many people choose their routes based on the prediction, then the route predicted to be clear will in fact be congested. The effect can also work in the opposite direction: the prediction might reinforce the outcome, resulting in feedback loops. To better understand how, let's talk about the final stage in our loop: feedback.

Feedback and feedback loops

Many systems receive feedback when they make predictions. When a search engine serves results, it typically records the links that the user clicks on and how long the user spends on those pages, and treats these as implicit signals about which results were found to be most relevant. When a video sharing website recommends a video, it uses the thumbs up/down feedback as an explicit signal. Such feedback is used to refine the model.

But feedback is tricky to interpret correctly. If a user clicked on the first link on a page of search results, is that simply because it was first, or because it was in fact the most relevant? This is again a case of the action (the ordering of search results) affecting the outcome (the link(s) the user clicks on). This is an active area of research; there are techniques that aim to learn accurately from this kind of biased feedback.[29]

[29] Joachims, Swaminathan, and Schnabel, "Unbiased Learning-to-Rank with Biased Feedback," in *Proc. 10Th International Conference on Web Search and Data Mining* (ACM, 2017), 781–89.

Bias in feedback might also reflect cultural prejudices, which is

of course much harder to characterize than the effects of the ordering of search results. For example, the clicks on the targeted ads that appear alongside search results might reflect gender and racial stereotypes. There's a well-known study that hints at this: Google searches for Black-sounding names such as "Latanya Farrell" were much more likely to results in ads for arrest records ("Latanya Farrell, Arrested?") than searches for white-sounding names ("Kristen Haring").[30] One potential explanation is that users are more likely to click on ads that conform to stereotypes, and the advertising system is optimized for maximizing clicks.

[30] Sweeney, "Discrimination in Online Ad Delivery," *Queue* 11, no. 3 (March 2013): 10:10–29.

In other words, even feedback that's designed into systems can lead to unexpected or undesirable biases. But there are many unintended ways in which feedback might arise, and these are more pernicious and harder to control. Let's look at three.

Self-fulfilling predictions. Suppose a predictive policing system determines certain areas of a city to be at high risk for crime. More police officers might be deployed to such areas. Alternatively, officers in areas predicted to be high risk might be subtly lowering their threshold for stopping, searching, or arresting people—perhaps even unconsciously. Either way, the prediction will appear to be validated, even if it had been made purely based on data biases.

Here's another example of how acting on a prediction can change the outcome. In the United States, some criminal defendants are released prior to trial, whereas for others, a bail amount is set as a precondition of release. Many defendants are unable to post bail. Does the release or detention affect the outcome of the case? Perhaps defendants who are detained face greater pressure to plead guilty. At any rate, how could one possibly test the causal impact of detention without doing an experiment? Intriguingly, we can take advantage of a pseudo-experiment, namely that defendants are assigned bail judges quasi-randomly, and some judges are stricter than others. Thus, pre-trial detention is partially random, in a quantifiable way. Studies using this technique have confirmed that detention indeed causes an increase in the likelihood of a conviction.[31] If bail were set based on risk predictions, whether human or algorithmic, and we evaluated its efficacy by examining case outcomes, we would see a self-fulfilling effect.

[31] Dobbie, Goldin, and Yang, "The Effects of Pre-Trial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges" (National Bureau of Economic Research, 2016).

Predictions that affect the training set. Continuing this example, predictive policing activity will leads to arrests, records of which might be added to the algorithm's training set. These areas might then continue to appear to be at high risk of crime, and perhaps also other areas with a similar demographic composition, depending on the feature set used for predictions. The biases might even compound over time.

A 2016 paper analyzed a predictive policing algorithm by PredPol, one of the few to be published in a peer-reviewed journal.[32] By applying it to data derived from Oakland police records, they found that Black people would be targeted for predictive policing of drug crimes at roughly twice the rate of whites, even though the two groups have roughly equal rates of drug use.[33] Their simulation showed that this initial bias would be amplified by a feedback loop, with policing increasingly concentrated on targeted areas. This is despite the fact that the PredPol algorithm does not explicitly take demographics into account.

[32] PredPol deserves praise for publicly releasing their algorithm, without which this research would not even have been possible.

[33] Lum and Isaac, "To Predict and Serve?" *Significance* 13, no. 5 (2016): 14–19.

A more recent paper built on this idea and showed mathematically how feedback loops occur when data discovered on the basis of predictions are used to update the model.[34] The paper also shows how to tweak the model to avoid feedback loops: by quantifying how surprising an observation of crime is given the predictions, and only updating the model in response to surprising events.

[34] Ensign et al., "Runaway Feedback Loops in Predictive Policing," *arXiv Preprint arXiv:1706.09847*, 2017.

Predictions that affect the phenomenon and society at large. Prejudicial policing on a large scale, algorithmic or not, will affect society over time, contributing to the cycle of poverty and crime. This is an extremely well-trodden thesis, and we'll briefly review the sociological literature on durable inequality and the persistence of stereotypes in Chapter 3.

Let us remind ourselves that we deploy machine learning so that we can act on its predictions. It is hard to even conceptually eliminate the effects of predictions on outcomes, future training sets, the phenomena themselves, or society at large. The more central machine learning becomes in our lives, the stronger this effect.

Returning to the example of a search engine, in the short term it might be possible to extract an unbiased signal from user clicks, but in the long run, results that are returned more often will be linked to and thus rank more highly. As a side effect of fulfilling its purpose of retrieving relevant information, a search engine will necessarily change the very thing that it aims to measure, sort, and rank. Similarly, most machine learning systems will affect the phenomena that they predict. This is why we've depicted the machine learning process as a loop.

Throughout this book we'll learn methods for mitigating societal biases in machine learning, but let us pause to consider that there are fundamental limits to what we can achieve, especially when we consider machine learning as a socio-technical system instead of a mathematical abstraction. The textbook model of training and test data being independent and identically distributed is a simplification, and might be unachievable in practice.

Getting concrete with a toy example

Now let's look at a concrete setting, albeit a toy problem, to illustrate many of the ideas discussed so far, and some new ones.

Let's say you're on a hiring committee, making decisions based on just two attributes of each applicant: their college GPA and their interview score (we did say it's a toy problem!). We formulate this as a machine-learning problem: the task is to use these two variables to predict some measure of the "quality" of an applicant. For example, it could be based on the average performance review score after two years at the company. We'll assume we have data from past candidates that allows us to train a model to predict performance scores based on GPA and interview score.

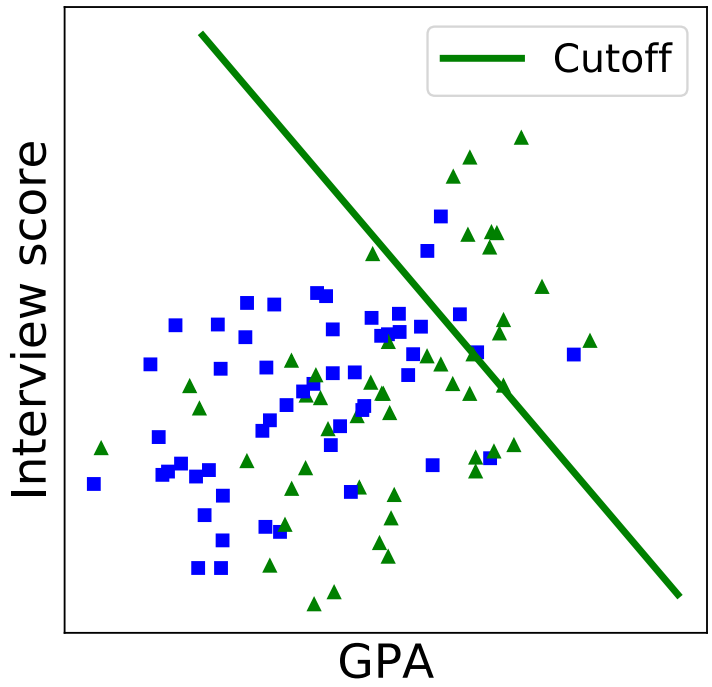


Figure 2: Toy example: a hiring classifier that predicts job performance (not shown) based on GPA and interview score, and then applies a cutoff.

Obviously, this is a reductive formulation—we're assuming that an applicant's worth can be reduced to a single number, and that we know how to measure that number. This is a valid criticism, and applies to most applications of data-driven decision-making today. But it has one big advantage: once we do formulate the decision as a prediction problem, statistical methods tend to do better than

humans, even domain experts with years of training, in making decisions based on noisy predictors. The subject has been well researched, and we'll study it in Chapter 3.

Given this formulation, the simplest thing we can do is to use linear regression to predict the average job performance rating from the two observed variables, and then use a cutoff based on the number of candidates we want to hire. The figure above shows what this might look like. In reality, the variables under consideration need not satisfy a linear relationship, thus suggesting the use of a non-linear model, which we avoid for simplicity.

As you can see in the figure, our candidates fall into two demographic groups, represented by triangles and squares.[35] Note that the classifier didn't take into account which group a candidate belonged to. Does this mean that the classifier is fair? We might hope that it is, based on the fairness-as-blindness idea, symbolized by the icon of Lady Justice wearing a blindfold. In this view, an impartial model—one that doesn't use the group membership in the regression—is fair; a model that gives different scores to otherwise-identical members of different groups is discriminatory.

We'll defer a richer understanding of what fairness means to Chapter 3, so let's ask a simpler question: are candidates from the two groups equally likely to be positively classified? The answer is no: the triangles are more likely to be selected than the squares. That's because data is a social mirror; the "ground truth" labels that we're predicting—job performance ratings—are systematically lower for the squares than the triangles.

There are many possible reasons for this disparity. First, the managers who score the employees' performance might have a bias against one group. Or the overall workplace might be biased against one group, preventing them from reaching their potential and leading to lower performance. Alternately, the disparity might originate before the candidates were hired. For example, it might arise from disparities in educational institutions attended by the two groups. Or there might be intrinsic differences between them. Of course, it might be a combination of these factors. We can't tell from our data how much of the disparity is attributable to these different factors. In general, such a determination is methodologically hard, and requires causal reasoning.[36]

For now, let's assume that we have evidence that the level of demographic disparity produced by our selection procedure is unjustified, and we're interested in intervening to decrease it. How could we do it? We observe that GPA is correlated with the demographic attribute—it's a proxy. Perhaps we could simply omit that variable as a predictor? Unfortunately, we'd also cripple the accuracy of our

[35] This binary categorization is a simplification for the purposes of our thought experiment. Such simplifications are also common in the research literature. Indeed, most proposed fairness interventions themselves start by assuming such a categorization. But when building real systems, enforcing rigid categories of people can be ethically questionable. This is not specific to machine learning, and a similar tension arises in many data-driven settings, such as the checkboxes for race on census forms or employment applications.

[36] Zhang and Bareinboim, "Fairness in Decision-Making — the Causal Explanation Formula," in *Proc. 32Nd AAAI*, 2018.

model. In real datasets, most attributes tend to be proxies for demographic variables, and dropping them may not be a reasonable option.

Another crude approach is to pick different cutoffs so that candidates from both groups have the same probability of being hired. Or we could mitigate the demographic disparity instead of eliminating it, by decreasing the difference in the cutoffs.

Given the available data, there is no mathematically principled way to know which cutoffs to pick. In some situations there is a legal baseline: for example, guidelines from the U.S. Equal Employment Opportunity Commission state that if the probability of selection for two groups differs by more than 20%, it might constitute a sufficient disparate impact to initiate a lawsuit. But a disparate impact alone is not illegal; the disparity needs to be unjustified or avoidable for courts to find liability. Even these quantitative guidelines do not provide easy answers or bright lines.

At any rate, the pick-different-thresholds approach to mitigating disparities seems unsatisfying. It is no longer blind, and two candidates with the same observable attributes may receive different decisions depending on which group they are in.

But there are other possible interventions, and we'll discuss one. To motivate it, let's take a step back and ask why the company wants to decrease the demographic disparity in hiring.

One answer is rooted in justice to individuals and the specific social groups to which they belong. But a different answer comes from the firm's selfish interests: diverse teams work better.[37] From this perspective, increasing the diversity of the cohort that is hired would benefit the firm and everyone in the cohort.

[37] Rock and Grant, "Why Diverse Teams Are Smarter" (Harvard Business Review. https://hbr.org/2016/11/why-diverse-teams-are-smarter, 2016).

How do we operationalize diversity in a selection task? If we had a distance function between pairs of candidates, we could measure the average distance between selected candidates. As a strawman, let's say we use the Euclidean distance based on the GPA and interview score. If we incorporated such a diversity criterion into the objective function, it would result in a model where the GPA is weighted less. This technique has the advantage of being blind: we didn't explicitly consider the group membership, but as a side-effect of insisting on diversity of the other observable attributes, we have also improved demographic diversity. However, a careless application of such an intervention can easily go wrong: for example, the model might give weight to attributes that are completely irrelevant to the task.

More generally, there are many possible algorithmic interventions beyond picking different thresholds for different groups. In particular, the idea of a similarity function between pairs of individuals is

a powerful one, and we'll see other interventions that make use of it. But coming up with a suitable similarity function in practice isn't easy: it may not be clear which attributes are relevant, how to weight them, and how to deal with correlations between attributes.

Other ethical considerations

So far we've been mostly concerned with ethical concerns that arise from demographic disparities in the outputs of machine learning systems. But a few other types of concerns are worth highlighting.

Predictions versus interventions

Fairly rendered decisions under unfair circumstances may do little to improve people's lives. In many cases, we cannot achieve any reasonable notion of fairness through changes to decision-making alone; we need to change the conditions under which these decisions are made.

Let's return to the hiring example above. When using machine learning to make predictions about how someone might fare in a specific workplace or occupation, we tend to treat the environment that people will confront in these roles as a constant and ask how people's performance will vary according to their observable characteristics. In other words, we treat the current state of the world as a given, leaving us to select the person who will do best under these circumstances. This approach risks overlooking more fundamental changes that we could make to the workplace (culture, family friendly policies, on-the-job training) that might make it a more welcoming and productive environment for people that have not flourished under previous conditions.[38]

The tendency with work on fairness in machine learning is to ask whether an employer is using a fair selection process, even though we might have the opportunity to intervene in the workplace dynamics that actually account for differences in predicted outcomes along the lines of race, gender, disability, and other characteristics.[39]

We can learn a lot from the so-called social model of disability, which views a predicted difference in a disabled person's ability to excel on the job as the result of a lack of appropriate accommodations (an accessible workplace, necessary equipment, flexible working arrangements) rather than any inherent capacity of the person himself. A person is only disabled in the sense that we have not built physical environments or adopted appropriate policies to ensure their equal participation.

The same might be true of people with other characteristics, and

[38] Barocas, "Putting Data to Work," in *Data and Discrimination: Collected Essays*, ed. Seeta Peña Gangadharan Virginia Eubanks and Solon Barocas (New America Foundation, 2014), 59–62.

[39] Jackson and VanderWeele, "Decomposition Analysis to Identify Intervention Targets for Reducing Disparities," *Epidemiology*, 2018, 825–35.

changes to the selection process alone will not help us address the fundamental injustice of conditions that keep certain people from contributing as effectively as others.

Accuracy

Accuracy is an underappreciated ethical issue. The reason that it doesn't get much attention in the technical literature is that we assume a setting where a decision maker has some notion of utility, which is almost always directly connected to maximizing accuracy. For example, a bank deciding who should receive a loan might use data to predict whether the recipient will pay it back; they would like to minimize both types of errors—false positives and false negatives—as they would lose money with false positives and forego potential profits with false negatives. Thus, machine learning problems are already framed in terms of maximizing accuracy, and the literature often talks about the accuracy-fairness trade-off.

Yet there are two reasons to separately consider accuracy as a criterion for responsible machine learning. We've already discussed one of these: errors might be unequally distributed between demographic groups, and a utility-maximizing decision maker might not take this into account.

The other, related reason is that whether to deploy the automated decision-making system at all is often a debate to be had, and one that we're not comfortable leaving to the logic (and whims) of the marketplace. Two such debates recently: should police use of facial recognition technology be regulated, and now?[40,41] What can go wrong with the use of DNA testing as a forensic tool? Understanding the error rate as well as the nature of errors of these technologies is critical to an informed debate.

At the same time, debating the merits of these technologies on the basis of their likely accuracy for different groups may distract from a more fundamental question: should we ever deploy such systems, even if they perform equally well for everyone? We may want to regulate the police's access to such tools, even if the tools are perfectly accurate. Our civil rights—freedom or movement and association—are equally threatened by these technologies when they fail and when they work well.

[40] Garvie, Bedoya, and Frankle, "The Perpetual Line-up," *Georgetown Law: Center on Privacy and Technology.*, 2016.

[41] This is not to say that accuracy is the sole criterion in determining the acceptability of police use of facial recognition. Rather, the primary concerns are about civil liberties and the unaccountability of police power.

Diversity

Diversity is a bit of a catch-all term. It is a criterion in selection systems, such as in the hiring example above. Another context in which we might care about diversity is in the construction of training datasets for machine learning that are representative of the world.

Let's discuss two more.

In information systems, low diversity can lead to a narrowing of opportunity. For example, one reason that students from poor backgrounds don't go to selective colleges is that they are simply unaware that the opportunity is available to them.[42] Online search and ads are valuable avenues for mitigating this problem; yet, doing so requires swimming against the current of targeting of ads (and sometimes searches) based on algorithmic profiling of users. There is evidence that ad targeting sometimes narrows opportunities in this way.[43]

A related concern arises in personalization systems: the infamous filter bubble.[44] This is the idea that when algorithmic systems learn our past activities to predict what we might click on, they feed us information that conforms to our existing views. Note that individual users may like the filter bubble—indeed, research suggests that our own choices result in a narrowing of what we consume online, compared to algorithmic recommendations—[45]but the worry is that an ideologically segregated populace may not be conducive to a functioning democracy. The filter bubble is a concern for search engines, news websites, and social media; the relevant machine learning techniques include information retrieval and collaborative filtering.

[42] Dillon and Smith, "The Determinants of Mismatch Between Students and Colleges" (National Bureau of Economic Research, 2013); Jaquette and Salazar, "Opinion | Colleges Recruit at Richer, Whiter High Schools - the New York Times" (https://www.nytimes.com/interactive/2018/04/13/opinion/college-recruitment-rich-white.html, 2018).

[43] Datta, Tschantz, and Datta, "Automated Experiments on Ad Privacy Settings," *Proceedings on Privacy Enhancing Technologies* 2015, no. 1 (2015): 92–112.

[44] Pariser, *The Filter Bubble: What the Internet Is Hiding from You* (Penguin UK, 2011).

[45] Bakshy, Messing, and Adamic, "Exposure to Ideologically Diverse News and Opinion on Facebook," *Science* 348, no. 6239 (2015): 1130–32.

Stereotype perpetuation and cultural denigration

Image search results for occupation terms such as CEO or software developer reflect (and arguably exaggerate) the prevailing gender composition and stereotypes about those occupations.[46] Should we care about such disparities in image search results? After all, these results don't affect hiring or any other consequential decisions. And what are the harms from gender stereotypes in online translation? These and other examples that are disturbing to varying degrees—such as Google's app labeling photos of Black Americans as "gorillas," or offensive results in autocomplete—seem to fall into a different moral category than, say, a discriminatory system used in criminal justice, which has immediate and tangible consequences.

A recent talk lays out the differences.[47] When decision-making systems in criminal justice, health care, etc. are discriminatory, they create *allocative harms*, which are caused when a system withholds certain groups an opportunity or a resource. In contrast, the other examples—stereotype perpetuation and cultural denigration—are examples of *representational harms*, which occur when systems reinforce the subordination of some groups along the lines of identity—race, class, gender, etc.

[46] Kay, Matuszek, and Munson, "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (ACM, 2015), 3819–28.

[47] Crawford, "The Trouble with Bias" (NIPS Keynote https://www.youtube.com/watch?v=fMym_BKWQzk, 2017).

Allocative harms have received much attention both because their effects are immediate, and because they are easier to formalize and study in computer science and in economics. Representational harms have long-term effects, and resist formal characterization. But as machine learning becomes a bigger part of how we make sense of the world—through technologies such as search, translation, voice assistants, and image labeling—representational harms will leave an imprint on our culture, and influence identity formation and stereotype perpetuation. Thus, these are critical concerns for the fields of natural language processing and computer vision.

Our outlook: limitations and opportunities

We've seen how machine learning propagates inequalities in the state of the world through the stages of measurement, learning, action, and feedback. Machine learning systems that affect people are best thought of as closed loops, since the actions we take based on predictions in turn affect the state of the world. One major goal of fair machine learning is to develop an understanding of when these disparities are harmful, unjustified, or otherwise unacceptable, and to develop interventions to mitigate such disparities.

There are fundamental challenges and limitations to this goal. Unbiased measurement might be infeasible even in principle, as we've seen through examples. There are additional practical limitations arising from the fact that the decision maker is typically not involved in the measurement stage. Further, observational data can be insufficient to identify the causes of disparities, which is needed in the design of meaningful interventions and in order to understand the effects of intervention. Most attempts to "debias" machine learning in the current research literature assume simplistic mathematical systems, often ignoring the effect of algorithmic interventions on individuals and on the long-term state of society.

Despite these important limitations, there are reasons to be cautiously optimistic about fairness and machine learning. First, data-driven decision-making has the potential to be more transparent compared to human decision-making. It forces us to articulate our decision-making objectives and enables us to clearly understand the tradeoffs between desiderata. However, there are challenges to overcome to achieve this potential for transparency. One challenge is improving the interpretability and explainability of modern machine learning methods, which is a topic of vigorous ongoing research. Another challenge is the proprietary nature of datasets and systems that are crucial to an informed public debate on this topic. Many commentators have called for a change in the status quo.[48]

[48] Reisman et al., "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability" (https://ainowinstitute.org/aiareport2018.pdf, 2018).

Second, effective interventions do exist in many machine learning applications, especially in natural-language processing and computer vision. Tasks in these domains (say, transcribing speech) are subject to less inherent uncertainty than traditional decision-making (say, predicting if a loan applicant will repay), removing some of the statistical constraints that we'll study in Chapter 2.

Our final and most important reason for optimism is that the turn to automated decision-making and machine learning offers an opportunity to reconnect with the moral foundations of fairness. Algorithms force us to be explicit about what we want to achieve with decision-making. And it's far more difficult to paper over our poorly specified or true intentions when we have to state these objectives formally. In this way, machine learning has the potential to help us debate the fairness of different policies and decision-making procedures more effectively.

We should not expect work on fairness in machine learning to deliver easy answers. And we should be suspicious of efforts that treat fairness as something that can be reduced to an algorithmic stamp of approval. At its best, this work will make it far more difficult to avoid the hard questions when it comes to debating and defining fairness, not easier. It may even force us to confront the meaningfulness and enforceability of existing approaches to discrimination in law and policy,[49] expanding the tools at our disposal to reason about fairness and seek out justice.

We hope that this book can play a small role in stimulating this nascent interdisciplinary inquiry.

[49] Barocas and Selbst, "Big Data's Disparate Impact."

Bibliographic notes and further reading

For an introduction to machine learning, we refer the reader to the text by Hardt and Recht[50] that is available online at mlstory.org. An excellent textbook by Wasserman[51] also provides additional statistical background.

This chapter draws from several taxonomies of biases in machine learning and data-driven decision-making: a blog post by Moritz Hardt,[52] a paper by Barocas and Selbst,[53] and a 2016 report by the White House Office of Science and Technology Policy.[54] For a broad survey of challenges raised by AI, machine learning, and algorithmic systems, see the AI Now report.[55]

An early work that investigated fairness in algorithmic systems is by Friedman and Nissenbaum in 1996.[56] Papers studying demographic disparities in classification began appearing regularly starting in 2008;[57] the locus of this research was in Europe, and in the data mining research community. With the establishment of the

[50] Hardt and Recht, *Patterns, Predictions, and Actions: A Story about Machine Learning* (https://mlstory.org, 2021), https://arxiv.org/abs/2102.05242.

[51] Wasserman, *All of Statistics: A Concise Course in Statistical Inference* (Springer, 2010).

[52] Hardt, "How Big Data Is Unfair."

[53] Barocas and Selbst, "Big Data's Disparate Impact."

[54] Munoz, Smith, and Patil, "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights," *Executive Office of the President. The White House*, 2016.

[55] Campolo et al., "AI Now 2017 Report," *AI Now Institute at New York University*, 2017.

[56] Friedman and Nissenbaum, "Bias in Computer Systems," *ACM Transactions on Information Systems (TOIS)* 14, no. 3 (1996): 330–47.

[57] Pedreschi, Ruggieri, and Turini, "Discrimination-Aware Data Mining," in *Proc. 14Th ACM SIGKDD* (ACM, 2008), 560–68.

FAT/ML workshop in 2014, a new community emerged, and the topic has since grown in popularity. Several popular-audience books have delivered critiques of algorithmic systems in modern society.[58]

[58] Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press, 2015); O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Broadway Books, 2016); Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St. Martin's Press, 2018); Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (nyu Press, 2018).

3
Classification

Simply put, the goal of classification is to determine a plausible value for an unknown variable Y given an observed variable X. For example, we might try to *predict* whether a loan applicant will pay back her loan by looking at various characteristics such as credit history, income, and net worth. Classification also applies in situations where the variable Y does not refer to an event that lies in the future. For example, we can try to determine if an image contains a *cat* by looking at the set of pixels encoding the image. This practice is also called *object recognition* or *image classification*. Object recognition might not even seem like a statistical problem, yet statistical methods came to be the method of choice for many important pattern recognition tasks in computer vision.

Supervised learning

A classifier is a mapping from the space of possible values for X to the space of values that the target variable Y can assume. *Supervised learning* is the prevalent method for constructing classifiers from observed data. The essential idea is very simple. Suppose we have labeled data, also called *training examples*, of the form $(x_1, y_1), ..., (x_n, y_n)$, where each *example* is a pair (x_i, y_i) of an *instance* x_i and a *label* y_i.

Instances are usually arranged as vectors of some dimension. You can think of them as arrays with numbers in them. In a classification problem, labels typically come from a discrete set such as $\{-1, 1\}$ in the case of binary classification. We interpret these labels as partitioning the set of instances into positive and negative instances depending on their label.[59] We can interpret such a classifier as a *decision rule* by equating a positive label with *acceptance* and a negative label with *rejection*.

In a *regression* problem, the label y is typically a real number. The goal is no longer to predict the exact value of y but rather to be close

[59] Multi-class prediction is the generalization to label sets with more than two values.

to it. The tools to solve classification and regression problems in practice are very similar. In both cases, roughly the same optimization approach is used to find a classifier f that maps an instance x to a label $\hat{y} = f(x)$ that we hope agrees with the correct label. This optimization process is often called *training*; its specifics are irrelevant for this chapter.

To turn supervised learning into a statistical problem, we assume that there is an underlying distribution from which the data were drawn. The distribution is fixed and each example is drawn independently of the others. We can express this underlying distribution as a pair of random variables (X, Y). For example, our training examples might be responses from a survey. Each survey participant is chosen independently at random from a fixed sampling frame that represents an underlying population. As we discussed in the introduction, the goal of supervised learning is to identify meaningful patterns in the population that aren't just artifacts of the sample.

At the population level, we can interpret our classifier as a random variable by considering $\hat{Y} = f(X)$. In doing so, we overload our terminology slightly by using the word *classifier* for both the random variable $\hat{Y}$ and mapping f. The distinction is mostly irrelevant for this chapter as we will focus on the statistical properties of the joint distribution of the data and the classifier, which we denote as a tuple of three random variables $(X, Y, \hat{Y})$. For now, we ignore how $\hat{Y}$ was learned from a finite sample, what the functional form of the classifier is, and how we estimate various statistical quantities from finite samples. While finite sample considerations are fundamental to machine learning, they are often not specific to the conceptual and technical questions around fairness that we will discuss.

Statistical classification criteria

What makes a classifier *good* for an application and how do we choose one out of many possible classifiers? This question often does not have a fully satisfying answer, but some formal criteria can help highlight different qualities of a classifier that can inform our choice.

Perhaps the most well known property of a classifier $\hat{Y}$ is its *accuracy* defined as $\mathbb{P}\{Y = \hat{Y}\}$, the probability of correctly predicting the target variable. It is common practice to apply the classifier that achieves highest accuracy among those available to us.[60]

[60] We typically don't know the classifier that maximizes accuracy among all possible classifiers, but rather we only have access to those that we can find with effective training procedures.

Accuracy is easy to define, but misses some important aspects. A classifier that always predicts *no traffic fatality in the next year* might have high accuracy, simply because individual accidents are highly unlikely. However, it's a constant function that has no value in assessing the risk that an individual experiences a fatal traffic accident.

Many other formal classification criteria highlight different aspects of a classifier. In a binary classification setting, we can consider the conditional probability $\mathbb{P}\{\text{event} \mid \text{condition}\}$ for various different settings.

Table 1: Common classification criteria

Event	Condition	Resulting notion ($\mathbb{P}\{\text{event} \mid \text{condition}\}$)
$\hat{Y} = 1$	$Y = 1$	True positive rate, recall
$\hat{Y} = 0$	$Y = 1$	False negative rate
$\hat{Y} = 1$	$Y = 0$	False positive rate
$\hat{Y} = 0$	$Y = 0$	True negative rate

To be clear, the true positive rate corresponds to the frequency with which the classifier correctly assigns a positive label to a positive instance. We call this a *true positive*. The other terms *false positive*, *false negative*, and *true negative* derive analogously from the respective definitions.

It is not important to memorize all these terms. They do, however, come up regularly in the classification setting so the table might come in handy.

Another family of classification criteria arises from swapping event and condition. We'll only highlight two of the four possible notions.

Table 2: Additional classification criteria

Event	Condition	Resulting notion ($\mathbb{P}\{\text{event} \mid \text{condition}\}$)
$Y = 1$	$\hat{Y} = 1$	Positive predictive value, precision
$Y = 0$	$\hat{Y} = 0$	Negative predictive value

We'll return to these criteria later on when we explore some of their properties and relationships.

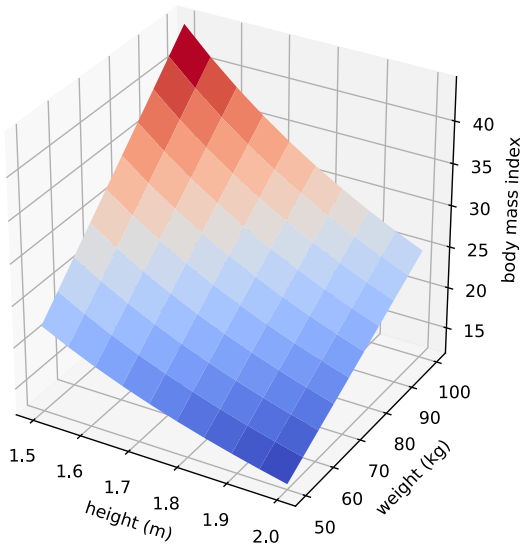
Score functions

Classification is often attacked by first solving a regression problem to summarize the data in a single real-valued variable. We will refer to such a variable as *score*. We can turn a score into a classifier by thresholding it somewhere on the real line.

For an illustrative example consider the well-known body mass index which summarizes *weight* and *height* of a person into a single real number. In our formal notation, the features are $X = (H, W)$ where H denotes height in meters and W denotes weight in kilo-

grams. The body mass index corresponds to the score function $R = W/H^2$.

Figure 3: Plot of the body mass index.



We could interpret the body mass index as measuring risk of heart disease. Thresholding it at the value *27*, we might decide that individuals with a body mass index above this value are at risk of developing heart disease while others are not. It does not take a medical degree to suspect that the resulting classifier may not be very accurate[61]. The body mass index has a number of known issues leading to errors when used for classification. We won't go into detail, but it's worth noting that these classification errors can systematically align with certain demographic groups. For instance, the body mass index tends to be inflated as a risk measure for taller people (due to its scaling issues).

[61] In fact, it seems to be quite poor.

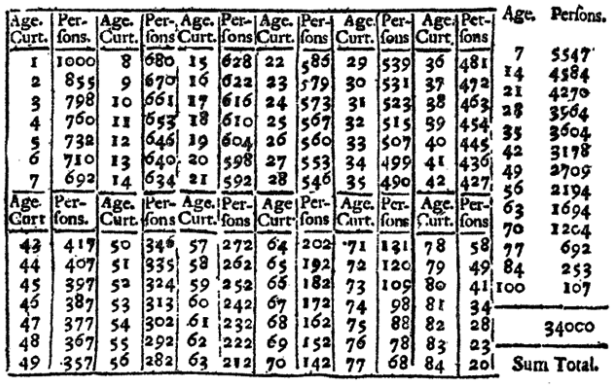
Score functions need not follow simple algebraic formulas such as the body mass index. In most cases, score functions are built by fitting regression models against historical data. Think of a credit score, as is common in some countries, which can be used to accept or deny loan applicants based on the score value. We will revisit this example in detail later.

The conditional expectation

A natural score function is the expectation of the target variable Y conditional on the features X we have observed. We can write this score as $R = r(X)$ where $r(x) = \mathbb{E}[Y \mid X = x]$, or more succinctly, $R = \mathbb{E}[Y \mid X]$. In a sense, this score function gives us the *best guess* for the

target variable given the observations we have. We can think of the conditional expectation as a *lookup table* that gives us for each setting of features the frequency of positive outcomes given these features.[62]

Such lookup tables have a long and fascinating history in applications of risk assessment such as insurance pricing.[63] One of the earliest examples is Halley's *life table* from 1693 that was used to estimate the life expectancy an individual in order to accurately price certain annuities.



Age.	Persons.
7	5547
14	4584
21	4270
28	3564
35	3604
42	3178
49	2709
56	2194
63	1694
70	1204
77	692
84	253
100	107
	34000
	Sum Total.

The conditional expectation also makes sense for our example of scoring risk of heart disease. What it would do here is to tell us for every setting of weight (say, rounded to the nearest kg unit) and every physical height (rounded to the nearest cm unit), the incidence rate of heart disease among individuals with these values of weight and height. The target variable in this case is a binary indicator of heart disease. So, $r((176, 68))$ would be the incidence rate of heart disease among individuals who are 1.76m tall and weigh 68kg. Intuitively, we can think of the conditional expectation as a big lookup table of incidence rates given some setting of characteristics.

The conditional expectation is likely more useful as a risk measure of heart disease than the body mass index we saw earlier. After all, the conditional expectation directly reflects the incidence rate of heart disease given the observed characteristics, while the body mass index is a general-purpose summary statistic.

That said, we can still spot a few issues with this score function. First, our definition of target variable was a bit fuzzy, lumping together all sorts of different kinds of heart disease with different characteristics. Second, in order to actually compute the conditional expectation in practice, we would have to collect incidence rate statistics by height and weight. These data points would only tell us about historical incidence rates. The extent to which they can tell us about future cases of heart disease is somewhat unclear. If our data comes from a time where people generally smoked more cigarettes, our

[62] We can make this statement more precise. This score is sometimes called the *Bayes optimal score* or *Bayes optimal score* as it minimizes the squared error $\mathbb{E}(g(X) - R)^2$ among all functions $g(X)$.

[63] Bouk, *How Our Days Became Numbered: Risk and the Rise of the Statistical Individual* (University of Chicago Press, 2015).

Figure 4: Halley's life table (1693)

statistics might overestimate future incidence rates. There are numerous other features that are relevant for the prediction of heart disease, including age and gender, but they are neglected in our data. We could include these additional features in our data; but as we increase the number of features, estimating the conditional expectation becomes increasingly difficult. Any feature set partitions the population into demographics. The more features we include, the fewer data points we can collect in each subgroup. As a result, the conditional expectation is generally hard to estimate in *high-dimensional* settings, where we have many attributes.

From scores to classifiers

We just saw how we can turn a score function into a discrete classifier by discretizing its values into buckets. In the case of a binary classifier, this corresponds to choosing a threshold t so that when the score is above t our classifier outputs 1 (*accept*) and otherwise -1 (*reject*).[64] Each choice of the threshold defines one binary classifier. Which threshold should we choose?

[64] The choice of the values 1 and -1 is arbitrary. Any two distinct values will do.

The answer to this question is surprisingly subtle. Roughly speaking, which threshold we choose depends on our notion of utility for the resulting classifier and the problem we're trying to solve. Our notion of utility could be complex and depend on many different considerations.

In classification, it is common to oversimplify the problem quite a bit by summarizing all considerations of utility with just two numbers: a cost for accepting a negative instance (false positive) and a cost for rejecting a positive instance (false negative). If in our problem we face a high cost for false positives, we want to choose a higher threshold than in other applications where false negatives are costly.

The choice of a threshold and its resulting trade-off between true positive rate and false positive rate can be neatly visualized with the help of an *ROC curve*[65]. Note that true positive rate equals 1 - false negative rate.

[65] ROC stands for receiver operating characteristic.

The ROC curve serves another purpose. It can be used to eyeball how predictive our score is of the target variable. A common measure of predictiveness is the area under the curve, which is the probability that a random positive instance gets a score higher than a random negative instance. An area of 1/2 corresponds to random guessing, and an area of 1 corresponds to perfect classification, or more formally, the score equals the target. Known disadvantages[66] make *area under the curve* a tool that must be interpreted with caution.

[66] Halligan, Altman, and Mallett, "Disadvantages of Using the Area Under the Receiver Operating Characteristic Curve to Assess Imaging Tests: A Discussion and Proposal for an Alternative Approach," *European Radiology* 25, no. 4 (April 2015): 932–39.

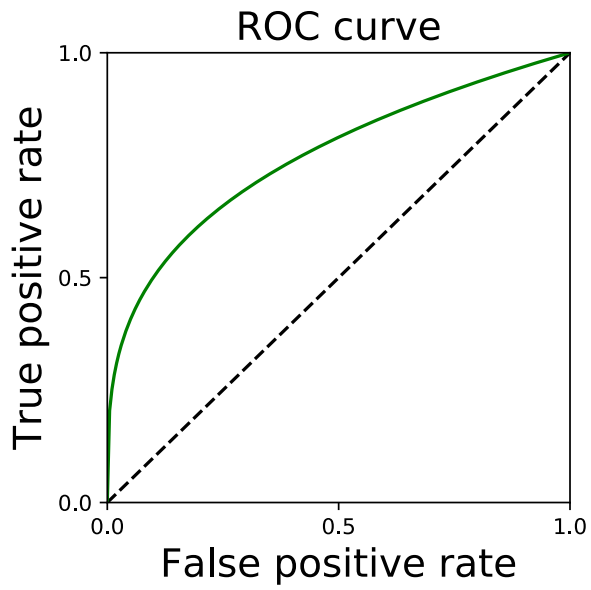


Figure 5: Example of an ROC curve. Each point on the solid curve is realized by thresholding the score function at some value. The dashed line shows the trade-offs achieved by randomly accepting an instance irrespective of its features with some probability $p \in [0, 1]$.

Sensitive characteristics

In many classification tasks, the features X contain or implicitly encode sensitive characteristics of an individual. We will set aside the letter A to designate a discrete random variable that captures one or multiple sensitive characteristics[67]. Different settings of A correspond to different groups of the population. This notational choice is not meant to suggest that we can cleanly partition the set of features into two independent categories such as "neutral" and "sensitive." In fact, we will see shortly that sufficiently many seemingly neutral features can often give high accuracy predictions of sensitive characteristics. This should not be surprising. After all, if we think of A as the target variable in a classification problem, there is reason to believe that the remaining features would give a non-trivial classifier for A.

[67] Note that formally we can always represent any number of discrete sensitive attributes as a single discrete attribute whose support corresponds to each of the possible settings of the original attributes.

The choice of sensitive attributes will generally have profound consequences as it decides which groups of the population we highlight, and what conclusions we draw from our investigation. The taxonomy induced by discretization can on its own be a source of harm if it is too coarse, too granular, misleading, or inaccurate. Even the act of introducing a sensitive attribute on its own can be problematic. We will revisit this important discussion in the next chapter.

No fairness through unawareness

Some have hoped that removing or ignoring sensitive attributes would somehow ensure the impartiality of the resulting classifier.

Unfortunately, this practice is usually somewhere on the spectrum between ineffective and harmful.

In a typical data set, we have many features that are slightly correlated with the sensitive attribute. Visiting the website `pinterest.com`, for example, has a small statistical correlation with being female.[68]

The correlation on its own is too small to predict someone's gender with high accuracy. However, if numerous such features are available, as is the case in a typical browsing history, the task of predicting gender becomes feasible at high accuracy levels.

In other words, several features that are slightly predictive of the sensitive attribute can be used to build high accuracy classifiers for that attribute.

[68] As of August 2017, 58.9% of Pinterest's users in the United States were female. See here (Retrieved 3-27-2018)

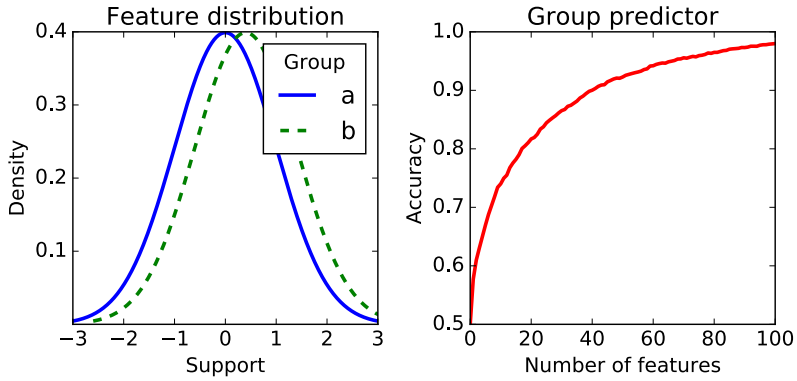


Figure 6: On the left, we see the distribution of a single feature that differs only very slightly between the two groups. In both groups the feature follows a normal distribution. Only the means are slightly different in each group. Multiple features like this can be used to build a high accuracy group membership classifier. On the right, we see how the accuracy grows as more and more features become available.

In large feature spaces sensitive attributes are generally *redundant* given the other features. If a classifier trained on the original data uses the sensitive attribute and we remove the attribute, the classifier will then find a redundant encoding in terms of the other features. This results in an essentially equivalent classifier, in the sense of implementing the same function.

To further illustrate the issue, consider a fictitious start-up that sets out to predict your income from your genome. At first, this task might seem impossible. How could someone's DNA reveal their income? However, we know that DNA encodes information about ancestry, which in turn correlates with income in some countries such as the United States. Hence, DNA can likely be used to predict income better than random guessing. The resulting classifier uses ancestry in an entirely implicit manner. Removing redundant encodings of ancestry from the genome is a difficult task that cannot be accomplished by removing a few individual genetic markers. What we learn from this is that machine learning can wind up building classifiers for sensitive attributes without explicitly being asked to,

simply because it is an available route to improving accuracy.

Redundant encodings typically abound in large feature spaces. What about small hand-curated feature spaces? In some studies, features are chosen carefully so as to be roughly statistically independent of each other. In such cases, the sensitive attribute may not have good redundant encodings. That does not mean that removing it is a good idea. Medication, for example, sometimes depends on race in legitimate ways if these correlate with underlying causal factors.[69] Forcing medications to be uncorrelated with race in such cases can harm the individual.

[69] Bonham, Callier, and Royal, "Will Precision Medicine Move Us Beyond Race?"

Formal non-discrimination criteria

Many *fairness criteria* have been proposed over the years, each aiming to formalize different desiderata. We'll start by jumping directly into the formal definitions of three representative fairness criteria that relate to many of the proposals that have been made.

Once we have acquired familiarity with the technical matter, we'll have a broader debate around the purpose, scope, and meaning of these fairness criteria in Chapter 3.

Most of the proposed fairness criteria are properties of the joint distribution of the sensitive attribute A, the target variable Y, and the classifier or score R.[70] This means that we can write them as some statement involving properties of these three random variables.

To a first approximation, most of these criteria fall into one of three different categories defined along the lines of different (conditional) independence[71] statements between the involved random variables.

[70] If all variables are binary, then the joint distribution is specified by 8 non-negative parameters that sum to 1. A non-trivial property of the joint distribution would restrict the way in which we can choose these parameters.

[71] Learn more about conditional independence here.

Table 3: Non-discrimination criteria

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Below we will introduce and discuss each of these conditions in detail. Variants of these criteria arise from different ways of relaxing them.

As an exercise, think about why we omitted the conditional independence statement $R \perp Y \mid A$ from our discussion here.

Independence

Our first formal criterion simply requires the sensitive characteristic to be statistically independent of the score.

Definition 1. *The random variables* (A, R) *satisfy* independence *if* $A \perp R$.

Independence has been explored through many equivalent terms or variants, referred to as *demographic parity*, *statistical parity*, *group fairness*, *disparate impact* and others. In the case of binary classification, independence simplifies to the condition

$$\mathbb{P}\{R = 1 \mid A = a\} = \mathbb{P}\{R = 1 \mid A = b\}\,,$$

for all groups a, b. Thinking of the event $R = 1$ as "acceptance," the condition requires the acceptance rate to be the same in all groups. A relaxation of the constraint introduces a positive amount of slack $\epsilon > 0$ and requires that

$$\mathbb{P}\{R = 1 \mid A = a\} \geq \mathbb{P}\{R = 1 \mid A = b\} - \epsilon\,.$$

Note that we can swap a and b to get an inequality in the other direction. An alternative relaxation is to consider a ratio condition, such as,

$$\frac{\mathbb{P}\{R = 1 \mid A = a\}}{\mathbb{P}\{R = 1 \mid A = b\}} \geq 1 - \epsilon\,.$$

Some have argued[72] that, for $\epsilon = 0.2$, this condition relates to the *80 percent rule* in disparate impact law.

Yet another way to state the independence condition in full generality is to require that A and R must have zero mutual information[73] $I(A; R) = 0$. The characterization in terms of mutual information leads to useful relaxations of the constraint. For example, we could require $I(A; R) \leq \epsilon$.

[72] Feldman et al., "Certifying and Removing Disparate Impact," in *Proc. 21St SIGKDD* (ACM, 2015).

[73] Mutual information is defined as $I(A; R) = H(A) + H(R) - H(A, R)$, where H denotes the entropy.

Limitations of independence

Independence is pursued as a criterion in many papers, for several reasons. For example, it may be an expression of a belief about human nature, namely that traits relevant for a job are independent of certain attributes. It also has convenient technical properties.

However, decisions based on a classifier that satisfies independence can have undesirable properties (and similar arguments apply to other statistical critiera). Here is one way in which this can happen, which is easiest to illustrate if we imagine a callous or ill-intentioned decision maker. Imagine a company that in group a hires diligently selected applicants at some rate $p > 0$. In group b, the company hires carelessly selected applicants at the same rate p. Even though the acceptance rates in both groups are identical, it is far more likely that unqualified applicants are selected in one group than in the other. As a result, it will appear in hindsight that members of group b performed worse than members of group a, thus establishing a negative track record for group b.[74]

This situation might arise without positing malice: the company might have historically hired employees primarily from group a, giving them a better understanding of this group. As a technical matter, the company might have substantially more training data in group a, thus potentially leading to lower error rates of a learned classifier within that group. The last point is a bit subtle. After all, if both groups were entirely homogenous in all ways relevant to the classification task, more training data in one group would equally benefit both. Then again, the mere fact that we chose to distinguish these two groups indicates that we believe they might be heterogeneous in relevant aspects.

Interlude: How to satisfy fairness criteria

A later chapter devoted to algorithmic interventions will go into detail, but we pause for a moment to think about how we can achieve the independence criterion when we actually build a classifier. We distinguish between three different techniques. While they generally apply to all the criteria and their relaxations that we review in this chapter, our discussion here focuses on independence.

- Pre-processing: Adjust the feature space to be uncorrelated with the sensitive attribute.
- At training time: Work the constraint into the optimization process that constructs a classifier from training data.
- Post-processing: Adjust a learned classifier so as to be uncorrelated with the sensitive attribute.

The three approaches have different strengths and weaknesses.

Pre-processing is a family of techniques to transform a feature space into a representation that as a whole is independent of the sensitive attribute. This approach is generally agnostic to what we do with the new feature space in downstream applications. After the pre-processing transformation ensures independence, any deterministic training process on the new space will also satisfy independence[75].

[75] Formally, this is a consequence of the data processing inequality from information theory.

Achieving independence at training time can lead to the highest utility since we get to optimize the classifier with this criterion in mind. The disadvantage is that we need access to the raw data and training pipeline. We also give up a fair bit of generality as this approach typically applies to specific model classes or optimization problems.

Post-processing refers to the process of taking a trained classifier and adjusting it possibly depending on the sensitive attribute and additional randomness in such a way that independence is achieved.

Formally, we say a *derived classifier* $\hat{Y} = F(R, A)$ is a possibly randomized function of a given score R and the sensitive attribute. Given a cost for false negatives and false positives, we can find the derived classifier that minimizes the expected cost of false positive and false negatives subject to the fairness constraint at hand. Post-processing has the advantage that it works for any *black-box* classifier regardless of its inner workings. There's no need for re-training, which is useful in cases where the training pipeline is complex. It's often also the only available option when we have access only to a trained model with no control over the training process. These advantages of post-processing are simultaneously also a weakness as it often leads to a significant loss in utility.

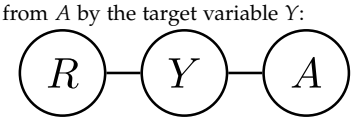
Separation

Our next criterion acknowledges that in many scenarios, the sensitive characteristic may be correlated with the target variable. For example, one group might have a higher default rate on loans than another. A bank might argue that it is a matter of business necessity to therefore have different lending rates for these groups.

Roughly speaking, the separation criterion allows correlation between the score and the sensitive attribute to the extent that it is *justified by the target variable*. This intuition can be made precise with a simple conditional independence statement.

Definition 2. *Random variables* (R, A, Y) *satisfy* separation *if* $R \perp A \mid Y$.[76]

[76] We can display separation as a graphical model in which R is separated from A by the target variable Y:



If you haven't seen graphical models before, don't worry. All this says is that R is conditionally independent of A given Y.

In the case where R is a binary classifier, separation is equivalent to requiring for all groups a, b the two constraints

$$\mathbb{P}\{R = 1 \mid Y = 1, A = a\} = \mathbb{P}\{R = 1 \mid Y = 1, A = b\}$$
$$\mathbb{P}\{R = 1 \mid Y = 0, A = a\} = \mathbb{P}\{R = 1 \mid Y = 0, A = b\}\,.$$

Recall that $\mathbb{P}\{R = 1 \mid Y = 1\}$ is called the *true positive rate* of the classifier. It is the rate at which the classifier correctly recognizes positive instances. The *false positive rate* $\mathbb{P}\{R = 1 \mid Y = 0\}$ highlights the rate at which the classifier mistakenly assigns positive outcomes to negative instances. What separation therefore requires is that all groups experience the same false negative rate and the same false positive rate.

This interpretation in terms of equality of error rates leads to natural relaxations. For example, we could only require equality of false negative rates. A false negative, intuitively speaking, corresponds to denied opportunity in scenarios where acceptance is desirable, such as in hiring.[77]

[77] In contrast, when the task is to identify high-risk individuals, as in the case of recidivism prediction, it is common to denote the undesirable outcome as the "positive" class. This inverts the meaning of false positives and false negatives, and is a frequent source of terminological confusion.

Achieving separation

As was the case with independence, we can achieve separation by post-processing a given score function without the need for retraining.[78]

[78] Recall, a derived classifier is a possible randomized mapping $\hat{Y} = F(R, A)$.

The post-processing step uses the ROC curve that we saw earlier and it's illustrative to go into a bit more detail. A binary classifier that satisfies separation must achieve the same true positive rates and the same false positive rates in all groups. This condition corresponds to taking the intersection of all group-level ROC curves. Within this constraint region, we can then choose the classifier that minimizes the given cost.

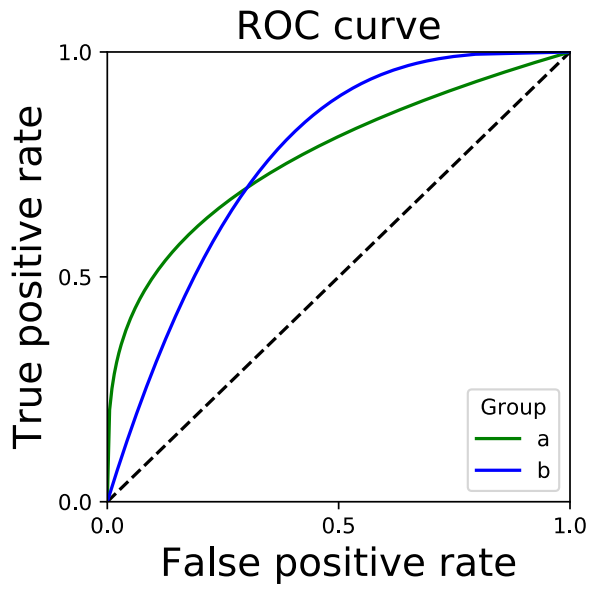


Figure 7: ROC curve by group.

We see the ROC curves of a score displayed for each group separately. The two groups have different curves indicating that not all trade-offs between true and false positive rate are achievable in both groups. The trade-offs that are achievable in both groups are precisely those that lie under both curves, corresponding to the intersection of the regions enclosed by the curves.

The highlighted region is the *feasible region* of trade-offs that we can achieve in all groups. There is a subtlety though. Points that are not exactly on the curves, but rather in the interior of the region, require *randomization*. To understand this point, consider a classifier that accepts everyone corresponding to true and false positive rate 1, the upper right corner of the plot. Consider another classifier that accepts no one, resulting in true and false positive rate 0, the lower left corner of the plot. Now, consider a third classifier that

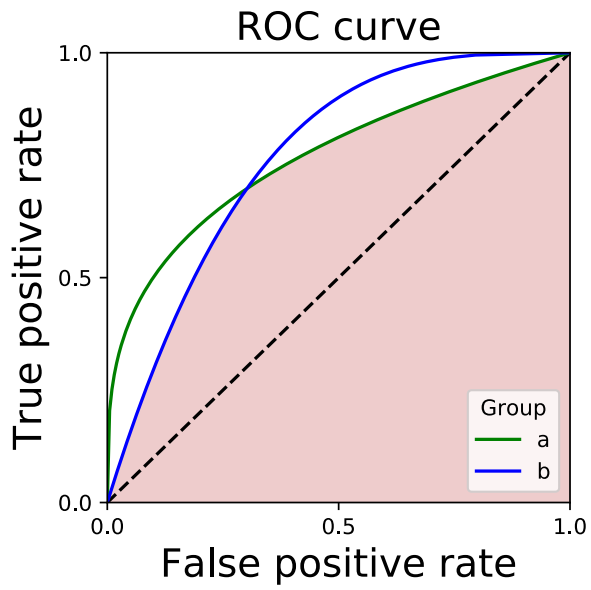


Figure 8: Intersection of area under the curves.

given an instance randomly picks and applies the first classifier with probability $1 - p$, and the second with probability p. This classifier achieves true and false positive rate p thus giving us one point on the dashed line in the plot. In the same manner, we could have picked any other pair of classifiers and randomized between them. We can fill out the entire shaded region in this way, because it is *convex*, meaning that every point in it lies on a line segment between two classifiers on the boundary.

Sufficiency

Our third criterion formalizes that the score already subsumes the sensitive characteristic for the purpose of predicting the target. This idea again boils down to a conditional independence statement.

Definition 3. *We say the random variables* (R, A, Y) *satisfy* sufficiency *if* $Y \perp A \mid R$.[79]

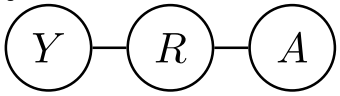
We will often just say that R satisfies *sufficiency* when the sensitive attribute A and target variable Y are clear from the context.

Let us write out the definition more explicitly in the binary case where $Y \in \{0, 1\}$. In this case, a random variable R is sufficient for A if and only if for all groups a, b and all values r in the support of R, we have

$$\mathbb{P}\{Y = 1 \mid R = r, A = a\} = \mathbb{P}\{Y = 1 \mid R = r, A = b\}\,.$$

When R has only two values we recognize this condition as requiring

[79] We can again display sufficiency as a graphical model as we did with separation before:



If you haven't seen graphical models before, feel free to ignore this interpretation.

a parity of positive/negative predictive values across all groups.

While it is often useful to think of sufficiency in terms of positive and negative predictive values, there's a useful alternative. Indeed, sufficiency turns out to be closely related to an important notion called *calibration*, as we will discuss next.

Calibration and sufficiency

In some applications it is desirable to be able to interpret the values of the score functions as probabilities. Formally, we say that a score R is *calibrated* if for all score values r in the support of R, we have

$$\mathbb{P}\{Y = 1 \mid R = r\} = r\,.$$

This condition means that the set of all instances assigned a score value r has an r fraction of positive instances among them. The condition refers to the group of all individuals receiving a particular score value. It does not mean that at the level of a single individual a score of r corresponds to a probability r of a positive outcome. The latter is a much stronger property that is satisfied by the conditional expectation $R = \mathbb{E}[Y \mid X]$.[80]

In practice, there are various heuristics to achieve calibration. For example, *Platt scaling* is a popular method that works as follows. Platt scaling takes a possibly uncalibrated score, treats it as a single feature, and fits a one variable regression model against the target variable based on this feature. More formally, given an uncalibrated score R, Platt scaling aims to find scalar parameters a, b such that the sigmoid function[81]

$$S = \frac{1}{1 + \exp(aR + b)}$$

fits the target variable Y with respect to the so-called *log loss*

$$-\mathbb{E}[Y \log S + (1 - Y) \log(1 - S)].$$

This objective can be minimized given labeled examples drawn from (R, Y) as is standard in supervised learning.

Calibration by group

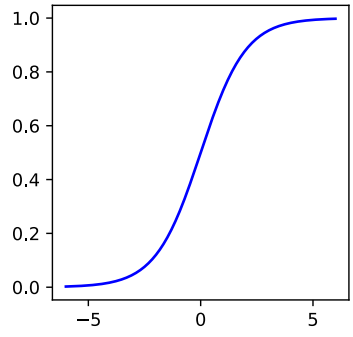
From the definition, we can see that sufficiency is closely related to the idea of calibration. To formalize the connection we say that the score R satisfies *calibration by group* if it satisfies

$$\mathbb{P}\{Y = 1 \mid R = r, A = a\} = r\,,$$

for all score values r and groups a. Recall that calibration is the same requirement at the population level without the conditioning on A.

[80] Formally, we have for every set S, $\mathbb{P}\{Y = 1 \mid R = r, X \in S\} = r$.

[81] A plot of the sigmoid function $1/(1 + \exp(-x))$.



Fact 1. *Calibration by group implies sufficiency.*

Conversely, sufficiency is only slightly weaker than calibration by group in the sense that a simple renaming of score values goes from one property to the other.

Proposition 1. *If a score R satisfies sufficiency, then there exists a function $\ell\colon [0,1] \to [0,1]$ so that $\ell(R)$ satisfies calibration by group.*

Proof. Fix any group a and put $\ell(r) = \mathbb{P}\{Y = 1 \mid R = r, A = a\}$. Since R satisfies sufficiency, this probability is the same for all groups a and hence this map ℓ is the same regardless of what value a we chose.

Now, consider any two groups a, b. We have,

$$\begin{aligned} r &= \mathbb{P}\{Y = 1 \mid \ell(R) = r, A = a\} \\ &= \mathbb{P}\{Y = 1 \mid R \in \ell^{-1}(r), A = a\} \\ &= \mathbb{P}\{Y = 1 \mid R \in \ell^{-1}(r), A = b\} \\ &= \mathbb{P}\{Y = 1 \mid \ell(R) = r, A = b\}\,, \end{aligned}$$

thus showing that $\ell(R)$ is calibrated by group.

□

We conclude that sufficiency and calibration by group are essentially equivalent notions. In particular, this gives us a large repertoire of methods for achieving sufficiency. We could, for example, apply Platt scaling for each of the groups defined by the sensitive attribute.

Calibration by group as a consequence of unconstrained learning

Sufficiency is often satisfied by default without the need for any explicit intervention. Indeed, we generally expect a learned score to satisfy sufficiency in cases where the sensitive attribute can be predicted from the other attributes.

To illustrate this point we look at the calibration values of a standard logistic regression model on the standard UCI adult data set.[82]

We fit a logistic regression model using Python's sklearn library on the UCI training data. The model is then applied to the UCI test data[83]. We make no effort to either tune or calibrate the model.

As we can see from the figure below, the model turns out to be fairly well calibrated by *gender* on its own without any explicit correction.

We see some deviation when we look at calibration by *race*.

The deviation we see in the mid deciles may be due to the scarcity of the test data in the corresponding group and deciles. For example, the 6th decile, corresponding to the score range $(0.5, 0.6]$, on the test

[82] Source

[83] Number of test samples in the UCI data set by group: 1561 Black, 13946 White; 5421 Female, 10860 Male

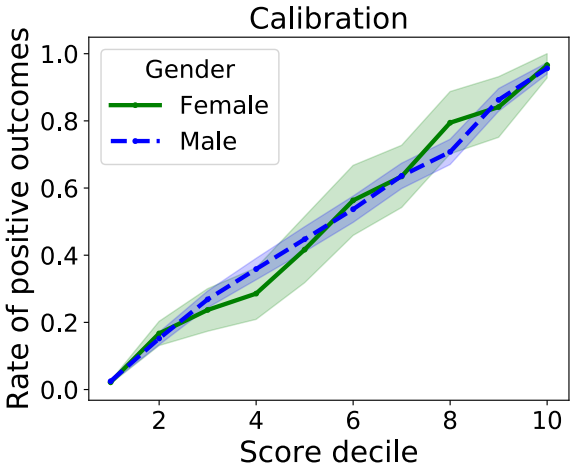


Figure 9: Calibration by gender on UCI adult data. A straight diagonal line would correspond to perfect calibration.

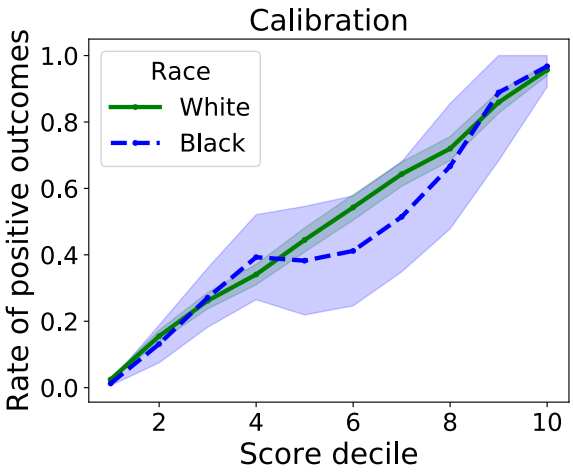


Figure 10: Calibration by race on UCI adult data.

data has only 34 instances with the 'Race' attribute set to 'Black.' As a result, the error bars[84] in this region are rather large.

Continue to explore the UCI Adult data in this code example.

The lesson is that sufficiency often comes for free (at least approximately) as a consequence of standard machine learning practices. The flip side is that imposing sufficiency as a constraint on a classification system may not be much of an intervention. In particular, it would not effect a substantial change in current practices.

Relationships between criteria

The criteria we reviewed constrain the joint distribution in non-trivial ways. We should therefore suspect that imposing any two of them simultaneously over-constrains the space to the point where only degenerate solutions remain. We will now see that this intuition is largely correct.

What this shows is that we cannot impose multiple criteria as hard constraints. This leaves open the possibility that meaningful trade-offs between these different criteria exist.

Independence versus Sufficiency

We begin with a simple proposition that shows how in general independence and sufficiency are mutually exclusive. The only assumption needed here is that the sensitive attribute A and the target variable Y are *not* independent. This is a different way of saying that group membership has an effect on the statistics of the target variable. In the binary case, this means one group has a higher rate of positive outcomes than another. Think of this as the typical case.

Proposition 2. *Assume that A and Y are not independent. Then sufficiency and independence cannot both hold.*

Proof. By the contraction rule for conditional independence,

$$A \perp R \quad \text{and} \quad A \perp Y \mid R \quad \Longrightarrow \quad A \perp (Y, R) \quad \Longrightarrow \quad A \perp Y .$$

To be clear, $A \perp (Y, R)$ means that A is independent of the pair of random variables (Y, R). Dropping R cannot introduce a dependence between A and Y.

In the contrapositive,

$$A \not\perp Y \quad \Longrightarrow \quad A \not\perp R \quad \text{or} \quad A \not\perp R \mid Y .$$

□

Independence versus Separation

An analogous result of mutual exclusion holds for independence and separation. The statement in this case is a bit more contrived and requires the additional assumption that the target variable Y is binary. We also additionally need that the score is not independent of the target. This is a rather mild assumption, since any useful score function should have correlation with the target variable.

Proposition 3. *Assume Y is binary, A is not independent of Y, and R is not independent of Y. Then, independence and separation cannot both hold.*

Proof. Assume $Y \in \{0, 1\}$. In its contrapositive form, the statement we need to show is

$$A \perp R \quad \text{and} \quad A \perp R \mid Y \quad \Longrightarrow \quad A \perp Y \quad \text{or} \quad R \perp Y$$

By the law of total probability,

$$\mathbb{P}\{R = r \mid A = a\} = \sum_y \mathbb{P}\{R = r \mid A = a, Y = y\}\mathbb{P}\{Y = y \mid A = a\}$$

Applying the assumption $A \perp R$ and $A \perp R \mid Y$, this equation simplifies to

$$\mathbb{P}\{R = r\} = \sum_y \mathbb{P}\{R = r \mid Y = y\}\mathbb{P}\{Y = y \mid A = a\}$$

Applied differently, the law of total probability also gives

$$\mathbb{P}\{R = r\} = \sum_y \mathbb{P}\{R = r \mid Y = y\}\mathbb{P}\{Y = y\}$$

Combining this with the previous equation, we have

$$\sum_y \mathbb{P}\{R = r \mid Y = y\}\mathbb{P}\{Y = y\} = \sum_y \mathbb{P}\{R = r \mid Y = y\}\mathbb{P}\{Y = y \mid A = a\}$$

Careful inspection reveals that when y ranges over only two values, this equation can only be satisfied if $A \perp Y$ or $R \perp Y$.

Indeed, we can rewrite the equation more compactly using the symbols $p = \mathbb{P}\{Y = 0\}$, $p_a = \mathbb{P}\{Y = 0 \mid A = a\}$, $r_y = \mathbb{P}\{R = r \mid Y = y\}$, as:

$$pr_0 + (1 - p)r_1 = p_a r_0 + (1 - p_a)r_1.$$

Equivalently, $p(r_0 - r_1) = p_a(r_0 - r_1)$.

This equation can only be satisfied if $r_0 = r_1$, in which case $R \perp Y$, or if $p = p_a$ for all a, in which case $Y \perp A$.

□

The claim is not true when the target variable can assume more than two values, which is a natural case to consider.

Exercise 1. *Give a counterexample to the claim in the previous proposition where the target variable Y assumes three distinct values.*

Separation versus Sufficiency

Finally, we turn to the relationship between separation and sufficiency. Both ask for a non-trivial conditional independence relationship between the three variables A, R, Y. Imposing both simultaneously leads to a degenerate solution space, as our next proposition confirms.

Proposition 4. *Assume that all events in the joint distribution of (A, R, Y) have positive probability, and assume $A \not\perp Y$. Then, separation and sufficiency cannot both hold.*

Proof. A standard fact[85] about conditional independence shows

$$A \perp R \mid Y \quad \text{and} \quad A \perp Y \mid R \quad \Longrightarrow \quad A \perp (R, Y)\,.$$

Moreover,

$$A \perp (R, Y) \quad \Longrightarrow \quad A \perp R \quad \text{and} \quad A \perp Y\,.$$

Taking the contrapositive completes the proof. □

[85] See Theorem 17.2 in (Wasserman, *All of Statistics*)

For a binary target, the non-degeneracy assumption in the previous proposition states that in all groups, at all score values, we have both positive and negative instances. In other words, the score value never fully resolves uncertainty regarding the outcome.

In case the classifier is also binary, we can weaken the assumption to require only that the classifier is imperfect in the sense of making at least one false positive prediction. What's appealing about the resulting claim is that its proof essentially only uses a well-known relationship between true positive rate (recall) and positive predictive value (precision). This trade-off is often called *precision-recall trade-off*.

Proposition 5. *Assume Y is not independent of A and assume $\hat{Y}$ is a binary classifier with nonzero false positive rate. Then, separation and sufficiency cannot both hold.*

Proof. Since Y is not independent of A there must be two groups, call them 0 and 1, such that

$$p_0 = \mathbb{P}\{Y = 1 \mid A = 0\} \neq \mathbb{P}\{Y = 1 \mid A = 1\} = p_1\,.$$

Now suppose that separation holds. Since the classifier is imperfect this means that all groups have the same non-zero false positive rate

FPR > 0, and the same true positive rate TPR ≥ 0. We will show that sufficiency does not hold.

Recall that in the binary case, sufficiency implies that all groups have the same positive predictive value. The positive predictive value in group a, denoted PPV_a satisfies

$$\text{PPV}_\text{a} = \frac{\text{TPR} p_a}{\text{TPR} p_a + \text{FPR}(1 - p_a)} \,.$$

From the expression we can see that $\text{PPV}_0 = \text{PPV}_1$ only if $\text{TPR} = 0$ or $\text{FPR} = 0$. The latter is ruled out by assumption. So it must be that $\text{TPR} = 0$. However, in this case, we can verify that the negative predictive value NPV_0 in group 0 must be different from the negative predictive value NPV_1 in group 1. This follows from the expression

$$\text{NPV}_\text{a} = \frac{(1 - \text{FPR})(1 - p_a)}{(1 - \text{TPR}) p_a + (1 - \text{FPR})(1 - p_a)} \,.$$

Hence, sufficiency does not hold.

□

A good exercise is to derive variants of these trade-offs such as the following.

Exercise 2. *Prove the following result: Assume Y is not independent of A and assume $\hat{Y}$ is a binary classifier with nonzero false positive rate and nonzero true positive rate. Then, if separation holds, there must be two groups with different positive predictive values.*

Inherent limitations of observational criteria

All criteria we've seen so far have one important aspect in common. They are properties of the joint distribution of the score, sensitive attribute, and the target variable. In other words, if we know the joint distribution of the random variables (R, A, Y), we can without ambiguity determine whether this joint distribution satisfies one of these criteria or not.[86]

We can broaden this notion a bit and also include all other features, not just the sensitive attribute. So, let's call a criterion *observational* if it is a property of the joint distribution of the features X, the sensitive attribute A, a score function R and an outcome variable Y.[87] Informally, a criterion is observational if we can express it using probability statements involving the random variables at hand.

Exercise 3. *Convince yourself that independence, separation, and sufficiency are all observational definitions. Come up with a criterion that is* not *observational.*

[86] For example, if all variables are binary, there are eight numbers specifying the joint distributions. We can verify the property by looking only at these eight numbers.

[87] Formally, this means an observational property is defined by set of joint distributions over a given set of variables.

Observational definitions have many appealing aspects. They're often easy to state and require only a lightweight formalism. They make no reference to the inner workings of the classifier, the decision maker's intent, the impact of the decisions on the population, or any notion of whether and how a feature actually influences the outcome. We can reason about them fairly conveniently as we saw earlier. In principle, observational definitions can always be verified given samples from the joint distribution—subject to statistical sampling error.

At the same time, all observational definitions share inherent limitations that we will explore now. Our starting point are two fictitious worlds with substantively different characteristics. We will see that despite their differences these two worlds can map to identical joint distributions. What follows is that all observational criteria will look the same in either world, thus glossing over whatever differences there are.

To develop these two worlds, we'll use the case of a fictitious advertising campaign that targets a hiring ad to software engineers. A score function estimates the likelihood that an individual is a software engineer given some available features.

Scenario I

Imagine we introduce the following random variables in our classification problem.

- A indicates gender
- X_1 indicates whether the user visited `pinterest.com`
- X_2 indicates whether the user visited `github.com`
- R^* is the optimal unconstrained score
- $\tilde{R}$ is the optimal score satisfying separation
- Y indicates whether the user is a software engineer

We can summarize the conditional independence relationships between the variables in a *directed graphical model*.[88] The main fact we need is that a node is conditionally independent of any node that is not a direct ancestor given its parents.

[88] Learn more about graphical models here.

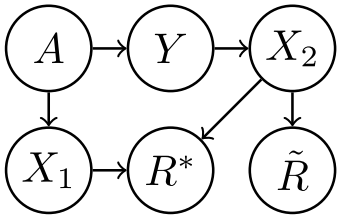


Figure 11: Directed graphical model for the variables in Scenario I

Let's imagine a situation that corresponds to this kind of graphical model. We could argue that gender influences the target variable, since currently software engineers are predominantly male. Gender also influences the first feature, since Pinterest's user base skews female.[89] We assume `github.com` has a male bias. However, this bias is explained by the target variable in the sense that conditional on being a software engineer, all genders are equally likely to visit `github.com`.

[89] As of August 2017, 58.9% of Pinterest's users in the United States were female. See here (Retrieved 3-27-2018)

Once we make these assumptions, we can work out what the optimal unconstrained classifier will do. Both features correlate with the target variable and are therefore useful for prediction. The first feature is predictive since (absent other information) visiting `pinterest.com` suggests female gender, which in turns makes "software engineer" less likely. The second feature is predictive in a more direct sense, as the website is specifically designed for software engineers.

The optimal classifier satisfying separation will refrain from using the first feature (visiting `pinterest.com`). After all, we can see from the graphical model that this feature is not conditionally independent of the sensitive attribute given the target. This score will only use the directly predictive feature `github.com`, which is indeed conditionally independent of gender given the target.

Scenario II

Our two features are different in Scenario II, but all other variables have the same interpretation.

- X_1 indicates whether the user studied computer science
- X_2 indicates whether the user visited the Grace Hopper conference

Although the other variables have the same names and interpretations, we now imagine a very different graphical model.

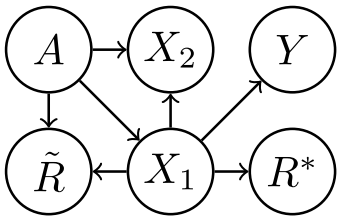


Figure 12: Directed graphical model for the variables in Scenario II

As before, we assume that gender influences the target variable, but now we assume that the target variable is conditionally independent from gender given the first feature. That is, conditional on having studied computer science, all genders are equally likely to go on to become software engineers.[90]

With these assumptions, we can again work out the optimal unconstrained classifier. This time, the optimal unconstrained classifier will only use one feature, namely the first. The reason is that, given the first feature, all remaining features (including the sensitive attribute) become conditionally independent of the target. Therefore, knowing the second feature does not help in predicting the target, once we have the first.

The optimal classifier under separation turns out to be a bit subtle in Scenario II. The issue is that neither of the two features is conditionally independent from the sensitive attribute given the target. The classifier will therefore actively take the sensitive attribute into account in order to *subtract* its influence on the other features.

Different interpretations

Interpreted in the concrete advertising context, the two scenarios don't seem very similar. In particular, the inner workings of the optimal unconstrained classifier in each scenario are rather different. In the first scenario it uses `pinterest.com` as a weak proxy for being *female*, which it then uses as a proxy for not being a software engineer. Software engineers who visit *pinterest.com* might be concerned about this kind of stereotyping, as they might miss out on seeing the ad, and hence the job opportunity. In the second scenario, unconstrained score leads to a classifier that is natural in the sense that it only considers the directly predictive educational information. Absent other features, this would seem agreeable.

Similarly, the optimal classifier satisfying separation behaves differently in the two scenarios. In the first, it corresponds to the natural classifier that only uses `github.com` when predicting *software engineer*. Since `github.com` is primarily a website for software engineers, this seems reasonable. In the second scenario, however, the optimal constrained score performs a subtle adjustment procedure that explicitly takes the sensitive attribute into account. These score functions are also not equivalent from a legal standpoint. One uses the sensitive attribute explicitly for an adjustment step, while the other does not.

Indistinguishability

Despite all their apparent differences, we can instantiate the random variables in each scenario in such a manner that the two scenarios map to identical joint distributions. This means that no property of the joint distribution will be able to distinguish the two scenarios. Whatever property holds for one scenario, it will inevitably also hold for the other. If by some observational criterion we call one scenario *unfair*, we will also have to call the other *unfair*.

Proposition 6. *The random variables in Scenario I and II admit identical joint distributions. In particular, no observational criterion distinguishes between the two scenarios.*

The indistinguishability result has nothing to do with sample sizes or sampling errors. No matter how many data points we have, the size of our data does not resolve the indistinguishability.

There's another interesting consequence of this result. Observational criteria cannot even determine if the sensitive attribute was fed into the classifier or not. To see this, recall that the optimal constrained score in one scenario directly uses *gender*, in the other it does not.

A forced perspective problem

To understand the indistinguishability result, it's useful to draw an analogy with a *forced perspective* problem. Two different objects can appear identical when looked at from a certain fixed perspective.

A data set always forces a particular perspective on reality. There is a possibility that this perspective makes it difficult to identify certain properties of the real world. Even if we have plenty of data, so long as this data comes from the same distribution, it still represents the same perspective. Having additional data is a bit like increasing the resolution of our camera. It helps with some problems, but it doesn't change the angle or the position of the camera.

The limitations of observational criteria are fundamentally the limitations of a single perspective. When analyzing a data set through the lens of observational criteria we do not evaluate alternatives to the data we have. Observational criteria do not tell us what is missing from our perspective.

What then is *not* observational and how do we go beyond observational criteria? This is a profound question that will be the focus of later chapters. In particular, we will introduce the technical repertoire of measurement and causality to augment the classification paradigm. Both measurement and causality give us mechanisms to interrogate, question, and change the perspective suggested by our data.

Case study: Credit scoring

We now apply some of the notions we saw to credit scoring. Credit scores support lending decisions by giving an estimate of the risk that a loan applicant will default on a loan. Credit scores are widely used in the United States and other countries when allocating credit, ranging from micro loans to jumbo mortgages. In the United States, there are three major credit-reporting agencies that collect data on

various lendees. These agencies are for-profit organizations that each offer risk scores based on the data they collected. FICO scores are a well-known family of proprietary scores developed by FICO and sold by the three credit reporting agencies.

Regulation of credit agencies in the United States started with the Fair Credit Reporting Act, first passed in 1970, that aims to promote the accuracy, fairness, and privacy of consumer of information collected by the reporting agencies. The Equal Credit Opportunity Act, a United States law enacted in 1974, makes it unlawful for any creditor to discriminate against any applicant the basis of race, color, religion, national origin, sex, marital status, or age.

Score distribution

Our analysis relies on data published by the Federal Reserve.[91] The data set provides aggregate statistics from 2003 about a credit score, demographic information (race or ethnicity, gender, marital status), and outcomes (to be defined shortly). We'll focus on the joint statistics of score, race, and outcome, where the race attributes assume four values detailed below.[92]

[91] The Federal Reserve Board, "Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit" (https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/, 2007).

[92] These numbers come from the "Estimation sample" column of Table 9 on this web page.

Table 4: Credit score distribution by ethnicity

Race or ethnicity	Samples with both score and outcome
White	133,165
Black	18,274
Hispanic	14,702
Asian	7,906
Total	174,047

The score used in the study is based on the TransUnion TransRisk score. TransUnion is a US credit-reporting agency. The TransRisk score is in turn based on a proprietary model created by FICO, hence often referred to as FICO scores. The Federal Reserve renormalized the scores for the study to vary from 0 to 100, with 0 being *least creditworthy*.

The information on race was provided by the Social Security Administration, thus relying on self-reported values.

The cumulative distribution of these credit scores strongly depends on the group as the next figure reveals.

For an extensive documentation of the data set see the Federal Reserve report.

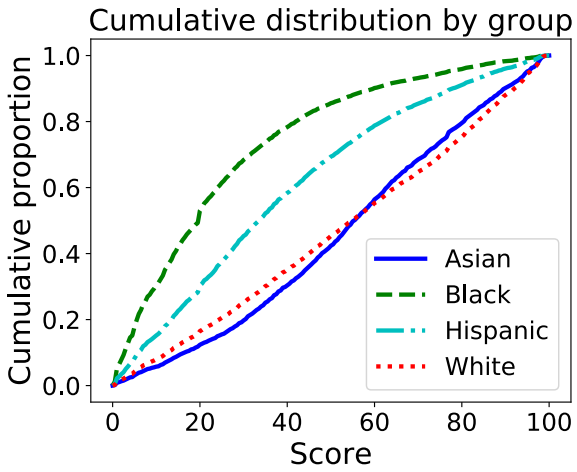


Figure 13: Cumulative density of scores by group.

Performance variables and ROC curves

As is often the case, the outcome variable is a subtle aspect of this data set. Its definition is worth emphasizing. Since the score model is proprietary, it is not clear what target variable was used during the training process. What is it then that the score is trying to predict? In a first reaction, we might say that the goal of a credit score is to predict a *default* outcome. However, that's not a clearly defined notion. Defaults vary in the amount of debt recovered, and the amount of time given for recovery. Any single binary performance indicator is typically an oversimplification.

What is available in the Federal Reserve data is a so-called *performance* variable that measures a *serious delinquency in at least one credit line of a certain time period*. More specifically,

> (the) measure is based on the performance of new or existing accounts and measures whether individuals have been late 90 days or more on one or more of their accounts or had a public record item or a new collection agency account during the performance period.[93]

[93] Quote from the Federal Reserve report.

With this performance variable at hand, we can look at the ROC curve to get a sense of how predictive the score is in different demographics.

The meaning of true positive rate is *the rate of predicted positive performance given positive performance.* Similarly, false positive rate is *the rate of predicted negative performance given a positive performance*.

We see that the shapes appear roughly visually similar in the groups, although the 'White' group encloses a noticeably larger area under the curve than the 'Black' group. Also note that even two ROC curves with the same shape can correspond to very different score

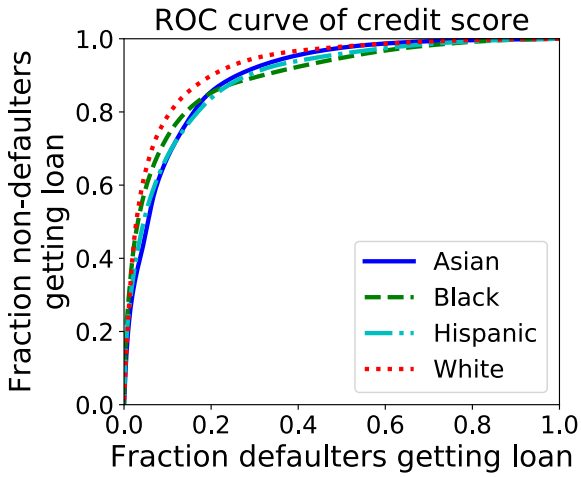


Figure 14: ROC curve of credit score by group.

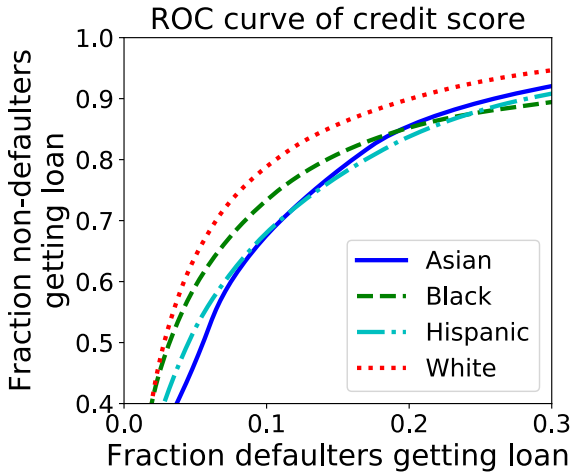


Figure 15: ROC curve of credit score by group zoomed in on region of large differences.

functions. A particular trade-off between true positive rate and false positive rate achieved at a threshold t in one group could require a different threshold t' in the other group.

Comparison of different criteria

With the score data at hand, we compare four different classification strategies:

- *Maximum profit:* Pick possibly group-dependent score thresholds in a way that maximizes profit.
- *Single threshold:* Pick a single uniform score threshold for all groups in a way that maximizes profit.
- *Separation:* Achieve an equal true/false positive rate in all groups. Subject to this constraint, maximize profit.
- *Independence:* Achieve an equal acceptance rate in all groups. Subject to this constraint, maximize profit.

To make sense of maximizing profit, we need to assume a reward for a true positive (correctly predicted positive performance), and a cost for false positives (negative performance predicted as positive). In lending, the cost of a false positive is typically many times greater than the reward for a true positive. In other words, the interest payments resulting from a loan are relatively small compared with the loan amount that could be lost. For illustrative purposes, we imagine that the cost of a false positive is 6 times greater than the return on a true positive. The absolute numbers don't matter. Only the ratio matters. This simple cost structure glosses over a number of details that are likely relevant for the lender such as the terms of the loan.

There is another major caveat to the kind of analysis we're about to do. Since we're only given aggregate statistics, we cannot retrain the score with a particular classification strategy in mind. The only thing we can do is to define a setting of thresholds that achieves a particular criterion. This approach may be overly pessimistic with regards to the profit achieved subject to each constraint. For this reason and the fact that our choice of cost function was rather arbitrary, we do not state the profit numbers. The numbers can be found in the original analysis,[94] which reports that 'single threshold' achieves higher profit than 'separation,' which in turn achieves higher profit than 'independence.'

[94] Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning," in *Proc.* 29*Th NIPS*, 2016, 3315–23.

What we do instead is to look at the different trade-offs between true and false positive rate that each criterion achieves in each group.

We can see that even though the ROC curves are somewhat similar, the resulting trade-offs can differ widely by group for some

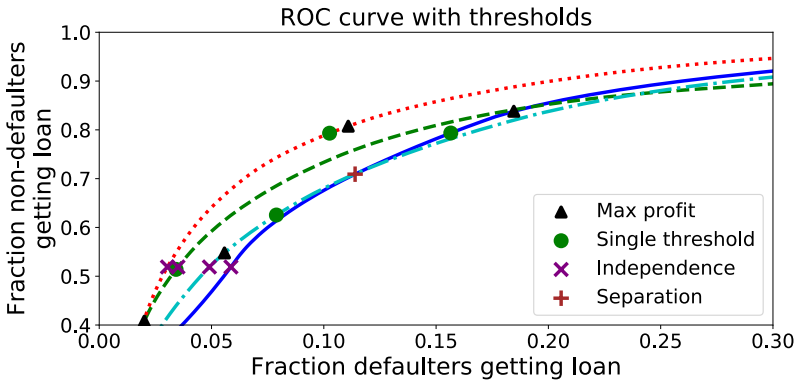
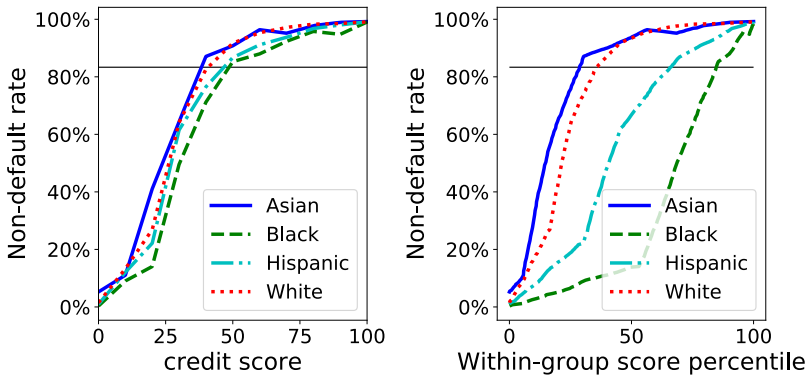


Figure 16: ROC curves with thresholds induced by different criteria.

of the criteria. The true positive rate achieved by *max profit* for the Asian group is twice of what it is for the Black group. The separation criterion, of course, results in the same trade-off in all groups. Independence equalizes acceptance rate, but leads to widely different trade-offs. For instance, the Asian group has a false positive rate more than three times the false positive rate within the Black group.

Calibration values

Finally, we consider the non-default rate by group. This corresponds to the calibration plot by group.[95]



[95] The error bars on the these plots were omitted as they are generally small except for very low score values (0-5) where few samples are available.

Figure 17: Calibration values of credit score by group.

We see that the performance curves by group are reasonably well aligned. This means that a monotonic transformation of the score values would result in a score that is roughly calibrated by group according to our earlier definition. Due to the differences in score distribution by group, it could nonetheless be the case that

thresholding the score leads to a classifier with different positive predictive values in each group.

Feel free to continue exploring the data in this code repository.

Problem set: Criminal justice case study

Risk assessment is an important component of the criminal justice system. In the United States, judges set bail and decide pre-trial detention based on their assessment of the risk that a released defendant would fail to appear at trial or cause harm to the public. While *actuarial risk assessment* is not new in this domain, there is increasing support for the use of learned risk scores to guide human judges in their decisions. Proponents argue that machine learning could lead to greater efficiency and less biased decisions compared with human judgment. Critical voices raise the concern that such scores can perpetuate inequalities found in historical data, and systematically harm historically disadvantaged groups.

In this problem set[96], we'll begin to scratch at the surface of the complex criminal justice domain. Our starting point is an investigation carried out by Propublica[97] of a proprietary risk score, called COMPAS score. These scores are intended to assess the risk that a defendant will re-offend, a task often called *recidivism prediction*. Within the academic community, the ProPublica article drew much attention to the trade-off between separation and sufficiency that we saw earlier.

We'll use data obtained and released by ProPublica as a result of a public records request in Broward Country, Florida, concerning the COMPAS recidivism prediction system. The data is available here. Following ProPublica's analysis, we'll filter out rows where `days_b_screening_arrest` is over 30 or under -30, leaving us with 6,172 rows.

[96] Solutions to these problems are available to course instructors on request.

[97] Angwin et al., "Machine Bias," *Pro Publica*, 2016.

Calibration/sufficiency

- Plot the fraction of defendants recidivating within two years (`two_year_recid == 1`) as a function of risk score (`decile_score`), for Black defendants (`race == "African-American"`) and White defendants (`race == "Caucasian"`).
- Based on these plots, does the risk score satisfy sufficiency across racial groups in this dataset? This is somewhat subjective, since we want to allow for approximate equality between groups; justify your answer in a sentence or two.

Error rates/separation

- Plot the distribution of scores received by the positive class (recidivists) and the distribution of scores received by the negative class (non-recidivists) for Black defendants and for White defendants.
- Based on these plots, does COMPAS achieve separation between the risk score and race?
- Report the Positive Predictive Value, False Positive Rate, and False Negative Rate for a risk threshold of 4 (i.e., defendants with `decile_score >= 4` are classified as high risk), for Black defendants and for White defendants.
- Can we pick two thresholds (one for Black defendants, one for White defendants) such that FPR and FNR are roughly equal for the two groups (say, within 1% of each other)? What is the PPV for the two groups in this case? Note: trivial thresholds of 0 or 11 don't count.

Risk factors and interventions

- Report the recidivism rate of defendants aged 25 or lower, and defendants aged 50 or higher. Note the stark difference between the two: younger defendants are far more likely to recidivate.

The following questions are best viewed as prompts for a class discussion.

- Suppose we are interested in taking a data-driven approach to changing the criminal justice system. Under a theory of incarceration as incapacitation (prevention of future crimes by removal of individuals from society), how might we act on the finding that younger defendants are more likely to reoffend?
- How might we act on this finding under a rehabilitative approach to justice, in which we seek to find interventions that minimize a defendant's risk of recidivism?
- Under a retributive theory of justice, punishment is based in part on culpability, or blameworthiness; this in turn depends on how much control the defendant had over their actions. Under such a theory, how might we act on the finding that younger defendants are more likely to reoffend (and, more generally, commit offenses at all)?

Problem set: Data modeling of traffic stops

For this problem we'll use data released by the Stanford Open Policing Project (SOPP) for the state of North Carolina, available here. It

contains records of 9.6 million police stops in the state between 2000 and 2015.

General notes and hints:

- The *stop rates* section of this problem requires linking SOPP data to census data, whereas the rest is based only on SOPP data and no external datasets. So you might want to work on *post-stop outcomes* and the following sections first, so that you can get familiar with the SOPP data before having to also deal with the census data.
- Throughout this problem, report any data cleaning steps (such as dropping some rows) that you took. Also report any ambiguities you encountered and how you resolved them.

Stop rates

Part A

- For each possible group defined by race, age, gender, location, and year, where:
  - race is one of "Asian," "Black," "Hispanic," "White"
  - age is one of the buckets 15–19, 20–29, 30–39, 40–49, and 50+.
  - gender is one of "female," "male"
  - location is a state patrol troop district
  - and year is between 2010 and 2015, inclusive
- report the following:
  - the population of the group from census data, and
  - the number of stops in that group from SOPP data.

The census data is available here and the fields are explained here. Your data should look like the table below.

Table 5: Census data

Race	Age	Gender	Location	Year	Population	Count
Hispanic	30-39	F	B5	2012	434	76
White	40-49	F	C8	2011	2053	213
Asian	15-19	M	A2	2012	2	0
White	20-29	M	A6	2011	8323	1464
Hispanic	20-29	F	D3	2010	393	56
Black	40-49	F	D7	2011	1832	252
Asian	30-39	M	E6	2013	503	34
Asian	15-19	F	B5	2015	12	4
White	20-29	M	A5	2012	12204	1852
Black	15-19	F	H1	2011	1281	55

Notes and hints:

- The table is a small sample of rows from the actual answer. You can use it to check your answers. There should be about 13,000 rows in the table in total.
- The relevant fields in the census data are `AA_[FE]MALE`, `BA_[FE]MALE`, `H_[FE]MALE`, `WA_[FE]MALE`.
- The relevant fields in the SOPP data are `driver_race`, `driver_age`, `driver_gender`, `district`, and `stop_date`.
- The census data is grouped by county, which is more granular than district.The mapping from county to district is available from SOPP here.

Part B

- Fit a negative binomial regression to your data from part (A) as given in page 5 of the SOPP paper. Report the coefficients of race, age, and gender, and the overdispersion parameter ϕ. Based on these coefficients, what is the ratio of stop rates of Hispanic drivers to White drivers, and Black drivers to White drivers, controlling for age, gender, location, and year?

Notes and hints:

- This and the following tasks will be easier using a data modeling framework such as R or statsmodels rather than an algorithmic modeling framework such as scikit-learn.
- The "Population" column in your data corresponds to the "exposure" variable in most frameworks. Equivalently, "offset" is the log of the exposure.
- The coefficients of the different values of each variable (e.g. female and male) are not interpretable individually; only the difference is interpretable.
- Treat `year` as a categorical rather than a continuous variable.

Part C

- Give three distinct potential reasons for the racial disparity in stop rate as measured in part B.

Post-stop outcomes

Part D

- Controlling for age (bucketed as in parts A & B), gender, year, and location, use logistic regression to estimate impact of race on
  - probability of a search (`search_conducted`)

  - probability of arrest (`is_arrested`),
  - probability of a citation (`stop_outcome == "Citation"`)

- For each of the three outcomes, report the coefficients of race, age, and gender along with standard errors of those coefficients. Feel free to sample the data for performance reasons, but if you do, make sure that all standard errors are < 0.1.

Part E

- Interpret the coefficients you reported in part D.
  - What is the ratio of the probability of search of Hispanic drivers to White drivers? Black drivers to White drivers?
  - Repeat the above for the probability of arrest instead of search.
  - What is the difference in citation probability between Hispanic drivers and White drivers? Black drivers and White drivers?
  - Comment on the age and gender coefficients in the regressions.

Notes and hints:

- Interpreting the coefficients is slightly subjective. Since the search and arrest rates are low, in those regressions we can approximate the $1/(1+e^{-\beta x})$ formula in logistic regression as $e^{\beta x}$, and thus we can use differences in β between groups to calculate approximate ratios of search/arrest probabilities.
- This trick doesn't work for citation rates, since those are not low. However, we can pick "typical" values for the control variables, calculate citation rates, and find the difference in citation rate between groups. The results will have little sensitivity to the values of the control variables that we pick.

Part F

Explain in a sentence or two why we control for variables such as gender and location in the regression, and why the results might not be what we want if we don't control for them. (In other words, explain the idea of a confound in this context.)

Part G

However, decisions about what to control are somewhat subjective. What is one reason we might *not* want to control for location in testing for discrimination? In other words, how might we underestimate discrimination if we control for location? (Hint: broaden the idea of discrimination from individual officers to the systemic aspects of policing.)

Data quality

Part H

The SOPP authors provide a README file in which they note the incompleteness, errors, and missing values in the data on a state-by-state level. Pick any two items from this list and briefly explain how each could lead to errors or biases in the analyses you performed (or in the other analyses performed in the paper).

Notes and hints:

- Here is one example: For North Carolina, stop time is not available for a subset of rows. Suppose we throw out the rows with missing stop time (which we might have to if that variable is one of the controls in our regression). These rows might not be a random subset of rows: they could be correlated with location, because officers in some districts don't record the stop time. If so, we might incorrectly estimate race coefficients, because officer behavior might also be correlated with location.

What is the purpose of a fairness criterion?

There is an important question we have neglected so far. Although we have seen several demographic classification criteria and explored their formal properties and the relationships between them, we haven't yet clarified the purpose of these criteria. This is a difficult normative question that will be a central concern of the next chapter. Let us address it briefly here.

Take the independence criterion as an example. Some support this criterion based on the belief that certain intrinsic human traits such as intelligence are independent of, say, race or gender. Others argue for independence based on their desire to live in a society where the sensitive attribute is statistically independent of outcomes such as financial well-being. In one case, independence serves as a proxy for a belief about human nature. In the other case, it represents a long-term societal goal. In either case, does it then make sense to impose independence as a constraint on a classification system?

In a lending setting, for example, independence would result in the same rate of lending in all demographic groups defined by the sensitive attribute, regardless of the fact that individuals' ability to repay might be distributed differently in different groups. This makes it hard to predict the long-term impact of an intervention that imposes independence as a hard classification constraint. It is not clear how to account for the impact of the fact that giving out loans to individuals who cannot repay them impoverishes the individual who defaults (in addition to diminishing profits for the bank).

Without an accurate model of long-term impact it is difficult to foresee the effect that a fairness criterion would have if implemented

as a hard classification constraint. However, if such a model of long-term impact model were available, directly optimizing for long-term benefit may be a more effective intervention than to impose a general and crude demographic criterion.[98]

[98] Liu et al., "Delayed Impact of Fair Machine Learning," in *Proc. 35Th ICML*, 2018, 3156–64.

If demographic criteria are not useful as direct guides to fairness interventions, how should we use them then? An alternative view is that classification criteria have *diagnostic value* in highlighting different social costs of the system. Disparities in true positive rates or false positive rates, for example, indicate that two or more demographic groups experience different costs of classification that are not necessarily reflected in the cost function that the decision maker optimized.

At the same time, the diagnostic value of fairness criteria is subject to the fundamental limitations that we saw. In particular, we cannot base a conclusive argument of fairness or unfairness on the value of any observational criterion alone. Furthermore, Corbett-Davies et al.[99] make the important point that statistics such as positive predictive values or false positive rates can be manipulated through external (and possibly harmful) changes to the real world processes reflected in the data. In the context of recidivism prediction in criminal justice, for example, we could artificially lower the false positive rate in one group by arresting innocent people and correctly classifying them as low risk. This external intervention will decrease the false positive rate at the expense of a clearly objectionable practice.

[99] Corbett-Davies et al., "Algorithmic Decision Making and the Cost of Fairness," *arXiv Preprint arXiv:1701.08230*, 2017.

Bibliographic notes and further reading

The fairness criteria reviewed in this chapter were already known in the 1960s and 70s, primarily in the education testing and psychometrics literature.[100] An important fairness criterion is due to Cleary[101] and compares regression lines between the test score and the outcome in different groups. A test is considered *fair* by the Cleary criterion if the slope of these regression lines is the same for each group. This turns out to be equivalent to the sufficiency criterion, since it means that at a given score value all groups have the same rate of positive outcomes.

Einhorn and Bass[102] considered equality of precision values, which is a relaxation of sufficiency as we saw earlier. Thorndike[103] considered a weak variant of calibration by which the frequency of positive predictions must equal the frequency of positive outcomes in each group, and proposed achieving it via a post-processing step that sets different thresholds in different groups. Thorndike's criterion is incomparable to sufficiency in general.

Darlington[104] stated four different criteria in terms of succinct

[100] We are greatful to Ben Hutchinson for bringing these to our attention.

[101] Cleary, "Test Bias: Validity of the Scholastic Aptitude Test for Negro and White Students in Integrated Colleges," *ETS Research Bulletin Series* 1966, no. 2 (1966): i–23; Cleary, "Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges," *Journal of Educational Measurement* 5, no. 2 (1968): 115–24.

[102] Einhorn and Bass, "Methodological Considerations Relevant to Discrimination in Employment Testing." *Psychological Bulletin* 75, no. 4 (1971): 261.

[103] Thorndike, "Concepts of Culture-Fairness," *Journal of Educational Measurement* 8, no. 2 (1971): 63–70.

[104] Darlington, "Another Look at 'Cultural Fairness'," *Journal of Educational Measurement* 8, no. 2 (1971): 71–82.

expressions involving the correlation coefficients between various pairs of random variables. These criteria include independence, a relaxation of sufficiency, a relaxation of separation, and Thorndike's criterion. Darlington included an intuitive visual argument showing that the four criteria are incompatible except in degenerate cases.

Lewis[105] reviewed three fairness criteria including equal precision and equal true/false positive rates.

These important early works were re-discovered later in the machine learning and data mining community. Numerous works considered variants of independence as a fairness constraint.[106] Feldman et al.[107] studied a relaxation of demographic parity in the context of disparate impact law. Zemel et al.[108] adopted the mutual information viewpoint and proposed a heuristic pre-processing approach for minimizing mutual information. Dwork et al.[109] argued that the independence criterion was inadequate as a fairness constraint.

The separation criterion appeared under the name *equalized odds*,[110] alongside the relaxation to equal false negative rates, called *equality of opportunity*. These criteria also appeared in an independent work[111] under different names. Woodworth et al.[112] studied a relaxation of separation stated in terms of correlation coefficients. This relaxation corresponds to the third criterion studied by Darlington.[113]

ProPublica[114] implicitly adopted equality of false positive rates as a fairness criterion in their article on COMPAS scores. Northpointe, the maker of the COMPAS software, emphasized the importance of calibration by group in their rebuttal[115] to ProPublica's article. Similar arguments were made quickly after the publication of ProPublica's article by bloggers including Abe Gong.[116] There has been extensive scholarship on the actuarial risk assessment in criminal justice that long predates the ProPublica debate; Berk et al.[117] provide a survey with commentary.

Variants of the trade-off between separation and sufficiency were shown by Chouldechova[118] and Kleinberg et al.[119] Each of them considered somewhat different criteria to trade off. Chouldechova's argument is very similar to the proof we presented that invokes the relationship between positive predictive value and true positive rate. Subsequent work[120] considers trade-offs between relaxed and approximate criteria. The other trade-off results presented in this chapter are new to this book. The proof of the proposition relating separation and independence for binary classifiers, as well as the counterexample for ternary classifiers, is due to Shira Mitchell and Jackie Shadlen, pointed out to us in personal communication.

The unidentifiability result for observational criteria is due to Hardt, Price, and Srebro,[121] except for minor changes in the choice of graphical models and their interpretation.

[105] Lewis, "A Comparison of Three Models for Determining Test Fairness" (Federal Aviation Administration Washington DC Office of Aviation Medicine, 1978).

[106] Calders, Kamiran, and Pechenizkiy, "Building Classifiers with Independency Constraints," in *In Proc. IEEE ICDMW*, 2009, 13–18; Kamiran and Calders, "Classifying Without Discriminating," in *Proc. 2Nd International Conference on Computer, Control and Communication*, 2009.

[107] Feldman et al., "Certifying and Removing Disparate Impact."

[108] Zemel et al., "Learning Fair Representations," in *Proc. 30Th ICML*, 2013.

[109] Dwork et al., "Fairness Through Awareness."

[110] Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning."

[111] Zafar et al., "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment," in *Proc. 26Th WWW*, 2017.

[112] Woodworth et al., "Learning Non-Discriminatory Predictors," in *Proc. 30Th COLT*, 2017, 1920–53.

[113] Darlington, "Another Look at 'Cultural Fairness'."

[114] Angwin et al., "Machine Bias."

[115] Dieterich, Mendoza, and Brennan, "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity," 2016, https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html.

[116] See this and subsequent posts.

[117] Berk et al., "Fairness in Criminal Justice Risk Assessments: The State of the Art," *ArXiv e-Prints* 1703.09207 (2017).

[118] Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," in *Proc. 3Rd FATML*, 2016.

[119] Kleinberg, Mullainathan, and Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," *arXiv Preprint arXiv:1609.05807*, 2016.

[120] Pleiss et al., "On Fairness and Calibration," in *Proc. 30Th NIPS*, 2017.

[121] Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning."

A dictionary of criteria

For convenience we collect some demographic fairness criteria below that have been proposed in the past (not necessarily including the original reference). We'll match them to their closest relative among the three criteria independence, separation, and sufficiency. This table is meant as a reference only and is not exhaustive. There is no need to memorize these different names.

Table 6: List of demographic fairness criteria

Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	Equivalent	Darlington (1971)
Equal opportunity	Separation	Relaxation	Hardt, Price, Srebro (2016)
Equalized odds	Separation	Equivalent	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Separation	Equivalent	Berk et al. (2017)
Avoiding disparate mistreatment	Separation	Equivalent	Zafar et al. (2017)
Balance for the negative class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Separation	Relaxation	Chouldechova (2016)
Equalized correlations	Separation	Relaxation	Woodworth (2017)
Darlington criterion (3)	Separation	Relaxation	Darlington (1971)
Cleary model	Sufficiency	Equivalent	Cleary (1966)
Conditional use accuracy	Sufficiency	Equivalent	Berk et al. (2017)
Predictive parity	Sufficiency	Relaxation	Chouldechova (2016)
Calibration within groups	Sufficiency	Equivalent	Chouldechova (2016)
Darlington criterion (1), (2)	Sufficiency	Relaxation	Darlington (1971)

4
Legal background and normative questions

Coming up soon!

5
Causality

Our starting point is the difference between an observation and an action. What we see in passive observation is how individuals follow their routine behavior, habits, and natural inclination. Passive observation reflects the state of the world projected to a set of features we chose to highlight. Data that we collect from passive observation show a snapshot of our world as it is.

There are many questions we can answer from passive observation alone: Do 16 year-old drivers have a higher incidence rate of traffic accidents than 18 year-old drivers? Formally, the answer corresponds to a difference of conditional probabilities assuming we model the population as a distribution as we did in the last chapter. We can calculate the conditional probability of a traffic accident given that the driver's age is 16 years and subtract from it the conditional probability of a traffic accident given the age is 18 years. Both conditional probabilities can be estimated from a large enough sample drawn from the distribution, assuming that there are both 16 year old and 18 year old drivers. The answer to the question we asked is solidly in the realm of observational statistics.

But important questions often are not observational in nature. Would traffic fatalities decrease if we raised the legal driving age by two years? Although the question seems similar on the surface, we quickly realize that it asks for a fundamentally different insight. Rather than asking for the frequency of an event in our manifested world, this question asks for the effect of a hypothetical action.

As a result, the answer is not so simple. Even if older drivers have a lower incidence rate of traffic accidents, this might simply be a consequence of additional driving experience. There is no obvious reason why an 18 year old with two months on the road would be any less likely to be involved in an accident than, say, a 16 year-old with the same experience. We can try to address this problem by holding the number of months of driving experience fixed, while comparing individuals of different ages. But we quickly

run into subtleties. What if 18 year-olds with two months of driving experience correspond to individuals who are exceptionally cautious and hence—by their natural inclination—not only drive less, but also more cautiously? What if such individuals predominantly live in regions where traffic conditions differ significantly from those in areas where people feel a greater need to drive at a younger age?

We can think of numerous other strategies to answer the original question of whether raising the legal driving age reduces traffic accidents. We could compare countries with different legal driving ages, say, the United States and Germany. But again, these countries differ in many other possibly relevant ways, such as, the legal drinking age.

At the outset, causal reasoning is a conceptual and technical framework for addressing questions about the effect of hypothetical actions or *interventions*. Once we understand what the effect of an action is, we can turn the question around and ask what action plausibly *caused* an event. This gives us a formal language to talk about cause and effect.

Not every question about cause is equally easy to address. Some questions are overly broad, such as, "What is the cause of success?" Other questions are too specific: "What caused your interest in 19th century German philosophy?" Neither question might have a clear answer. Causal inference gives us a formal language to ask these questions, in principle, but it does not make it easy to choose the right questions. Nor does it trivialize the task of finding and interpreting the answer to a question. Especially in the context of fairness, the difficulty is often in deciding what the question is that causal inference is the answer to.

In this chapter, we will develop sufficient technical understanding of causality to support at least three different purposes.

The first is to conceptualize and address some limitations of the observational techniques we saw in Chapter 2. The second is to provide tools that help in the design of interventions that reliably achieve a desired effect. The third is to engage with the important normative debate about when and to which extent reasoning about discrimination and fairness requires causal understanding. We will also see that causality forces us to grapple with some difficult questions that we have not encountered so far.

The limitations of observation

Before we develop any new formalism, it is important to understand why we need it in the first place.

To see why we turn to the venerable example of graduate admissions at the University of California, Berkeley in 1973.[122] Historical

[122] Bickel et al., "Sex Bias in Graduate Admissions: Data from Berkeley," *Science* 187, no. 4175 (1975): 398–404.

data show that 12763 applicants were considered for admission to one of 101 departments and inter-departmental majors. Of the 4321 women who applied roughly 35 percent were admitted, while 44 percent of the 8442 men who applied were admitted. Standard statistical significance tests suggest that the observed difference would be highly unlikely to be the outcome of sample fluctuation if there were no difference in underlying acceptance rates.

A similar pattern exists if we look at the aggregate admission decisions of the six largest departments. The acceptance rate across all six departments for men is about 44%, while it is only roughly 30% for women, again, a significant difference. Recognizing that departments have autonomy over who to admit, we can look at the gender bias of each department.[123]

[123] Source (Note: There is some discrepancy with a Wikipedia page. Retrieved: Dec 27, 2018.)

Table 7: UC Berkeley admissions data from 1973.

	Men		Women	
Department	Applied	Admitted (%)	Applied	Admitted (%)
A	825	62	108	**82**
B	520	60	25	**68**
C	325	**37**	593	34
D	417	33	375	**35**
E	191	**28**	393	24
F	373	6	341	**7**

What we can see from the table is that four of the six largest departments show a higher acceptance ratio among women, while two show a higher acceptance rate for men. However, these two departments cannot account for the large difference in acceptance rates that we observed in aggregate. So, it appears that the higher acceptance rate for men that we observed in aggregate seems to have reversed at the department level.

Such reversals are sometimes called *Simpson's paradox*[124], even though mathematically they are no surprise. It's a fact of conditional probability that there can be events Y (here, acceptance), A (here, female gender taken to be a binary variable) and a random variable Z (here, department choice) such that:

1. $\mathbb{P}\{Y \mid A\} < \mathbb{P}\{Y \mid \neg A\}$
2. $\mathbb{P}\{Y \mid A, Z = z\} > \mathbb{P}\{Y \mid \neg A, Z = z\}$ for all values z that the random variable Z assumes.

Simpson's paradox nonetheless causes discomfort to some, because intuition suggests that a trend which holds for all subpopulations should also hold at the population level.

[124] For clarifications regarding the popular interpretation of Simpson's original article (Simpson, "The Interpretation of Interaction in Contingency Tables," *Journal of the Royal Statistical Society: Series B (Methodological)* 13, no. 2 (1951): 238–41), see (Hernán, Clayton, and Keiding, "The Simpson's paradox unraveled," *International Journal of Epidemiology* 40, no. 3 (March 2011): 780–85, https://doi.org/10.1093/ije/dyr041) and (Pearl, *Causality* (Cambridge University Press, 2009)).

The reason why Simpson's paradox is relevant to our discussion is that it's a consequence of how we tend to misinterpret what information conditional probabilities encode. Recall that a statement of conditional probability corresponds to passive observation. What we see here is a snapshot of the normal behavior of women and men applying to graduate school at UC Berkeley in 1973.

What is evident from the data is that gender influences department choice. Women and men appear to have different preferences for different fields of study. Moreover, different departments have different admission criteria. Some have lower acceptance rates, some higher. Therefore, one explanation for the data we see is that women *chose* to apply to more competitive departments, hence getting rejected at a higher rate than men.

Indeed, this is the conclusion the original study drew:

> *The bias in the aggregated data stems not from any pattern of discrimination on the part of admissions committees, which seems quite fair on the whole, but apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.*[125]

[125] Bickel et al., "Sex Bias in Graduate Admissions."

In other words, the article concluded that the source of gender bias in admissions was a *pipeline problem*: Without any wrongdoing by the departments, women were "shunted by their socialization" that happened at an earlier stage in their lives.

It is difficult to debate this conclusion on the basis of the available data alone. The question of discrimination, however, is far from resolved.[126] We can ask why women applied to more competitive departments in the first place. There are several possible reasons. Perhaps less competitive departments, such as engineering schools, were unwelcoming of women at the time. This may have been a general pattern at the time or specific to the university. Perhaps some departments had a track record of poor treatment of women that was known to the applicants. Perhaps the department advertised the program in a manner that discouraged women from applying.

[126] The example has been heavily discussed in various other writings, such as Pearl's recent discussion (Pearl and Mackenzie, *The Book of Why: The New Science of Cause and Effect* (Basic Books, 2018)). However, the development throughout this chapter will differ significantly in its arguments and conclusions.

The data we have also shows no measurement of *qualification* of an applicant. It's possible that due to self-selection women applying to engineering schools in 1973 were over-qualified relative to their peers. In this case, an equal acceptance rate between men and women might actually be a sign of discrimination.

There is no way of knowing what was the case from the data we have. We see that at best the original analysis leads to a number of follow-up questions.

What is encoded in the UC Berkeley admissions example is a

variant of the impossibility result we saw in Chapter 2. There are multiple scenarios with fundamentally different interpretations and consequences that we cannot distinguish from the data at hand.

At this point, we have two choices. One is to design a new study and collect more data in a manner that might lead to a more conclusive outcome. The other is to argue over which scenario is more likely based on our beliefs and plausible assumptions about the world.

Causal inference is helpful in either case. On the one hand, it can be used as a guide in the design of new studies. It can help us choose which variables to include, which to exclude, and which to hold constant. On the other hand, causal models can serve as a mechanism to incorporate scientific domain knowledge and exchange plausible assumptions for plausible conclusions.

Causal models

We will develop just enough formal concepts to engage with the technical and normative debate around causality and discrimination. The topic is much deeper than what we can explore in this chapter.

We choose *structural causal models* as the basis of our formal discussion as they have the advantage of giving a sound foundation for various causal notions we will encounter. The easiest way to conceptualize a structural causal model is as a program for generating a distribution from independent noise variables through a sequence of formal instructions. Let's unpack this statement. Imagine instead of samples from a distribution, somebody gave you a step-by-step computer program to generate samples on your own starting from a random seed. The process is not unlike how you would write code. You start from a simple random seed and build up increasingly more complex constructs. That is basically what a structural causal model is, except that each assignment uses the language of mathematics rather than any concrete programming syntax.

A first example

Let's start with a toy example not intended to capture the real world. Imagine a hypothetical population in which an individual exercises regularly with probability $1/2$. With probability $1/3$, the individual has a latent disposition to develop overweight that manifests in the absence of regular exercise. Similarly, in the absence of exercise, heart disease occurs with probability $1/3$. Denote by X the indicator variable of regular exercise, by W that of excessive weight, and by H the indicator of heart disease. Below is a structural causal model to

generate samples from this hypothetical population.

1. Sample independent Bernoulli[127] random variables, i.e., biased coin flips: $U_1 \sim \mathrm{B}(1/2), U_2 \sim \mathrm{B}(1/3), U_3 \sim \mathrm{B}(1/3)$.
2. $X := U_1$
3. $W :=$ if $X = 1$ then 0 else U_2
4. $H :=$ if $X = 1$ then 0 else U_3

[127] A Bernoulli random variable $\mathrm{B}(p)$ with bias p is a biased coin toss that assumes value 1 with probability p and value 0 with probability $1 - p$.

Contrast this generative description of the population with a usual random sample drawn from the population that might look like this:

X	W	H
0	1	1
1	0	0
1	1	1
1	1	0
0	1	0
...	...	...

From the program description, we can immediately see that in our hypothetical population *exercise* averts both *overweight* and *heart disease*, but in the absence of exercise the two are independent. At the outset, our program generates a joint distribution over the random variables (X, W, H). We can calculate probabilities under this distribution. For example, the probability of heart disease under the distribution specified by our model is $1/2 \cdot 1/3 = 1/6$. We can also calculate the conditional probability of heart diseases given overweight. From the event $W = 1$ we can infer that the individual does not exercise so that the probability of heart disease given overweight increases to $1/3$ compared with the baseline of $1/6$.

Does this mean that overweight causes heart disease in our model? The answer is *no* as is intuitive given the program to generate the distribution. But let's see how we would go about arguing this point formally. Having a program to generate a distribution is substantially more powerful than just having sampling access. One reason is that we can manipulate the program in whichever way we want, assuming we still end up with a valid program. We could, for example, set $W := 1$, resulting in a new distribution. The resulting program looks like this:

2. $X := U_1$
3. $W := 1$
4. $H :=$ if $X = 1$ then 0 else U_3

This new program specifies a new distribution. We can again

calculate the probability of heart disease under this new distribution. We still get $1/6$. This simple calculation reveals a significant insight. The substitution $W := 1$ does not correspond to a conditioning on $W = 1$. One is an action, albeit inconsequential in this case. The other is an observation from which we can draw inferences. If we observe that an individual is overweight, we can infer that they have a higher risk of heart disease (in our toy example). However, this does not mean that lowering body weight would avoid heart disease. It wouldn't in our example. The active substitution $W := 1$ in contrast creates a new hypothetical population in which all individuals are overweight with all that it entails in our model.

Let us belabor this point a bit more by considering another hypothetical population, specified by the equations:

2. $W := U_2$
3. $X :=$ if $W = 0$ then 0 else U_1
4. $H :=$ if $X = 1$ then 0 else U_3

In this population exercise habits are driven by body weight. Overweight individuals choose to exercise with some probability, but that's the only reason anyone would exercise. Heart disease develops in the absence of exercise. The substitution $W := 1$ in this model leads to an increased probability of exercise, hence lowering the probability of heart disease. In this case, the conditioning on $W = 1$ has the same affect. Both lead to a probability of $1/6$.

What we see is that fixing a variable by substitution may or may not correspond to a conditional probability. This is a formal rendering of our earlier point that observation isn't action. A substitution corresponds to an action we perform. By substituting a value we break the natural course of action our model captures. This is the reason why the substitution operation is sometimes called the *do-operator*, written as $\mathrm{do}(W := 1)$.

Structural causal models give us a formal calculus to reason about the effect of hypothetical actions. We will see how this creates a formal basis for all the different causal notions that we will encounter in this chapter.

Structural causal models, more formally

Formally, a structural causal model is a sequence of assignments for generating a joint distribution starting from independent noise variables. By executing the sequence of assignments we incrementally build a set of jointly distributed random variables. A structural causal model therefore not only provides a joint distribution, but also a description of how the joint distribution can be generated from ele-

mentary noise variables. The formal definition is a bit cumbersome compared with the intuitive notion.

Definition 4. *A* structural causal model *M is given by a set of variables $X_1, ..., X_d$ and corresponding assignments of the form*

$$X_i := f_i(P_i, U_i), \qquad i = 1, ..., d\,.$$

Here, $P_i \subseteq \{X_1, ..., X_d\}$ is a subset of the variables that we call the parents *of X_i. The random variables $U_1, ..., U_d$ are called* noise variables*, which we require to be jointly independent.*

The directed graph corresponding to the model has one node for each variable X_i, which has incoming edges from all the parents P_i. We will call such a graph the causal graph *corresponding to the structural causal model.*

Let's walk through the formal concepts introduced in this definition in a bit more detail.

The noise variables that appear in the definition model *exogenous factors* that influence the system. Consider, for example, how the weather influences the delay on a traffic route you choose. Due to the difficulty of modeling the influence of weather more precisely, we could take the weather induced delay to be an exogenous factor that enters the model as a noise variable. The choice of exogenous variables and their distribution can have important consequences for what conclusions we draw from a model.

The parent nodes P_i of node i in a structural causal model are often called the *direct causes* of X_i. Similarly, we call X_i the direct effect of its direct causes P_i. Recall our hypothetical population in which weight gain was determined by lack of exercise via the assignment $W := \min\{U_1, 1 - X\}$. Here we would say that exercise (or lack thereof) is a direct cause of weight gain.

Structural causal model are a collection of formal *assumptions* about how certain variables interact. Each assignment specifies a *response function*. We can think of nodes as receiving messages from their parents and acting according to these messages as well as the influence of an exogenous noise variable.

To which extent a structural causal model conforms to reality is a separate and difficult question that we will return to in more detail later. For now, think of a structural causal model as formalizing and exposing a set of assumptions about a data generating process. As such different models can expose different hypothetical scenarios and serve as a basis for discussion. When we make statements about cause and effect in reference to a model, we don't mean to suggest that these relationship necessarily hold in the real world. Whether

they do depends on the scope, purpose, and validity of our model, which may be difficult to substantiate.

It's not hard to show that a structural causal model defines a unique joint distribution over the variables $(X_1, ..., X_d)$ such that $X_i = f_i(P_i, U_i)$. It's convenient to introduce a notion for probabilities under this distribution. When M denotes a structural causal model, we will write the probability of an event E under the entailed joint distribution as $\mathbb{P}_M\{E\}$. To gain familiarity with the notation, let M denote the structural causal model for the hypothetical population in which both weight gain and heart disease are directly caused by an absence of exercise. We calculated earlier that the probability of heart disease in this model is $\mathbb{P}_M\{H\} = 1/6$.

In what follows we will derive from this single definition of a structural causal model all the different notions and terminology that we'll need in this chapter.

Throughout, we restrict our attention to acyclic assignments. Many real-world systems are naturally described as stateful dynamical system with feedback loops. At the end of the chapter, we discuss some of the options for dealing with such closed loop systems. For example, often cycles can be broken up by introducing time dependent variables, such as, investments at time 0 grow the economy at time 1 which in turn grows investments at time 2, continuing so forth until some chosen time horizon t.

Causal graphs

We saw how structural causal models naturally give rise to *causal graphs* that represent the assignment structure of the model graphically. We can go the other way as well by simply looking at directed graphs as placeholders for an unspecified structural causal model which has the assignment structure given by the graph. Causal graphs are often called *causal diagrams*. We'll use these terms interchangeably.

Below we see causal graphs for the two hypothetical populations from our heart disease example.

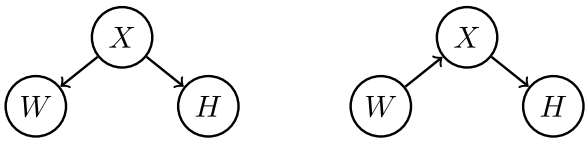


Figure 18: Causal diagrams for the heart disease examples.

The scenarios differ in the direction of the link between exercise and weight gain.

Causal graphs are convenient when the exact assignments in a

structural causal models are of secondary importance, but what matters are the paths present and absent in the graph. Graphs also let us import the established language of graph theory to discuss causal notions. We can say, for example, that an *indirect cause* of a node is any ancestor of the node in a given causal graph. In particular, causal graphs allow us to distinguish cause and effect based on whether a node is an ancestor or descendant of another node.

Let's take a first glimpse at a few important graph structures.

Forks

A *fork* is a node Z in a graph that has outgoing edges to two other variables X and Y. Put differently, the node Z is a common cause of X and Y.

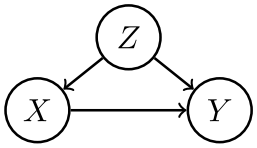


Figure 19: Example of a fork.

We already saw an example of a fork in our weight and exercise example: $W \leftarrow X \rightarrow H$. Here, exercise X influences both weight and heart disease. We also learned from the example that Z has a *confounding* effect: Ignoring exercise X, we saw that W and H appear to be positively correlated. However, the correlation is a mere result of confounding. Once we hold exercise levels constant (via the do-operation), weight has no effect on heart disease in our example.

Confounding leads to a disagreement between the calculus of conditional probabilities (observation) and do-interventions (actions).

Real-world examples of confounding are a common threat to the validity of conclusions drawn from data. For example, in a well known medical study a suspected beneficial effect of *hormone replacement therapy* in reducing cardiovascular disease disappeared after identifying *socioeconomic status* as a confounding variable.[128]

[128] Humphrey, Chan, and Sox, "Postmenopausal Hormone Replacement Therapy and the Primary Prevention of Cardiovascular Disease," *Annals of Internal Medicine* 137, no. 4 (August 2002): 273–84.

Mediators

The case of a fork is quite different from the situation where Z lies on a directed path from X to Y:

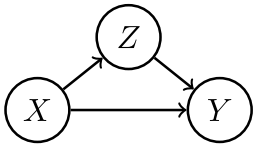


Figure 20: Example of a chain.

In this case, the path $X \to Z \to Y$ contributes to the total effect of X on Y. It's a causal path and thus one of the ways in which X causally influences Y. That's why Z is not a confounder. We call Z a *mediator* instead.

We saw a plausible example of a mediator in our UC Berkeley admissions example. In one plausible causal graph, department choice mediates the influences of gender on the admissions decision.

The notion of a mediator is particularly relevant to the topic of discrimination analysis and we will return to this discussion in more detail again.

Colliders

Finally, let's consider another common situation: the case of a *collider*.

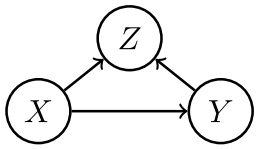


Figure 21: Example of a collider.

Colliders aren't confounders. In fact, in the above graph, X and Y are unconfounded, meaning that we can replace do-statements by conditional probabilities. However, something interesting happens when we condition on a collider. The conditioning step can create correlation between X and Y, a phenomenon called *explaining away*. A good example of the explaining away effect, or *collider bias*, is due to Berkson. Two independent diseases can become negatively correlated when analyzing hospitalized patients. The reason is that when either disease (X or Y) is sufficient for admission to the hospital (indicated by variable Z), observing that a patient has one disease makes the other statistically less likely.[129]

Berkson's law is a cautionary tale for statistical analysis when we're studying a cohort that has been subjected to a selection rule. For example, there's an ongoing debate about the effectiveness of GRE scores in higher education. Recent studies[130] argue that GRE scores are not predictive of various success outcomes in a graduate student population. However, care must be taken when studying the effectiveness of educational tests, such as the GRE, by examining a sample of admitted students. After all, students were in part admitted on the basis of the test score. It's the selection rule that introduces the potential for collider bias.

[129] See the Wikipedia article and the reprint of Berkson's original article (Berkson, "Limitations of the Application of Fourfold Table Analysis to Hospital Data," *International Journal of Epidemiology* 43, no. 2 (2014): 511–15).

[130] Moneta-Koehler, "The Limitations of the GRE in Predicting Success in Biomedical Graduate School," *PLOS ONE* 12, no. 1 (January 2017): 1–17; Hall, "Predictors of Student Productivity in Biomedical Graduate School Applications," *PLOS ONE* 12, no. 1 (January 2017): 1–14.

Interventions and causal effects

Structural causal models give us a way to formalize the effect of hypothetical actions or interventions on the population within the assumptions of our model. As we saw earlier all we needed was the ability to do substitutions.

Substitutions and the do-operator

Given a structural causal model M we can take any assignment of the form

$$X := f(P, U)$$

and replace it by another assignment. The most common substitution is to assign X a constant value x:

$$X := x$$

We will denote the resulting model by $M' = M[X := x]$ to indicate the surgery we performed on the original model M. Under this assignment we hold X constant by removing the influence of its parent nodes and thereby any other variables in the model.

Graphically, the operation corresponds to eliminating all incoming edges to the node X. The children of X in the graph now receive a fixed message x from X when they query the node's value.

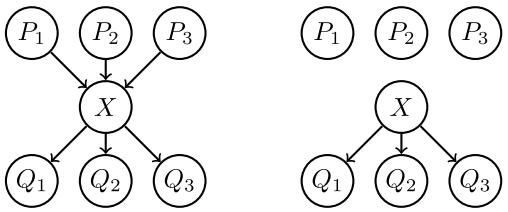


Figure 22: Graph before and after substitution.

The assignment operator is also called the *do-operator* to emphasize that it corresponds to performing an action or intervention. We already have notation to compute probabilities after applying the do-operator, namely, $\mathbb{P}_{M[X:=x]}(E)$.

Another notation is popular and common:

$$\mathbb{P}\{E \mid \text{do}(X := x)\} = \mathbb{P}_{M[X:=x]}(E)$$

This notation analogizes the do-operation with the usual notation for conditional probabilities, and is often convenient when doing calculations involving the do-operator. Keep in mind, however, that the do-operator (action) is fundamentally different from the conditioning operator (observation).

Causal effects

The *causal effect* of an action $X := x$ on a variable Y refers to the distribution of the variable Y in the model $M[X := x]$. When we speak of the causal effect of a variable X on another variable Y we refer to all the ways in which setting X to any possible value x affects the distribution of Y.

Often we think of X as a binary treatment variable and are interested in a quantity such as

$$\mathbb{E}_{M[X:=1]}[Y] - \mathbb{E}_{M[X:=0]}[Y] \,.$$

This quantity is called the *average treatment effect*. It tells us how much treatment (action $X := 1$) increases the expectation of Y relative to no treatment (action $X := 0$).

Causal effects are population quantities. They refer to effects averaged over the whole population. Often the effect of treatment varies greatly from one individual or group of individuals to another. Such treatment effects are called *heterogeneous*.

Confounding

Important questions in causality relate to when we can rewrite a do-operation in terms of conditional probabilities. When this is possible, we can estimate the effect of the do-operation from conventional conditional probabilities that we can estimate from data.

The simplest question of this kind asks when a causal effect $\mathbb{P}\{Y = y \mid \mathrm{do}(X := x)\}$ coincides with the condition probability $\mathbb{P}\{Y = y \mid X = x\}$. In general, this is not true. After all, the difference between observation (conditional probability) and action (interventional calculus) is what motivated the development of causality.

The disagreement between interventional statements and conditional statements is so important that it has a well-known name: *confounding*. We say that X and Y are confounded when the causal effect of action $X := x$ on Y does not coincide with the corresponding conditional probability.

When X and Y are confounded, we can ask if there is some combination of conditional probability statements that give us the desired effect of a do-intervention. This is generally possible given a causal graph by conditioning on the parent nodes PA of the node X:

$$\mathbb{P}\{Y = y \mid \mathrm{do}(X := x)\} = \sum_z \mathbb{P}\{Y = y \mid X = x, PA = z\}\mathbb{P}\{PA = z\}$$

This formula is called the *adjustment formula*. It gives us one way of estimating the effect of a do-intervention in terms of conditional probabilities.

The adjustment formula is one example of what is often called *controlling for* a set of variables: We estimate the effect of X on Y separately in every slice of the population defined by a condition $Z = z$ for every possible value of z. We then average these estimated sub-population effects weighted by the probability of $Z = z$ in the population. To give an example, when we control for age, we mean that we estimate an effect separately in each possible age group and then average out the results so that each age group is weighted by the fraction of the population that falls into the age group.

Controlling for more variables in a study isn't always the right choice. It depends on the graph structure. Let's consider what happens when we control for the variable Z in the three causal graphs we discussed above.

- Controlling for a confounding variable Z in a fork $X \leftarrow Z \rightarrow Y$ will deconfound the effect of X on Y.
- Controlling for a mediator Z on a chain $X \rightarrow Z \rightarrow Y$ will eliminate some of the causal influence of X on Y.
- Controlling for a collider will create correlation between X and Y. That is the opposite of what controlling for Z accomplishes in the case of a fork. The same is true if we control for a descendant of a collider.

The backdoor criterion

At this point, we might worry that things get increasingly complicated. As we introduce more nodes in our graph, we might fear a combinatorial explosion of possible scenarios to discuss. Fortunately, there are simple sufficient criteria for choosing a set of deconfounding variables that is safe to control for.

A well known graph-theoretic notion is the *backdoor* criterion.[131] Two variables are confounded if there is a so-called *backdoor* path between them. A *backdoor path* from X to Y is any path starting at X with a backward edge "$\leftarrow$" into X such as:

$$X \leftarrow A \rightarrow B \leftarrow C \rightarrow Y$$

Intuitively, backdoor paths allow information flow from X to Y in a way that is not causal. To deconfound a pair of variables we need to select a *backdoor set* of variables that "blocks" all backdoor paths between the two nodes. A backdoor path involving a chain $A \rightarrow B \rightarrow C$ can be blocked by controlling for B. Information by

[131] Pearl, *Causality*.

default cannot flow through a collider $A \rightarrow B \leftarrow C$. So we only have to be careful not to open information flow through a collider by conditioning on the collider, or descendant of a collider.[132]

[132] For additional discussion of backdoor paths and confounding, see (Pearl).

Unobserved confounding

The adjustment formula might suggest that we can always eliminate confounding bias by conditioning on the parent nodes. However, this is only true in the absence of *unobserved confounding*. In practice often there are variables that are hard to measure, or were simply left unrecorded. We can still include such unobserved nodes in a graph, typically denoting their influence with dashed lines, instead of solid lines.

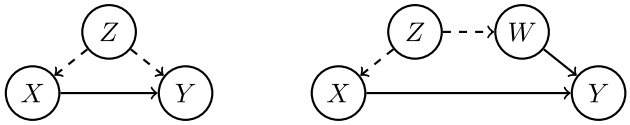


Figure 23: Two cases of unobserved confounding.

The above figure shows two cases of unobserved confounding. In the first example, the causal effect of X on Y is unidentifiable. In the second case, we can block the confounding backdoor path $X \leftarrow Z \rightarrow W \rightarrow Y$ by controlling for W even though Z is not observed. The backdoor criterion lets us work around unobserved confounders in some cases where the adjustment formula alone wouldn't suffice.

Unobserved confounding nonetheless remains a major obstacle in practice. The issue is not just lack of measurement, but often lack of anticipation or awareness of a counfounding variable. We can try to combat unobserved confounding by increasing the number of variables under consideration. But as we introduce more variables into our study, we also increase the burden of coming up with a valid causal model for all variables under consideration. In practice, it is not uncommon to control for as many variables as possible in a hope to disable confounding bias. However, as we saw, controlling for mediators or colliders can be harmful.

Randomization

The backdoor criterion gives a non-experimental way of eliminating confounding bias given a causal model and a sufficient amount of observational data from the joint distribution of the variables. An alternative experimental method of eliminating confounding bias is the well-known *randomized controlled trial*.

In a *randomized controlled trial* a group of subjects is randomly partitioned into a *control group* and a *treatment group*. Participants

do not know which group they were assigned to and neither do the staff administering the trial. The treatment group receives an actual treatment, such as a drug that is being tested for efficacy, while the control group receives a placebo identical in appearance. An outcome variable is measured for all subjects.

The goal of a randomized controlled trial is to break natural inclination. Rather than observing who chose to be treated on their own, we assign treatment randomly. Thinking in terms of causal models, what this means is that we eliminate all incoming edges into the treatment variable. In particular, this closes all backdoor paths and hence avoids confounding bias.

There are many reasons why often randomized controlled trials are difficult or impossible to administer. Treatment might be physically or legally impossible, too costly, or too dangerous. As we saw, randomized controlled trials are not always necessary for avoiding confounding bias and for reasoning about cause and effect. Nor are they free of issues and pitfalls.[133]

[133] Deaton and Cartwright, "Understanding and Misunderstanding Randomized Controlled Trials," *Social Science & Medicine* 210 (2018): 2–21.

Graphical discrimination analysis

We now explore how we can bring causal graphs to bear on discussions of discrimination. We return to the example of graduate admissions at Berkeley and develop a causal perspective on the earlier analysis.

The first step is to come up with a plausible causal graph consistent with the data that we saw earlier. The data contained only three variables, sex A, department choice Z, and admission decision Y. It makes sense to draw two arrows $A \to Y$ and $Z \to Y$, because both features A and Z are available to the institution when making the admissions decision.

We'll draw one more arrow, for now, simply because we have to. If we only included the two arrows $A \to Y$ and $Z \to Y$, our graph would claim that A and Z are statistically independent. However, this claim is inconsistent with the data.[134]

This means we need to include either the arrow $A \to Z$ or $Z \to A$.[135] Deciding between the two isn't as straightforward as it might first appear.

If we interpreted A in the narrowest possible sense as the applicant's *reported sex*, i.e., literally which box they checked on the application form, we could imagine a scenario where some applicants choose to (mis-)report their sex in a certain way that depends in part on their department choice. Even if we assume no misreporting occurs, it's hard to substantiate *reported sex* as a plausible cause of department choice. The fact that an applicant checked a box labeled

[134] We can see from the table that several departments have a statistically significant gender bias among applicants.

[135] There is also the possibility of an unobserved confounder that we will discuss later.

male certainly isn't the cause for their interest in engineering.

The causal story in Bickel's argument is a different one. It alludes to a complex socialization and preference formation process that took place in the applicant's life before they applied which in part depended on the applicant's sex. To align this story with our causal graph, we need the variable A to reference whatever ontological entity it is that through this "socialization process" influences intellectual and professional preferences, and hence, department choice.

It is difficult to maintain that this ontological entity coincides with sex as a biological trait. There is no scientific basis to support that the biological trait *sex* is what determines our intellectual preferences. Few scholars (if any) would currently attempt to maintain a claim such as *two X chromosomes cause an interest in English literature*.

The truth is that we don't know the exact mechanism by which the thing referenced by A influences department choice. In drawing the arrow A to Z we assert—perhaps with some naivety or ignorance—that there exists such a mechanism.

We will discuss the important difficulty we encountered here in depth later on. For now, we commit to this modeling choice and thus arrive at the following graph.

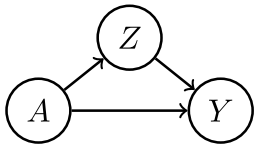


Figure 24: Possible causal graph for the UC Berkeley graduate admissions scenario.

In this graph, department choice mediates the influence of gender on admissions. There's a direct path from A to Y and an indirect path that goes through Z.

We will use this model to put pressure on the central claim in the original study, namely, that *there is no evidence of sex discrimination*.

In causal language, Bickel's argument had two components[136]:

1. There appears to be no direct effect of sex A on the admissions decision Y that favors men.
2. The indirect effect of A on Y that is mediated by department choice should not be counted as evidence of discrimination.

We will discuss both arguments in turn.

[136] In fact, this is Pearl's proffered causal interpretation of Bickel's analysis. See (Pearl, *Causality*; Pearl and Mackenzie, *The Book of Why*).

Direct effects

To obtain the direct effect of A on Y we need to disable all paths between A and Y except for the direct link. In our model, we can accomplish this by holding department choice Z constant and evaluating the conditional distribution of Y given A. Recall that holding

a variable constant is generally not the same as conditioning on the variable.

Specifically, a problem would arise if department choice and admissions outcome were confounded by another variable, such as, state of residence R

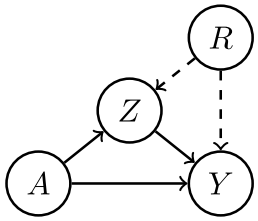


Figure 25: Alternative causal graph for the UC Berkeley graduate admissions scenario showing influence of residence.

Department choice is now a collider between A and R. Conditioning on a collider opens the backdoor path $A \rightarrow Z \leftarrow R \rightarrow Y$. In this graph, conditioning on department choice does *not* give us the desired direct effect.[137]

If we assume, however, that department choice and admissions decisions are unconfounded, then Bickel's approach indeed supports the first claim.

Unfortunately, the direct effect of a protected variable on a decision is a poor measure of discrimination on its own. At a technical level, it is rather brittle as it cannot detect any form of *proxy discrimination*. The department could, for example, use the applicant's personal statement to make inferences about their gender, which are then used to discriminate.

It's best to think of the direct effect as whether or not the decision maker explicitly *uses* the attribute in its decision rule. The absence of a direct effect corresponds to the somewhat troubled notion of a *blind* decision rule that doesn't have explicit access to the sensitive attribute. As we argued in all preceding chapters, blind decision rules can still be the basis of discriminatory practices.

As we saw in the previous chapter, direct effects don't cleanly map onto a legal framework. However, it's possible to see semblance between what a direct effect captures and what kind of discrimination the legal doctrine of *disparate treatment* describes.

[137] The possibility of a confounder between department choice and decision was the subject of an exchange between Bickel and Kruskal, as Pearl discusses in (Pearl and Mackenzie, *The Book of Why*).

Indirect paths

Let's turn to the indirect effect of sex on admission that goes through department choice.

It's tempting to think of the the node Z as referencing the applicant's inherent department preferences that stem from a process of *socialization* alluded to earlier. In this view, the department is not responsible for the applicant's preferences and so the mediating

influence of department preferences is not interpreted as a sign of discrimination. This, however, is a normative judgment that does not follow as formal matter.

We can easily think of natural alternative scenarios that are consistent with both the data and our causal model, in which the indirect path encodes a pattern of discrimination.

For example, the department may have advertised the program in a manner that strongly discouraged women from applying. In this case, department preference in part measures exposure to this hostile advertising campaign. Or, the department could have a track record of hostile behavior against women and it is awareness of such that shapes preferences in an applicant. Finally, blatant discriminatory practices, such as compensating women at a lower rate than equally qualified male graduate students can obviously shape an applicant's preference.

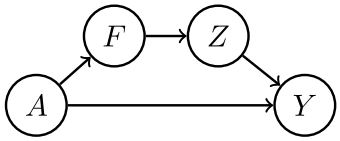


Figure 26: Alternative causal graph for the UC Berkeley graduate admissions scenario where department preferences are shaped by fear of discrimination.

Accepting the indirect path as *non-discriminatory* is to assert that all these scenarios we described are deemed implausible. Fundamentally, we are confronted with a normative question. The path $A \to Z \to Y$ could either be where discrimination occurs or what explains the absence thereof. Which case we're in isn't a purely technical matter and cannot be resolved without subject matter knowledge. Causal modeling gives us a framework for exposing these normative questions, but not necessarily one to resolve them.

Path inspection

To summarize, discrimination may not only occur on the direct pathway from the sensitive category to the outcome. Seemingly innocuous mediating paths can hide discriminatory practices. We have to carefully discuss what pathways we consider evidence for or against discrimination.

To appreciate this point, contrast our Berkeley scenario with the important legal case *Griggs v. Duke Power Co.*[138] that was argued before the U.S. Supreme Court in 1970. Duke Power Company had introduced the requirement of a high school diploma for certain higher paying jobs. We could draw a causal graph for this scenario not unlike the one for the Berkeley case. There's a mediating variable (here, level of education), a sensitive category (here, race) and an employment outcome (here, employment in a higher paying job).

[138] Griggs v. Duke Power Co., 401 U.S. 424 (1971)

The company didn't directly make employment decisions based on race, but rather used the mediating variable. The court ruled that the requirement of a high school diploma was not justified by business necessity, but rather had adverse impact on ethnic minority groups where the prevalence of high school diplomas is lower. Put differently, the court decided that the use of this mediating variable was not an argument against, but rather for discrimination.

Glymour[139] makes another related and important point about the moral character of mediation analysis:

> Implicitly, the question of what mediates observed social effects informs our view of which types of inequalities are socially acceptable and which types require remediation by social policies. For example, a conclusion that women are "biologically programmed" to be depressed more than men may ameliorate the social obligation to try to reduce gender inequalities in depression. Yet if people get depressed whenever they are, say, sexually harassed—and women are more frequently sexually harassed than men—this suggests a very strong social obligation to reduce the depression disparity by reducing the sexual harassment disparity.

[139] Glymour, "Using Causal Diagrams to Understand Common Problems in Social Epidemiology," *Methods in Social Epidemiology*, 2006, 393–428.

Ending on a technical note, it's worth noting that we currently do not have a method to estimate indirect effects. Estimating an indirect effect somehow requires us to *disable* the direct influence. There is no way of doing this with the do-operation that we've seen so far. However, we will shortly introduce *counterfactuals*, which among other applications will give us a way of estimating path-specific effects.

Structural discrimination

There's an additional problem we neglected so far. Imagine a spiteful university administration that systematically defunds graduate programs that attract more female applicants. This structural pattern of discrimination is invisible from the causal model we drew. There is a kind of type mismatch here. Our model talks about individual applicants, their department preferences, and their outcomes. Put differently, individuals are the *units* of our investigation. University policy is not one of the mechanisms that our model exposes. As a result we cannot talk about university policy as a cause of discrimination in our model.

The model we chose commits us to an individualistic perspective that frames discrimination as the consequence of how decision makers respond to information about individuals.

An analogy is helpful. In epidemiology, scientists can seek the cause of health outcomes in biomedical aspects and lifestyle choices

of individuals, such as whether or not an individual smokes, exercises, maintains a balanced diet etc. The growing field of social epidemiology criticizes the view of individual choices as causes of health outcomes, and instead draws attention to social and structural causes,[140] such as poverty and inequality.

[140] Krieger, "Epidemiology and the People's Health: Theory and Context," 2011.

Similarly, we can contrast the individualistic perspective on discrimination with structural discrimination. Causal modeling can in principle be used to study the causes of structural discrimination, as well. But it requires a different perspective than the one we chose for our Berkeley scenario.

Counterfactuals

Fully specified structural causal models allow us to ask causal questions that are more delicate than the mere effect of an action. Specifically, we can ask *counterfactual* questions such as: Would I have avoided the traffic jam had I taken a different route this morning? Counterfactual questions are common. We can answer them given a structural causal model. However, the procedure for extracting the answer from the model looks a bit subtle at first. It helps to start with a simple example.

A simple counterfactual

To understand counterfactuals, we first need to convince ourselves that they aren't quite as straightforward as a single substitution in our model.

Assume every morning we need to decide between two routes $X = 0$ and $X = 1$. On bad traffic days, indicated by $U = 1$, both routes are bad. On good days, indicated by $U = 0$, the traffic on either route is good unless there was an accident on the route.

Let's say that $U \sim B(1/2)$ follows the distribution of an unbiased coin toss. Accidents occur independently on either route with probability $1/2$. So, choose two Bernoulli random variables $U_0, U_1 \sim B(1/2)$ that tell us if there is an accident on route 0 and route 1, respectively.

We reject all external route guidance and instead decide on which route to take uniformly at random. That is, $X := U_X \sim B(1/2)$ is also an unbiased coin toss.

Introduce a variable $Y \in \{0, 1\}$ that tells us whether the traffic on the chosen route is good ($Y = 0$) or bad ($Y = 1$). Reflecting our discussion above, we can express Y as

$$Y := X \cdot \max\{U, U_1\} + (1 - X) \max\{U, U_0\} .$$

In words, when $X = 0$ the first term disappears and so traffic is determined by the larger of the two values U and U_0. Similarly, when $X = 1$ traffic is determined by the larger of U and U_1.

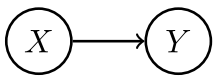


Figure 27: Causal diagram for our traffic scenario.

Now, suppose one morning we have $X = 1$ and we observe bad traffic $Y = 1$. Would we have been better off taking the alternative route this morning?

A natural attempt to answer this question is to compute the likelihood of $Y = 0$ after the do-operation $X := 0$, that is, $\mathbb{P}_{M[X:=0]}(Y = 0)$. A quick calculation reveals that this probability is $\frac{1}{2} \cdot \frac{1}{2} = 1/4$. Indeed, given the substitution $X := 0$ in our model, for the traffic to be good we need that $\max\{U, U_0\} = 0$. This can only happen when both $U = 0$ (probability $1/2$) and $U_0 = 0$ (probability $1/2$).

But this isn't the correct answer to our question. The reason is that we took route $X = 1$ and observed that $Y = 1$. From this observation, we can deduce that certain background conditions did not manifest for they are inconsistent with the observed outcome. Formally, this means that certain settings of the noise variables (U, U_0, U_1) are no longer feasible given the observed event $\{Y = 1, X = 1\}$. Specifically, if U and U_1 had both been zero, we would have seen no bad traffic on route $X = 1$, but this is contrary to our observation. In fact, the available evidence $\{Y = 1, X = 1\}$ leaves only the following settings for U and U_1:[141]

[141] We leave out U_0 from the table, since its distribution is unaffected by our observation.

Table 9: Possible noise settings after observing evidence

U	U_1
0	1
1	1
1	0

Each of these three cases is equally likely, which in particular means that the event $U = 1$ now has probability $2/3$. In the absence of any additional evidence, recall, $U = 1$ had probability $1/2$. What this means is that the observed evidence $\{Y = 1, X = 1\}$ has biased the distribution of the noise variable U toward 1. Let's use the letter U' to refer to this biased version of U.[142]

[142] Formally, U' is distributed according to the distribution of U conditional on the event $\{Y = 1, X = 1\}$.

Working with this biased noise variable, we can again entertain the effect of the action $X := 0$ on the outcome Y. For $Y = 0$ we need that $\max\{U', U_0\} = 0$. This means that $U' = 0$, an event that now has probability $1/3$, and $U_0 = 0$ (probability $1/2$ as before). Hence, we

get the probability $1/6 = 1/2 \cdot 1/3$ for the event that $Y = 0$ under our do-operation $X := 0$, and after updating the noise variables to account for the observation $\{Y = 1, X = 1\}$.

To summarize, incorporating available evidence into our calculation decreased the probability of no traffic ($Y = 0$) when choosing route 0 from $1/4$ to $1/6$. The intuitive reason is that the evidence made it more likely that it was generally a bad traffic day, and even the alternative route would've been clogged. More formally, the event that we observed biases the distribution of exogenous noise variables.

We think of the result we just calculated as the *counterfactual* of choosing the alternative route given the route we chose had bad traffic.

The general recipe

We can generalize our discussion of computing counterfactuals from the previous example to a general procedure. There were three essential steps. First, we incorporated available observational evidence by biasing the exogenous noise variables through a conditioning operation. Second, we performed a do-operation in the structural causal model after we substituted the biased noise variables. Third, we computed the distribution of a target variable.

These three steps are typically called *abduction*, *action*, and *prediction*, as can be described as follows.

Definition 5. *Given a structural causal model M, an observed event E, an action $X := x$ and target variable Y, we define the* counterfactual *$Y_{X:=x}(E)$ by the following three step procedure:*

1. ***Abduction:*** *Adjust noise variables to be consistent with the observed event. Formally, condition the joint distribution of $U = (U_1, ..., U_d)$ on the event E. This results in a biased distribution U'.*
2. ***Action:*** *Perform do-intervention $X := x$ in the structural causal model M resulting in the model $M' = M[X := x]$.*
3. ***Prediction:*** *Compute target counterfactual $Y_{X:=x}(E)$ by using U' as the random seed in M'.*

It's important to realize that this procedure *defines* what a counterfactual is in a structural causal model. The notation $Y_{X:=x}(E)$ denotes the outcome of the procedure and is part of the definition. We haven't encountered this notation before.

Put in words, we interpret the formal counterfactual $Y_{X:=x}(E)$ as the value Y would've taken had the variable X been set to value x in the circumstances described by the event E.

In general, the counterfactual $Y_{X:=x}(E)$ is a random variable that varies with U'. But counterfactuals can also be deterministic. When

the event E narrows down the distribution of U to a single point mass, called *unit*, the variable U' is constant and hence the counterfactual $Y_{X:=x}(E)$ reduces to a single number. In this case, it's common to use the shorthand notation $Y_x(u) = Y_{X:=x}(\{U = u\})$, where we make the variable X implicit, and let u refer to a single unit.

The motivation for the name *unit* derives from the common situation where the structural causal model describes a population of entities that form the atomic units of our study. It's common for a unit to be an individual (or the description of a single individual). However, depending on application, the choice of units can vary. In our traffic example, the noise variables dictate which route we take and what the road conditions are.

Answers to counterfactual questions strongly depend on the specifics of the structural causal model, including the precise model of how the exogenous noise variables come into play. It's possible to construct two models that have identical graph structures, and behave identically under interventions, yet give different answers to counterfactual queries.[143]

[143] Peters, Janzing, and Schölkopf, *Elements of Causal Inference* (MIT Press, 2017).

Potential outcomes

The *potential outcomes* framework is a popular formal basis for causal inference, which goes about counterfactuals differently. Rather than deriving them from a structural causal model, we assume their existence as ordinary random variables, albeit some unobserved.

Specifically, we assume that for every unit u there exist random variables $Y_x(u)$ for every possible value of the assignment x. In the potential outcomes model, it's customary to think of a binary *treatment* variable X so that x assumes only two values, 0 for *untreated*, and 1 for *treated*. This gives us two potential outcome variables $Y_0(u)$ and $Y_1(u)$ for each unit u.[144]

[144] There is some potential for notational confusion here. Readers familiar with the potential outcomes model may be used to the notation "$Y_i(0), Y_i(1)$" for the two potential outcomes corresponding to unit i. In our notation the unit (or, more generally, set of units) appears in the parentheses and the subscript denotes the substituted value for the variable we intervene on.

The key point about the potential outcomes model is that we only observe the potential outcome $Y_1(u)$ for units that were treated. For untreated units we observe $Y_0(u)$. In other words, we can never simultaneously observe both, although they're both assumed to exist in a formal sense. Formally, the outcome $Y(u)$ for unit u that we observe depends on the binary treatment $T(u)$ and is given by the expression:

$$Y(u) = Y_0(u) \cdot (1 - T(u)) + Y_1(u) \cdot T(u)$$

It's often convenient to omit the parentheses from our notation for counterfactuals so that this expression would read $Y = Y_0 \cdot (1 - T) + Y_1 \cdot T$.

We can revisit our traffic example in this framework. The next table summarizes what information is observable in the potential outcomes model. We think of the route we choose as the treatment variable, and the observed traffic as reflecting one of the two potential outcomes.

Table 10: Traffic example in the potential outcomes model

Route X	Outcome Y_0	Outcome Y_1	Probability
0	0	?	1/8
0	1	?	3/8
1	?	0	1/8
1	?	1	3/8

Often this information comes in the form of samples. For example, we might observe the traffic on different days. With sufficiently many samples, we can estimate the above frequencies with arbitrary accuracy.

Table 11: Traffic data in the potential outcomes model

Day	Route X	Outcome Y_0	Outcome Y_1
1	0	1	?
2	0	0	?
3	1	?	1
4	0	1	?
5	1	?	0
...	...	...	...

A typical query in the potential outcomes model is the *average treatment effect* $\mathbb{E}[Y_1 - Y_0]$. Here the expectation is taken over the properly weighted units in our study. If units correspond to equally weighted individuals, the expectation is an average over these individuals.

In our original traffic example, there were 16 units corresponding to the background conditions given by the four binary variables U, U_0, U_1, U_X. When the units in the potential outcome model agree with those of a structural causal model, then causal effects computed in the potential outcomes model agree with those computed in the structural equation model. The two formal frameworks are perfectly consistent with each other.

As is intuitive from the table above, causal inference in the potential outcomes framework can be thought of as filling in the missing

entries ("?") in the table above. This is sometimes called *missing data imputation* and there are numerous statistical methods for this task. If we could *reveal* what's behind the question marks, estimating the average treatment effect would be as easy as counting rows.

There is a set of established conditions under which causal inference becomes possible:

1. **Stable Unit Treatment Value Assumption** (SUTVA): The treatment that one unit receives does not change the effect of treatment for any other unit.
2. **Consistency**: Formally, $Y = Y_0(1 - T) + Y_1 T$. That is, $Y = Y_0$ if $T = 0$ and $Y = Y_1$ if $T = 1$. In words, the outcome Y agrees with the potential outcome corresponding to the treatment indicator.
3. **Ignorability**: The potential outcomes are independent of treatment given some deconfounding variables Z, i.e., $T \perp (Y_0, Y_1) \mid Z$. In words, the potential outcomes are conditionally independent of treatment given some set of deconfounding variables.

The first two assumptions automatically hold for counterfactual variables derived from structural causal models according to the procedure described above. This assumes that the units in the potential outcomes framework correspond to the atomic values of the background variables in the structural causal model.

The third assumption is a major one. It's easiest to think of it as aiming to formalize the guarantees of a perfectly executed randomized controlled trial. The assumption on its own cannot be verified or falsified, since we never have access to samples with both potential outcomes manifested. However, we can verify if the assumption is consistent with a given structural causal model by checking if the set Z blocks all backdoor paths from treatment T to outcome Y.

There's no tension between structural causal models and potential outcomes and there's no harm in having familiarity with both. It nonetheless makes sense to say a few words about the differences of the two approaches.

We can derive potential outcomes from a structural causal model as we did above, but we cannot derive a structural causal model from potential outcomes alone. A structural causal model in general encodes more assumptions about the relationships of the variables. This has several consequences. On the one hand, a structural causal model gives us a broader set of formal concepts (causal graphs, mediating paths, counterfactuals for every variable, and so on). On the other hand, coming up with a plausibly valid structural causal model is often a daunting task that might require knowledge that is simply not available. We will dive deeper into questions of validity below. Difficulty to come up with a plausible causal model often

exposes unsettled substantive questions that require resolution first.

The potential outcomes model, in contrast, is generally easier to apply. There's a broad set of statistical estimators of causal effects that can be readily applied to observational data. But the ease of application can also lead to abuse. The assumptions underpinning the validity of such estimators are experimentally unverifiable. Frivolous application of causal effect estimators in situations where crucial assumptions do not hold can lead to false results, and consequently to ineffective or harmful interventions.

Counterfactual discrimination analysis

Counterfactuals serve at least two purposes for us. On the technical side, counterfactuals give us a way to compute path-specific causal effects. This allows us to make path analysis a quantitative matter. On the conceptual side, counterfactuals let us engage with the important normative debate about whether discrimination can be captured by counterfactual criteria. We will discuss each of these in turn.

Quantitative path analysis

Mediation analysis is a venerable subject dating back decades.[145] Generally speaking, the goal of mediation analysis is to identify a mechanism through which a cause has an effect. We will review some recent developments and how they relate to questions of discrimination.

[145] Baron and Kenny, "The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51, no. 6 (1986): 1173.

In the language of our formal framework, mediation analysis aims to decompose a total causal effect into path-specific components. We will illustrate the concepts in the basic three variable case of a mediator, although the ideas extend to more complicated structures.

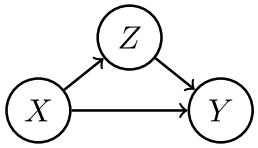


Figure 28: Causal graph with mediator Z.

There are two different paths from X to Y. A direct path and a path through the mediator Z. The conditional expectation $\mathbb{E}[Y \mid X = x]$ lumps together influence from both paths. If there were another confounding variable in our graph influencing both X and Y, then the conditional expectation would also include whatever correlation is the result of confounding. We can eliminate the confounding path by virtue of the do-operator $\mathbb{E}[Y \mid \mathrm{do}(X := x)]$. This gives us the total effect of the action $X := x$ on Y. But the total effect still conflates the two causal pathways, the direct effect and the indirect effect.

We will now see how we can identify the direct and indirect effects separately.

The direct effect we already dealt with earlier as it did not require any counterfactuals. Recall, we can hold the mediator fixed at level $Z := z$ and consider the effect of treatment $X := 1$ compared with no treatment $X := 0$ as follows:

$$\mathbb{E}\left[Y \mid \text{do}(X := 1, Z := z)\right] - \mathbb{E}\left[Y \mid \text{do}(X := 0, Z := z)\right] .$$

We can rewrite this expression in terms of counterfactuals equivalently as:

$$\mathbb{E}\left[Y_{X:=1,Z:=z} - Y_{X:=0,Z:=z}\right] .$$

To be clear, the expectation is taken over the background variables in our structural causal models.[146]

[146] In other words, the counterfactuals inside the expectation are invoked with an elementary setting u of the background variables, i.e., $Y_{X:=1,Z:=z}(u) - Y_{X:=0,Z:=x}(u)$ and the expectation averages over all possible settings.

The formula for the direct effect above is usually called *controlled direct effect*, since it requires setting the mediating variable to a specified level. Sometimes it is desirable to allow the mediating variable to vary as it would had no treatment occurred. This too is possible with counterfactuals and it leads to a notion called *natural direct effect*, defined as:

$$\mathbb{E}\left[Y_{X:=1,Z:=Z_{X:=0}} - Y_{X:=0,Z:=Z_{X:=0}}\right] .$$

The counterfactual $Y_{X:=1,Z:=Z_{X:=0}}$ is the value that Y would obtain had X been set to 1 and had Z been set to the value Z would've assumed had X been set to 0.

The advantage of this slightly mind-bending construction is that it gives us an analogous notion of *natural indirect effect*:

$$\mathbb{E}\left[Y_{X:=0,Z:=Z_{X:=1}} - Y_{X:=0,Z:=Z_{X:=0}}\right] .$$

Here we hold the treatment variable constant at level $X := 0$, but let the mediator variable change to the value it would've attained had treatment $X := 1$ occurred.

In our three node example, the effect of X on Y is unconfounded. In the absence of confounding, the natural indirect effect corresponds to the following statement of conditional probability (involving neither counterfactuals nor do-interventions):

$$\sum_z \mathbb{E}\left[Y \mid X = 0, Z = z\right]\big(\mathbb{P}(Z = z \mid X = 1) - \mathbb{P}(Z = z \mid X = 0)\big) .$$

In this case, we can estimate the natural direct and indirect effect from observational data.

The technical possibilities go beyond the case discussed here. In principle, counterfactuals allow us to compute all sorts of path-specific effects even in the presence of (observed) confounders. We can also design decision rules that eliminate path-specific effects we deem undesirable.

Counterfactual discrimination criteria

Beyond their application to path analysis, counterfactuals can also be used as a tool to put forward normative fairness criteria.

Consider the typical setup of Chapter 2. We have features X, a sensitive attribute A, an outcome variable Y and a predictor $\hat{Y}$.

One criterion that is technically natural would say the following: For every possible demographic described by the event $E := \{X := x, A := a\}$ and every possible setting a' of A we ask that the counterfactual $\hat{Y}_{A:=a}(E)$ and the counterfactual $\hat{Y}_{A:=a'}(E)$ follow the same distribution.

We will refer to it as *counterfactual demographic parity*[147], since it's closely related to the observational criterion *conditional demographic parity*. Recall, conditional demographic parity requires that in each demographic defined by a feature setting $X = x$, the sensitive attribute is independent of the predictor. Formally, we have the conditional independence relation $\hat{Y} \perp A \mid X$. In the case of a binary predictor, this condition is equivalent to requiring for all feature settings x and groups a, a':

$$\mathbb{E}[\hat{Y} \mid X = x, A = a] = \mathbb{E}[\hat{Y} \mid X = x, A = a']$$

The easiest way to satisfy counterfactual demographic parity is for the predictor $\hat{Y}$ to only use non-descendants of A in the causal graph. This is analogous to the statistical condition of only using features that are independent of A.

In the same way that we defined a counterfactual analog of demographic parity, we can explore causal analogs of other statistical criteria in Chapter 2.

In doing so, we need to be careful in separating technical questions about the difference between observational and causal criteria from the normative content of the criterion. Just because a causal variant of a criterion might get around some statistical issues of non-causal correlations does not mean that the causal criterion resolves normative concerns or questions with its observational cousin.

[147] This criterion was introduced as *counterfactual fairness* (Kusner et al., "Counterfactual Fairness," in *Proc. 30Th NIPS*, 2017, 4069–79).

Counterfactuals in the law

Many scholars see support for a counterfactual interpretation of United States discrimination law in various rulings by judges that seemed to have invoked counterfactual language. Here's a quote from a popular recent textbook on causal inference:[148]

> U.S. courts have issued clear directives as to what constitutes employment discrimination. According to law makers, "The central question in any employment-discrimination case is whether the employer would have taken the same action had the employee been of a different race (age, sex, religion, national origin etc.) and everything else had been the same." (In Carson vs Bethlehem Steel Corp., 70 FEP Cases 921, 7th Cir. (1996).)

[148] Pearl, Glymour, and Jewell, *Causal Inference in Statistics: A Primer* (Wiley, 2016), 114.

Unfortunately, the situation is not so simple. This quote invoked here—and in several other technical papers on the topic—expresses the opinion of judges in the 7th Circuit Court at the time. This court is one of thirteen United States courts of appeals. The case has little to no precedential value; the quote cannot be considered a definitive statement on what employment discrimination means under either Title VII or Equal Protection law.

Even the U.S. Supreme Court has not issued "clear directives" that cleanly map onto technical criteria, as we examined in Chapter 3. There is currently no strong basis to support a claim such as U.S. law and legal precedents necessitate a formal counterfactual criterion to determine discrimination. Whether formal counterfactual reasoning *should* become the legal basis of deciding discrimination cases is a separate question.

Harvard college admissions

The language of counterfactuals regularly enters discussions of discrimination in high stakes scenarios. A recent example is the important Harvard admissions lawsuit.

In a trial dating back to 2015, the plaintiff *Students for Fair Admissions* (SFFA)[149] allege discrimination in Harvard undergraduate admissions against Asian-Americans.

[149] Plaintiff SFFA is an offshoot of Edward Blum's Project on Fair Representation, a legal defense fund which aims to end the use of race in voting, education, contracting, and employment.

The trial entailed unprecedented discovery regarding higher education admissions processes and decision-making, including statistical analyses of individual-level applicant data from the past five admissions cycles.

The plaintiff's expert report claims:

> Race plays a significant role in admissions decisions. Consider the example of an Asian-American applicant who is male, is not disadvantaged, and has other characteristics that result in a 25% chance

> of admission. Simply changing the race of the applicant to white—and leaving all his other characteristics the same—would increase his chance of admission to 36%. Changing his race to Hispanic (and leaving all other characteristics the same) would increase his chance of admission to 77%. Changing his race to African-American (again, leaving all other characteristics the same) would increase his chance of admission to 95%.[150]

[150] Plaintiff's expert report of Peter S. Arcidiacono, Professor of Economics at Duke University.

The plaintiff's charge, summarized above, is based on the technical argument that conditional statistical parity is not satisfied by a model of Harvard's admissions decisions.[151] Formally, denote by $\hat{Y}$ a model of Harvard's admissions decisions, by X a set of applicant features deemed relevant for admission, and denoting by A the applicant's reported race we have that

$$\mathbb{E}[\hat{Y} \mid X = x, A = a] < \mathbb{E}[\hat{Y} \mid X = x, A = a'] - \delta\,,$$

for some groups a, a' and some significant value of $\delta > 0$.

[151] Harvard's decision process isn't codified as a formal decision rule. Hence, to talk about Harvard's decision rule formally, we first need to model Harvard's decision rule. The plaintiff's expert Arcidiacono did so by fitting a logistic regression model against Harvard's past admissions decisions in terms of variables deemed relevant for the admission decision.

The violation of this condition certainly depends on which features we deem relevant for admissions. Indeed, this is to a large extent the response of the defendant's expert.[152]

[152] Defendant's expert report of David Card, Professor of Economics at the University of California, Berkeley.

The selection and discussion of what constitute relevant features is certainly important for the interpretation of conditional statistical parity. But arguably a bigger question is whether a violation of conditional statistical parity constitutes evidence of discrimination in the first place. This isn't merely a question of having selected the right features to condition on.

What is it the plaintiff's expert report means by "changing his race?" The literal interpretation is to "flip" the race attribute in the input to the model without changing any of the other features of the input. But a formal interpretation in terms of attribute swapping is not necessarily what triggers our moral intuition.

A stronger moral intuition derives from the interpretation of such statements as alluding to a hypothetical world in which the applicant had been of a different race at the point of application. The construction of such a hypothetical world is closer to the semantics of counterfactual reasoning.

As we know now, attribute flipping generally does not produce valid counterfactuals. Indeed, if we assume a causal graph in which some of the relevant features are influenced by race, then computing counterfactuals with respect to race would require adjusting downstream features. Changing the race attribute without a change in any other attribute only corresponds to a counterfactual in the case where race does not have any descendant nodes—an implausible assumption.

Attribute flipping is often mistakenly given a counterfactual

causal interpretation. Obtaining valid counterfactuals is in general substantially more involved than flipping a single attribute independently of the others. In particular, we cannot meaningfully talk about counterfactuals without bringing clarity to how the variables under consideration interact.

This raises the question what the substantive knowledge is that is required in creating valid causal models and counterfactuals. We turn to this important topic next.

Validity of causal models

So far we have neglected a central question: What makes a causal model *valid* and how do we create valid causal models?

To approach this question, we start with the position that a *valid* causal model is one that represents scientific knowledge. As Pearl put it:

> The causal diagrams are simply dot-and-arrow pictures that summarize our existing scientific knowledge. The dots represent variables of interest, and the arrows represent known or suspected causal relationships between those variables, namely, which variable "listens" to which others.[153]

[153] Pearl and Mackenzie, *The Book of Why*.

This definition immediately puts some constraints on what we can put in a valid causal model. For example, we couldn't substantiate statements about witchcraft in a valid causal model, since we believe that witchcraft has no scientific reality. We might argue that we could still make causal statements about witchcraft in reference to a fantasy novel. Such causal statements do not represent scientific knowledge, but rather reference an implicit ontology of what witchcraft is in some fictional universe.

Causal statements that lack grounding in established scientific knowledge are better thought of as *convenient mathematical fiction*[154] valid causal models. Such *narrative* causal models can still be useful as the basis of exploring hypothetical scenarios as we did in the Berkeley case.

[154] To borrow an expression from (Glymour, "Comment: Statistics and Metaphysics," *Journal of the American Statistical Association* 81, no. 396 (1986): 964–66).

The line between convenient mathematical fiction and scientifically valid causal models is often blurry. This is especially true in cases where causal models reference constructs with limited established scientific knowledge.

To discuss these issues, a little bit of additional background and terminology will be helpful. Simply put, what things and relationships "dot-and-arrow pictures" reference is in part a matter of *ontology*. What beliefs of knowledge about these things are justified is a matter of *epistemology*.

Let's get acquainted with these terms. Webster's 1913 dictionary defines *ontology* as:

> That department of the science of metaphysics which investigates and explains the nature and essential properties and relations of all beings, as such, or the principles and causes of being.

In order to create a valid causal model, we need to provide clarity about what ontological entities each node references, and what relationships exist between these entities. This task may no longer be difficult for certain objects of study. We might have strong scientifically justified beliefs on how certain mechanical parts in an airplane interact. We can use this knowledge to reliably diagnose the cause of an airplane crash.

In other domains, our subject matter knowledge is less stable and subject to debate. This is where *epistemology* comes in, which the Oxford dictionary defines as:

> The theory of knowledge, especially with regard to its methods, validity, and scope, and the distinction between justified belief and opinion.

Against this backdrop, we will explore some of the ontological and epistemic challenges that causal models relating to questions of discrimination surface.

A motivating example

Consider a claim of employment discrimination of the kind: *The company's hiring practices discriminated against applicants of a certain religion.*

What we will see is how two different ontological representations of *religious affiliation* lead to fundamentally different causal models. Each model in turn leads to different conclusions and comes with modeling challenges that raise difficult questions.

Our first attempt is to model *religious affiliation* as a personal trait or characteristic that someone either does or does not possess. This trait, call it A, may influence choices relating to one's appearance, social practices, and variables relevant to the job, such as, the person's level of education Z. So, we might like to start with a model such as the following:

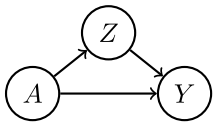


Figure 29: Religion as a root node.

Religious affiliation A is a source node in this graph, which influences the person's level of education Z. Members of certain religions may be steered away from or encouraged towards obtaining a higher level of education by their social peer group. This story is similar to how in our Berkeley admissions graph *sex* influences *department choice*.

This view of religion places burden on understanding the possible indirect pathways, such as $A \to Z \to Y$, through which religion can influence the outcome. There may be insufficient understanding of how a religious affiliation affects numerous other relevant variables throughout life. If we think of religion as a source node in a causal graph, changing it will potentially affect all downstream nodes. For each such downstream node we would need a clear understanding of the mechanisms by which religion influence the node. Where would such *scientific knowledge* of such relationships come from?

But the causal story around religion might also be different. It could be that obtaining a higher level of education causes an individual to lose their religious beliefs.[155] From this perspective, religious affiliation becomes an ancestor of level of education and so the graph might look like this:

[155] In fact this modeling choice was put forward in a recent paper on this topic. (Zhang and Bareinboim, "Fairness in Decision-Making — the Causal Explanation Formula").

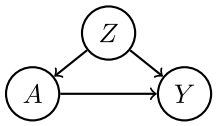


Figure 30: Religion as ancestor.

This view of religion forces us to correctly identify the variables that influence religious affiliation and are also relevant to the decision. Perhaps it is not just level of education, but also socioeconomic status and other factors that have a similar confounding influence.

What is troubling is that in our first graph level of education is a mediator, while in our second graph it is a confounder. The difference is important; to quote Pearl:

> As you surely know by now, mistaking a mediator for a confounder is one of the deadliest sins in causal inference and may lead to the most outrageous error. The latter invites adjustment; the former forbids it.[156]

[156] Pearl and Mackenzie, *The Book of Why*, 276.

Either of our modeling choices follows a natural causal story. Identifying which one is more accurate and applicable to the problem at hand is no easy task. Where do we turn for subject matter knowledge that confirms one model and rejects the other?

The point is not that these are the only two possible modeling choices for how religious affiliation might interact with decision making processes. Rather, our point is that there exist multiple

plausible choices. Moreover, different ontological representations lead to different results emphasizing that how we think of categories such as religion is not just a philosophical debate that practitioners can afford to ignore.

Social construction and human categories

The difficulties we encountered in our motivating example arise routinely when making causal statements involving human kinds and categories, such as, race, religion, or gender, and how these interact with consequential decisions.

Let's take a closer look at the case of race. The ontology of race is complex and still subject of debate, as an article on this topic noted not long ago:

> In recent years, there has been a flurry of work on the metaphysics of race. While it is now widely accepted that races do not share robust, biobehavioral essences, opinions differ over what, if anything, race is.[157]

[157] Mallon, "'Race': Normative, Not Metaphysical or Semantic," *Ethics* 116, no. 3 (2006): 525–51.

Moreover, the act of assigning racial categories to individuals is inextricably tied to a long history of oppression, segregation, and discriminatory practices.[158]

[158] For an entry point to this topic, see for example (Bowker and Leigh Star, *Sorting Things Out: Classification and Its Consequences* (MIT Press, 2000); Fields and Fields, *Racecraft: The Soul of Inequality in American Life* (Verso, 2014); Benjamin, *Race After Technology* (Polity, 2019)).

In the technical literature around discrimination and causality, it's common for researchers to model *race* as a source node in a causal graph, which is to say that race has no incoming arrows. As a source node it can directly and indirectly influence an outcome variable, say, *getting a job offer*.[159]

[159] A recent example of this kind of modeling approach can be found in (Kusner et al., "Counterfactual Fairness").

Implicit in this modeling choice is a kind of naturalistic perspective that views race as a biologically grounded trait, similar to *sex*. The trait exists at the beginning of one's life. Other variables that come later in life, education and income, for example, thus become ancestors in the causal graph.

This view of race challenges us to identify all the possible indirect pathways through which race can influence the outcome. But it's not just this modeling challenge that we need to confront. The view of race as a biologically grounded trait stands in contrast with the *social constructivist* account of race.[160]

[160] For an entry points to this literature, see (Hacking, *The Social Construction of What?* (Harvard University Press, 2000); Haslanger, *Resisting Reality: Social Construction and Social Critique* (Oxford University Press, 2012); Mallon, *The Construction of Human Kinds* (Oxford University Press, 2018); Glasgow et al., "What Is Race?: Four Philosophical Views," 2019).

In this view, roughly speaking, race has no strong biological grounding but rather is a social construct. Race stems from a particular classification of individuals by society, and the shared experiences that stem from the classification. As such, the surrounding social system of an individual influences what race is and how it is perceived. In the constructivist view, *race* is a socially constructed category that individuals are assigned to.

The challenge with adopting this view is that it is difficult to tease out a set of nodes that faithfully represent the influence that society has on race, and perceptions of race. The social constructivist perspective does not come with a simple operational guide for identifying causal structures.

Lack of modularity

Another challenge with socially constructed categories is that they often lack the kind of modularity that a causal diagram requires. Suppose that group membership is constructed from a set of social facts about the group and practices of individuals within the group. We might have some understanding of how these facts and practices constitutively identify group membership. But we may not have an understanding of how each factor individually interacts with each other factor, or whether such a decomposition is even possible.[161]

[161] For a deeper discussion of modularity in causality, see (Cartwright, *Hunting Causes and Using Them, Too* (Cambridge University Press, 2006)).

Causal models inevitably must draw "bounding boxes" around a subset of the possible variables that we could include in a study. Models are therefore systems of reduced complexity that isolate a set variables from their possibly chaotic surroundings in the universe. Pearl summarizes this philosophical point crisply:

> If you wish to include the entire universe in the model, causality disappears because interventions disappear—the manipulator and the manipulated lose their distinction. However, scientists rarely consider the entirety of the universe as an object of investigation. In most cases the scientist carves out a piece from the universe and proclaims that piece *in*—namely, the *focus* of investigation. The rest of the universe is then considered *out* or *background* and is summarized by what we call *boundary conditions*. This choice of *ins* and *outs* creates asymmetry in the way we look at things, and it is this asymmetry that permits us to talk about "outside intervention" and hence about causality and cause-effect directionality.[162]

[162] Pearl, *Causality*, 420.

What we learn from a causal model therefore depends on what bounding box we chose. Too broad a bounding box can lead to a Byzantine model that requires too many interactions to be modeled. Too narrow a bounding box can fail to capture salient aspects of the object of our study. Drawing adequate bounding boxes around socially constructed categories is a delicate task.

Ontological instability

The previous arguments notwithstanding, pragmatist might accuse our discussion of adding unnecessary complexity to what might seem like a matter of common sense to some. Surely, we could also

find subtlety in other characteristics, such as, smoking habits or physical exercise. How is race different from other things we reference in causal models?

An important difference is a matter of ontological stability. When we say *rain caused the grass to be wet* we also refer to an implicit understanding of what rain is, what grass is, and what wet means. However, we find that acceptable in this instance, because all three things we refer to in our causal statement have *stable enough* ontologies. We know what we reference when we invoke them. To be sure, there could be subtleties in what we call grass. Perhaps the colloquial term *grass* does not correspond to a precise botanical category, or one that has changed over time and will again change in the future. However, by making the causal claim, we implicitly assert that these subtleties are irrelevant for the claim we made. We know that grass is a plant and that other plants would also get wet from rain. In short, we believe the ontologies we reference are *stable enough* for the claim we make.

This is not always an easy judgment to make. There are, broadly speaking, at least two sources of ontological instability. One stems from the fact that the world changes over time. Both social progress, political events, and our own epistemic activities may obsolete theories, create new categories, or disrupt existing ones.[163]

[163] Mallon calls this form of instability *Taylor instability* in reference to work by the philosopher Charles Taylor. (Mallon, *The Construction of Human Kinds*)

Hacking's work describes another important source of instability. Categories lead people who putatively fall into such categories to change their behavior in possibly unexpected ways. Individuals might conform or disconform to the categories they are confronted with. As a result, the responses of people, individually or collectively, invalidate the theory underlying the categorization. Hacking calls this a "looping effect."[164] As such, social categories are moving targets that need constant revision.[165]

[164] Hacking, "Making up People," *London Review of Books* 28, no. 16 (2006).

[165] Related feedback effects are well known in policy-making and sometimes go by the name *Goodhart's law* or *Campbell's law*. Patterns observed in a population tend to break down when used for consequential classification or control purposes.

Certificates of ontological stability

The debate around human categories in causal models is by no means new. But it often surfaces in a seemingly unrelated, yet long-standing discussion around causation and manipulation.

One school of thought in causal inference aligns with the mantra *no causation without manipulation*, a view expressed by Holland in an influential article from 1986:

> Put as bluntly and as contentiously as possible, in this article I take the position that causes are only those things that could, in principle, be treatments in experiments.[166]

[166] Holland, "Statistics and Causal Inference," *Journal of the American Statistical Association (JASA)* 81 (1986): 945–70.

Holland goes further by arguing that statements involving "attributes" are necessarily statements of association:

> The only way for an attribute to change its value is for the unit to change in some way and no longer be the same unit. Statements of "causation" that involve attributes as "causes" are always statements of association between the values of an attribute and a response variable across the units in a population.[167]

[167] Holland.

To give an example, Holland maintains that the sentence "She did well on the exam because she is a woman" means nothing but "the performance of women on the exam exceeds, in some sense, that of men."[168]

[168] Holland.

If we believed that there is no causation without manipulation, we would have to refrain from including immutable characteristics in causal models altogether. After all, there is by definition no experimental mechanism that turns immutable attributes into treatments.

Holland's view remains popular among practitioners of the potential outcomes model. The assumptions common in the potential outcomes model are easiest to conceptualize by analogy with a well-designed randomized trial. Practitioners in this framework are therefore used to conceptualizing causes as things that could, in principle, be a treatment in randomized controlled trials.

The desire or need to make causal statements involving race in one way or the other not only arises in the context of discrimination. Epidemiologists encounter the same difficulties when confronting health disparities,[169] as do social scientists when reasoning about inequality in poverty, crime, and education.

[169] Jackson and VanderWeele, "Decomposition Analysis to Identify Intervention Targets for Reducing Disparities"; VanderWeele and Robinson, "On Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables," *Epidemiology*, 2014.

Practitioners facing the need of making causal statements about race often turn to a particular conceptual trick. The idea is to change object of study from the *effect of race* to the effect of *perceptions of race*.[170] What this boils down to is that we change the units of the study from individuals with a race attribute to *decision makers*. The treatment becomes *exposure to race* through some observable trait, like the name on a CV in a job application setting. The target of the study is then how decision makers respond to such *racial stimuli* in the decision-making process. The hope behind this maneuver is that exposure to race, unlike race itself, may be something that we can control, manipulate, and experiment with.

[170] Greiner and Rubin, "Causal Effects of Perceived Immutable Characteristics," *The Review of Economics and Statistics* 93, no. 3 (2011): 775–85.

While this approach superficially avoids the difficulty of conceptualizing manipulation of immutable characteristics, it shifts the burden elsewhere. We now have to sort out all the different ways in which we think that race could possibly be perceived: through names, speech, style, and all sorts of other characteristics and combinations thereof. But not only that. To make a counterfactual statements viz-a-viz *exposure to race*, we would have to be able to create the authentic background conditions under which all these perceptible characteristics would've come out in a manner that's consistent with a different

racial category. There is no way to construct such counterfactuals accurately without a clear understanding of what we mean by the category of race.[171]

Just as we cannot talk about witchcraft in a valid causal model for lack of any scientific basis, we also cannot talk about perceptions of witchcraft in a valid causal model for the very same reason. Similarly, if we lack the ontological and epistemic basis for talking about race in a valid causal model, there is no easy remedy to be found in moving to perceptions of race.[172]

In opposition to Holland's view, other scholars, including Pearl, argue that causation does not require manipulability but rather an understanding of *interactions*.

We can reason about hypothetical Volcano eruptions without being able to manipulate Volcanoes. We can explain the mechanism that causes tides without being able to manipulate the moon by any feasible intervention. What is required is an understanding of the ways in which a variable interacts with other variables in the model. Structural equations in a causal model are *response functions*. We can think of a node in a causal graph as receiving messages from its parent nodes and responding to those messages. Causality is thus about who *listens* to whom. We can form a causal model once we know how the nodes in it interact.

But as we saw the conceptual shift to *interaction*—who *listens* to whom—by no means makes it straightforward to come up with valid causal models. If causal models organize available scientific or empirical information, there are inevitably limitations to what constructs we can include in a causal model without running danger of divorcing the model from reality. Especially in sociotechnical systems, scientific knowledge may not be available in terms of precise modular response functions.

We take the position that causes need not be experimentally manipulable. However, our discussion motivates that constructs referenced in causal models need a certificate of ontological and epistemic stability. Manipulation can be interpreted as a somewhat heavy-handed approach to clarify the ontological nature of a node by specifying an explicit experimental mechanism for manipulating the node. This is one way, but not the only way, to clarify what it is that the node references.

Looking ahead

We did not resolve the question of validity in causal modeling around discrimination. Nor do we expect that these questions can be resolved at generality. Questions of validity depend on the purpose

[171] For a deeper discussion of this point and an insightful critique of counterfactual arguments about race more broadly, see (Kohler-Hausmann, "Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination," *SSRN*, 2019).

[172] Note the term *racecraft* coined in (Fields and Fields, *Racecraft*).

and scope of the model.[173] We will return to questions of validity in our next chapter on measurement that provides a helpful complementary perspective.

[173] On this point of purpose and scope, see Cartwright's essay in (Peschard and Van Fraassen, *The Experimental Side of Modeling* (University of Minnesota Press, 2018)).

Problem set

Bibliographic notes and further reading

Introductions to causality

There are several excellent introductory textbooks on the topic of causality. For an introduction to causality turn to Pearl's primer,[174] or the more comprehensive text.[175] At the technical level, Pearl's text emphasizes causal graphs and structural causal models. Our exposition of Simpson's paradox and the UC Berkeley was influenced by Pearl's discussion, updated for a new popular audience book.[176] All of these texts touch on the topic of discrimination. In these books, Pearl takes the position that discrimination corresponds to the direct effect of the sensitive category on a decision.

[174] Pearl, Glymour, and Jewell, *Causal Inference in Statistics*.

[175] Pearl, *Causality*.

[176] Pearl and Mackenzie, *The Book of Why*.

The technically-minded reader will enjoy complementing Pearl's book with the recent open access text by Peters, Janzing, and Schölkopf[177] that is available online. The text emphasizes two variable causal models and applications to machine learning. See Spirtes, Glymour and Scheines[178] for a general introduction based on causal graphs with an emphasis on *graph discovery*, i.e., inferring causal graphs from observational data.

[177] Peters, Janzing, and Schölkopf, *Elements of Causal Inference*.

[178] Spirtes et al., *Causation, Prediction, and Search* (MIT press, 2000).

Morgan and Winship[179] focus on applications in the social sciences. Imbens and Rubin[180] give a comprehensive overview of the technical repertoire of causal inference in the potential outcomes model. Angrist and Pischke[181] focus on causal inference and potential outcomes in econometrics.

[179] Morgan and Winship, *Counterfactuals and Causal Inference* (Cambridge University Press, 2014).

[180] Imbens and Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences* (Cambridge University Press, 2015).

[181] Angrist and Jörn-Steffen, *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press, 2009).

Hernan and Robins[182] give another detailed introduction to causal inference that draws on the authors' experience in epidemiology.

[182] Hernán and Robins, *Causal Inference* (Boca Raton: Chapman & Hall/CRC, forthcoming, 2019).

Pearl[183] already considered the example of gender discrimination in UC Berkeley graduate admissions that we discussed at length. In his discussion, he implicitly advocates for a view of discussing discrimination based on the causal graphs by inspecting which paths in the graph go from the sensitive variable to the decision point.

[183] Pearl, *Causality*.

Recent technical work

The topic of causal reasoning and discrimination gained significant momentum in the computer science and statistics community around

2017. Zhang et al.[184] previously also considered discrimination analysis via path-specific causal effects. Kusner et al.[185] introduced a notion of *counterfactual fairness*. The authors extend this line of thought in another work.[186] Chiappa introduces a path-specific notion of counterfactual fairness.[187]

Kilbertus et al.[188] distinguish between two graphical causal criteria, called *unresolved discrimination* and *proxy discrimination*. Both notions correspond to either allowing or disallowing paths in causal models. Razieh and Shpitser[189] conceptualize discrimination as the influence of the sensitive attribute on the outcome along certain *disallowed* causal paths. Chiappa and Isaac[190] give a tutorial on causality and fairness with an emphasis on the COMPAS debate.

There is also extensive relevant scholarship in other disciplines that we cannot fully survey here. Of relevance is the vast literature in epidemiology on health disparities. In particular, epidemiologists have grappled with race and gender in causal models. See, for example, the article by VanderWeele and Robinson,[191] as well as Krieger's comment on the article,[192] and Krieger's article on discrimination and health inequalities[193] for a starting point.

Sociological and philosophical debate

A recent article by Hirschman and Reed[194] does an excellent job at summarizing and systematizing sociological accounts of causality. Hirschman and Reed argue that traditional sociological and formal accounts of causality conceptualize causes as forces that govern fixed entities with variable attributes and the modular relationships between fixed entities. The article distinguishes between three main strands of such *forcing cause* accounts: *variable causality*, *treatment and manipulation*, and *mechanisms*. To connect this distinction with the content of this chapter, it may be helpful to analogize variable causality with drawing causal diagrams, treatment and manipulation with Rubin's potential outcomes model, and mechanisms with mediation analysis.

In this chapter we took a rather pragmatic perspective on causality by developing the minimal conceptual and technical tools to understand ongoing research on causal inference. In doing so, we've ignored centuries of philosophical debate around causality. Cartwright's work[195] is a good starting point in this direction.

Systems, dynamics, feedback loops

So far we have assumed that our causal models are always acyclic. Variables cannot simultaneously cause each other. In many applications it does make sense to talk about cyclic dependencies. For

[184] Zhang, Wu, and Wu, "A Causal Framework for Discovering and Removing Direct and Indirect Discrimination," in *Proc.* 26*Th IJCAI*, 2017, 3929–35.

[185] Kusner et al., "Counterfactual Fairness."

[186] Russell et al., "When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness," in *Proc.* 30*Th NIPS*, 2017, 6417–26.

[187] Chiappa, "Path-Specific Counterfactual Fairness," in *Proc.* 33*Rd AAAI*, vol. 33, 2019, 7801–8.

[188] Kilbertus et al., "Avoiding Discrimination Through Causal Reasoning," in *Proc.* 30*Th NIPS*, 2017, 656–66.

[189] Nabi and Shpitser, "Fair Inference on Outcomes," in *Proc.* 32*Nd AAAI*, 2018, 1931–40.

[190] Chiappa and Isaac, "A Causal Bayesian Networks Viewpoint on Fairness," *Arxiv.org* arXiv:1907.06430 (2019).

[191] VanderWeele and Robinson, "On Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables."

[192] Krieger, "On the Causal Interpretation of Race," *Epidemiology* 25, no. 6 (2014): 937.

[193] Krieger, "Discrimination and Health Inequities," *International Journal of Health Services* 44, no. 4 (2014): 643–710.

[194] Hirschman and Reed, "Formation Stories and Causality in Sociology," *Sociological Theory* 32, no. 4 (2014): 259–82.

[195] Cartwright, *Hunting Causes and Using Them, Too*.

example, we might reason that the economy grew, because of an increase investments, and that investments grew, because of a growing economy. The formalisms we encountered do not directly apply to such *closed loop* dynamics.

There are a few ways of coping. One is to *unroll* the system into discrete time steps. What this means is that we repeat the causal graph for some number of discrete time steps in such a manner that each node appears multiple times indexed by a time step.

An alternative route is to develop formalisms that directly deal with actions in closed loop dynamics. See, work by Bongers et al.[196] on extending the structural causal model perspective to cyclic models.

Traditionally, feedback systems are the focus of control theory, an area with a long history and vast technical repertoire.[197] While much of control theory focuses on physical systems, concepts from control theory also influenced policy and decision making in other domains. A well-known example is the area of *system dynamics* pioneered by Forrester[198] in the 60s and 70s that lead to some politically influential works such as *Limits to Growth*.[199] But see Baker's thesis[200] for a history of system modeling and its pitfalls.

[196] Bongers et al., "Theoretical Aspects of Cyclic Structural Causal Models," *arXiv.org Preprint* arXiv:1611.06221v2 (2018).

[197] Aström and Murray, *Feedback Systems: An Introduction for Scientists and Engineers* (Princeton university press, 2010).

[198] Forrester, "Urban Dynamics," *IMR; Industrial Management Review (Pre-1986)* 11, no. 3 (1970): 67; Forrester, "Counterintuitive Behavior of Social Systems," *Technological Forecasting and Social Change* 3 (1971): 1–22; Forrester, "System Dynamics, Systems Thinking, and Soft OR," *System Dynamics Review* 10, no. 2–3 (1994): 245–56.

[199] Meadows, Randers, and Meadows, *The Limits to Growth: The 30-Year Update* (Routledge, 2012).

[200] Baker, "World Processors: Computer Modeling, Global Environmentalism, and the Birth of Sustainable Development" (Northwestern University, 2019).

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.[201] 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.[202]

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.[203]

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

[201] 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.

[202] 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.

[203] Bogen and Rieke, "Help wanted: an examination of hiring algorithms, equity, and bias" (Technical report, Upturn, 2018).

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 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 2. 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 4). 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.[204] They have since been successfully applied to many other domains.

[204] Wienk et al., "Measuring Racial Discrimination in American Housing Markets: The Housing Market Practices Survey." 1979.

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.[205] 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.

[205] Ayres and Siegelman, "Race and Gender Discrimination in Bargaining for a New Car," *The American Economic Review*, 1995, 304–21.

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.[206]

[206] 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.

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 4, 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.[207]

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 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."[208]

Another famous audit study tested discrimination in the labor market.[209] 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.

[207] Freeman et al., "Looking the Part: Social Status Cues Shape Race Perception," *PloS One* 6, no. 9 (2011): e25107.

[208] 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.

[209] Bertrand and Mullainathan, "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review* 94, no. 4 (2004): 991–1013.

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.[210]

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 intuition.[211] 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.[212] The method has also been used to study discrimination based on gender, sexual orientation, and physical appearance.[213] It has also been used outside the labor market, in retail and academia.[214] Finally, trends over time have been studied: a meta-

[210] 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, no. 1 (2007): 104–33.

[211] Kohler-Hausmann, "Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination," *Nw. UL Rev.* 113 (2018): 1163.

[212] Bertrand and Duflo, "Field Experiments on Discrimination," in *Handbook of Economic Field Experiments*, vol. 1 (Elsevier, 2017), 309–93.

[213] Bertrand and Duflo.

[214] Bertrand and Duflo.

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.[215]

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.

[215] Quillian et al., "Meta-Analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring over Time," *Proceedings of the National Academy of Sciences* 114, no. 41 (2017): 10870–75.

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,[216] but in general even this strategy can be impractical.

[216] Blank, "The Effects of Double-Blind Versus Single-Blind Reviewing: Experimental Evidence from the American Economic Review," *The American Economic Review*, 1991, 1041–67.

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.[217] 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.[218] 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.

[217] Pischke, "Empirical Methods in Applied Economics: Lecture Notes," 2005.

[218] Bertrand, Duflo, and Mullainathan, "How Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics* 119, no. 1 (2004): 249–75.

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.[219] The effects reported in the study are meaningful to job seekers who engage in this practice.

[219] Kang et al., "Whitened Resumes: Race and Self-Presentation in the Labor Market," *Administrative Science Quarterly* 61, no. 3 (2016): 469–502.

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 3). 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.[220] 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.[221]

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.[222] 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.

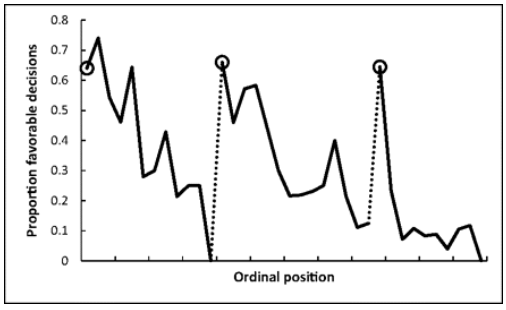


Figure 31: (from Danziger et al.): fraction of favorable rulings over the course of a day. The dotted lines indicate food breaks.

The finding would be extraordinary if the order of cases was truly random.[223] 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.[224] This invalidates the conclusion of the original study.

[220] Eren and Mocan, "Emotional Judges and Unlucky Juveniles," *American Economic Journal: Applied Economics* 10, no. 3 (2018): 171–205.

[221] 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."

[222] Danziger, Levav, and Avnaim-Pesso, "Extraneous Factors in Judicial Decisions," *Proceedings of the National Academy of Sciences* 108, no. 17 (2011): 6889–92.

[223] In fact, it would be so extraordinary that it has been argued that the study should be dismissed 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 (Lakens, "Impossibly Hungry Judges" (https://daniellakens.blogspot.com/2017/07/impossibly-hungry-judges.html, 2017))

[224] Weinshall-Margel and Shapard, "Overlooked Factors in the Analysis of Parole Decisions," *Proceedings of the National Academy of Sciences* 108, no. 42 (2011): E833–33.

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.[225] 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.

[225] Huq, "Racial Equity in Algorithmic Criminal Justice," *Duke LJ* 68 (2018): 1043.

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.[226] Thus we can estimate the impact of the scholarship as if we did a randomized controlled trial.

[226] 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.[227]

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[228] (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 2, 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 4.

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).[229]

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 causal framework to make our claims more precise. In this case, causal modeling might

[227] Norton, "The Supreme Court's Post-Racial Turn Towards a Zero-Sum Understanding of Equality," *Wm. & Mary L. Rev.* 52 (2010): 197.

[228] Ayres, "Three Tests for Measuring Unjustified Disparate Impacts in Organ Transplantation: The Problem of" Included Variable" Bias," *Perspectives in Biology and Medicine* 48, no. 1 (2005): 68–S87.

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

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 $\hat{Y}$ 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 ($\hat{Y} = 1$) 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). 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.

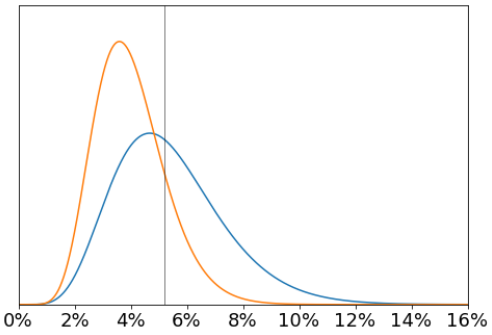


Figure 32: Hypothetical probability density of loan default for two groups, women (orange) and men (blue).

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 $(\hat{Y}, A, Y)$ 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.[230] 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.

[230] Simoiu, Corbett-Davies, and Goel, "The Problem of Infra-Marginality in Outcome Tests for Discrimination," *The Annals of Applied Statistics* 11, no. 3 (2017): 1193–1216.

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.[231]

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.[232]

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.

[231] Lakkaraju et al., "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* (ACM, 2017), 275–84.

[232] Bird et al., "Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI," in *Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2016.

Summary of traditional tests and methods

Table 12: 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 / limitations
1	Audit study	Blindness	A-exp $:=$, $X :=$, R	Difficult to interpret
2	Natural experiment especially diff-in-diff	Impact of blinding	A-exp $\sim$, R	Confounding; SUTVA violations; other
3	Natural experiment	Arbitrariness	$W \sim$, R	Unobserved confounders
4	Natural experiment especially regr. disc.	Impact of decision	R, Y or Y'	Sample size; confounding; other technical difficulties
5	Regression analysis	Blindness	X, A, R	Unreliable due to proxies
6	Regression analysis	Cond. demographic parity	X, A, R	Weak moral justification
7	Outcome test	Predictive parity	$A, Y \mid \hat{Y} = 1$	Infra-marginality
8	Threshold test	Sufficiency	$X', A, Y \mid \hat{Y} = 1$	Model-specific
9	Experiment	Separation/error rate parity	$A, R, \hat{Y} :=$, Y	Often unethical or impractical
10	Observational test	Demographic parity	A, R	See Chapter 2
11	Mediation analysis	"Relevant" mechanism	X, A, R	See Chapter 4

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.[233]

A statistical discriminator, in contrast, aims to make optimal predictions about the target variable using all available information, including the protected attribute.[234] 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

[233] Becker, *The Economics of Discrimination* (University of Chicago Press, 1957).

[234] Phelps, "The Statistical Theory of Racism and Sexism," *The American Economic Review* 62, no. 4 (1972): 659–61; Arrow, "The Theory of Discrimination," *Discrimination in Labor Markets* 3, no. 10 (1973): 3–33.

they only consider causes that are relatively proximate to the decision point). We will discuss structural discrimination in Chapter 6.

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.[235] 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.

[235] For example, laws restricting employers from asking about applicants' criminal history resulted in employers using race as a proxy for it. See (Agan and Starr, "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment," *The Quarterly Journal of Economics* 133, no. 1 (2017): 191–235).

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.

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.[236] 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.[237]

[236] Williams and Ceci, "National Hiring Experiments Reveal 2: 1 Faculty Preference for Women on STEM Tenure Track," *Proceedings of the National Academy of Sciences* 112, no. 17 (2015): 5360–65.

[237] 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.

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.[238] Over the decades, however, such overt attitudes have become less common, or at least less likely to be expressed.[239] 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.[240] 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 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*

[238] Neckerman and Kirschenman, "Hiring Strategies, Racial Bias, and Inner-City Workers," *Social Problems* 38, no. 4 (1991): 433–47.

[239] Pager and Shepherd, "The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets," *Annu. Rev. Sociol* 34 (2008): 181–209.

[240] Rivera, *Pedigree: How Elite Students Get Elite Jobs* (Princeton University Press, 2016).

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.[241] 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.

[241] Posselt, *Inside Graduate Admissions* (Harvard University Press, 2016).

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.[242]

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.[243] 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.[244] 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.

[242] 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.

[243] Roth, "The Origins, History, and Design of the Resident Match," *Jama* 289, no. 7 (2003): 909–12; Friedman and Nissenbaum, "Bias in Computer Systems."

[244] A 2014 paper issued a call to action towards this type of research. Most of the studies that we cite postdate that piece. (Sandvig et al., "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms," *ICA Pre-Conference on Data and Discrimination*, 2014)

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)[245]. 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[246] and from applicants,[247] but it is limited to anecdotes. There is also evidence from the lab experiments on the task of predicting occupation from online biographies.[248]

We briefly survey other findings. Automated essay grading software tends to assign systematically lower scores to some demographic groups[249] compared to human graders, who may themselves provide biased ratings.[250] Hate speech detection models use markers of dialect as predictors of toxicity, according to a lab study,[251] 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.[252] Speech-to-text systems perform worse for speakers with certain accents.[253] In all these cases, the author or speaker of the text is potentially harmed. In other NLP systems, i.e., those involving natural language generation or translation, there is a different type of fairness concern, namely the generation of text reflecting cultural

[245] For a treatise on AAE, see (Green, *African American English: A Linguistic Introduction* (Cambridge University Press, 2002)). 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.

[246] Dastin, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women," *Reuters*, 2018.

[247] Buranyi, "How to Persuade a Robot That You Should Get the Job" (Guardian, 2018).

[248] De-Arteaga et al., "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* (ACM, 2019), 120–28.

[249] Ramineni and 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, no. 1 (2018): 1–31.

[250] Amorim, Cançado, and 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)*, 2018, 229–37.

[251] Sap et al., "The Risk of Racial Bias in Hate Speech Detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, 1668–78.

[252] Kiritchenko and Mohammad, "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems," *arXiv Preprint arXiv:1805.04508*, 2018.

[253] Tatman, "Gender and Dialect Bias in YouTube's Automatic Captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (Valencia, Spain: Association for Computational Linguistics, 2017), 53–59, https://doi.org/10.18653/v1/W17-1606.

prejudices resulting in representational harm to a group of people.[254] 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.

[254] Solaiman et al., "Release Strategies and the Social Impacts of Language Models," *arXiv Preprint arXiv:1908.09203*, 2019.

Table 13: **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	Harm
Perception	Language id speech-to-text	Underrep. in training corpus	Degraded service
Automating judgment	Toxicity detection essay grading	Human labels, underrep. in training corpus	Adverse decisions
Predicting outcomes	Resume filtering	Various, including human labels	Adverse decisions
Sequence prediction	Language generation translation	Cultural stereotypes, historical prejudices	Repres. harm

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.[255] The study found that all three classifiers perform better on male faces than female faces (8.1% –

[255] Buolamwini and Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Conference on Fairness, Accountability and Transparency*, 2018, 77–91.

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.[256] 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).[257]

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.[258] 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.[259] This possibility was shown in a simple linear model of face recognition as early as 1991.[260] Many commercial products have had difficulty detecting faces of darker-skinned people McEntegart.[261] Similar results are known from lab studies of publicly available object detection models.[262]

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

[256] Vries et al., "Does Object Recognition Work for Everyone?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, 52–59.

[257] Shankar et al., "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.

[258] Simonite, "When It Comes to Gorillas, Google Photos Remains Blind," *Wired, January* 13 (2018); Hern, "Flickr Faces Complaints over 'Offensive'auto-Tagging for Photos," *The Guardian* 20 (2015).

[259] Martineau, "Cities Examine Proper—and Improper—Uses of Facial Recognition | WIRED" (https://www.wired.com/story/cities-examine-proper-improper-facial-recognition/ 2019).

[260] O'TOOLE et al., "Simulating the 'Other-Race Effect'as a Problem in Perceptual Learning," *Connection Science* 3, no. 2 (1991): 163–78.

[261] "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).

[262] Wilson, Hoffman, and Morgenstern, "Predictive Inequity in Object Detection," *arXiv Preprint arXiv:1902.11097*, 2019.

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.[263]

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.[264]

[263] Turow et al., "Americans Reject Tailored Advertising and Three Activities That Enable It," *Available at SSRN 1478214*, 2009.

[264] Raghavan et al., "Mitigating Bias in Algorithmic Employment Screening: Evaluating Claims and Practices," *arXiv Preprint arXiv:1906.09208*, 2019.

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).[265] 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

[265] Yao and Huang, "Beyond Parity: Fairness Objectives for Collaborative Filtering," in *Advances in Neural Information Processing Systems*, 2017, 2921–30.

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.[266] 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.

[266] Mehrotra et al., "Auditing Search Engines for Differential Satisfaction Across Demographics," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, 626–33.

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.[267]

Representational harms. The book *Algorithms of Oppression* drew attention to the ways in which search engines reinforce harmful racial, gender, and intersectional stereotypes.[268] 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.[269] This can be seen as test for calibration.[270] 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.

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

[267] For in-depth treatments of the history and politics of information platforms, see: (Wu, *The Master Switch: The Rise and Fall of Information Empires* (Vintage, 2010); Gillespie, "The Politics of 'Platforms'," *New Media & Society* 12, no. 3 (2010): 347–64; Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (Yale University Press, 2018); Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech," *Harv. L. Rev.* 131 (2017): 1598)

[268] Noble, *Algorithms of Oppression*.

[269] Kay, Matuszek, and Munson, "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations."

[270] 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.

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 3.

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.[271] 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.[272]

There is a vast literature on how race and gender are portrayed in ads that we consider to fall under the persuasion framework.[273] 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 based on users' free-form text descriptions of their interests. These categories were found to include "Jew haters" and many other antisemitic terms.[274] 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

[271] Susser, Roessler, and Nissenbaum, "Online Manipulation: Hidden Influences in a Digital World," *Available at SSRN 3306006*, 2018.

[272] The economic analysis of advertising includes a third category, complementary, that's related to persuasive or manipulative category (Bagwell, "The Economic Analysis of Advertising," *Handbook of Industrial Organization* 3 (2007): 1701–844).

[273] See, e.g. (Coltrane and Messineo, "The Perpetuation of Subtle Prejudice: Race and Gender Imagery in 1990s Television Advertising," *Sex Roles* 42, no. 5–6 (2000): 363–89)

[274] Angwin, Varner, and Tobin, "Facebook Enabled Advertisers to Reach 'Jew Haters'" (ProPublica. https://www.%20propublica.%20org/article/facebook-enabled-advertisers-to-reach-jew-haters, 2017).

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.[275] 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.[276] 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.[277] 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.[278] 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.[279]

[275] Ali et al., "Discrimination Through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes," *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (2019): 199; Lambrecht and Tucker, "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads," *Management Science*, 2019.

[276] Datta, Tschantz, and Datta, "Automated Experiments on Ad Privacy Settings."

[277] Andreou et al., "AdAnalyst" (https://adanalyst.mpi-sws.org/, 2017).

[278] Ali et al., "Discrimination Through Optimization."

[279] Ali et al.

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.[280] 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.[281]

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.[282]

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.[283] 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

[280] Hutson et al., "Debiasing Desire: Addressing Bias & Discrimination on Intimate Platforms," *Proceedings of the ACM on Human-Computer Interaction* 2, no. CSCW (2018): 73.

[281] This is not to say that discrimination is nonexistent. See, e.g., (Ayres, Banaji, and Jolls, "Race Effects on eBay," *The RAND Journal of Economics* 46, no. 4 (2015): 891–917).

[282] Lee et al., "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* (ACM, 2015), 1603–12.

[283] Edelman, Luca, and Svirsky, "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment," *American Economic Journal: Applied Economics* 9, no. 2 (2017): 1–22.

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 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.[284] 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[285] 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).[286] 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.)[287] 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.

[284] Thebault-Spieker, Terveen, and Hecht, "Toward a Geographic Understanding of the Sharing Economy: Systemic Biases in UberX and TaskRabbit," *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, no. 3 (2017): 21.

[285] Ge et al., "Racial and Gender Discrimination in Transportation Network Companies" (National Bureau of Economic Research, 2016).

[286] Levy and Barocas, "Designing Against Discrimination in Online Markets," *Berkeley Tech. LJ* 32 (2017): 1183.

[287] Tjaden, Schwemmer, and Khadjavi, "Ride with Me—Ethnic Discrimination, Social Markets, and the Sharing Economy," *European Sociological Review* 34, no. 4 (2018): 418–32.

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.[288] 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.[289] 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).[290] 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[291] and prior research.[292]

However, a more plausible hypothesis for the filter bubble effect in search comes from a qualitative study.[293] Simplified somewhat for our purposes, it goes as follows: when an event with political significance unfolds, key influencers (politicians, partisan news

[288] Muthukumar et al., "Understanding Unequal Gender Classification Accuracy from Face Images," *arXiv Preprint arXiv:1812.00099*, 2018.

[289] 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 (Harvey and LaPlace, "MegaPixels: Origins, Ethics, and Privacy Implications of Publicly Available Face Recognition Image Datasets," 2019, https://megapixels.cc/).

[290] Robertson et al., "Auditing Partisan Audience Bias Within Google Search," *Proceedings of the ACM on Human-Computer Interaction* 2, no. CSCW (2018): 148.

[291] 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-s html, 2019).

[292] Hannak et al., "Measuring Personalization of Web Search," in *Proceedings of the 22nd International Conference on World Wide Web* (ACM, 2013), 527–38.

[293] Tripodi, "Searching for Alternative Facts: Analyzing Scriptural Inference in Conservative News Practices," *Data & Society*, 2018.

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.[294] 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.[295]

A final example to reinforce the fact that disparity-producing mechanisms can 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.[296] 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.

[294] 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."

[295] But see (Golebiewski and Boyd, "Data Voids: Where Missing Data Can Easily Be Exploited," *Data & Society* 29 (2018)) ("Data Void Type #4: Fragmented Concepts") for an argument that search engines' decision not to collapse related concepts contributes to this fragmentation.

[296] Valentino-Devries, Singer-Vine, and Soltani, "Websites Vary Prices, Deals Based on Users' Information," *Wall Street Journal* 10 (2012): 60–68.

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 3), 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.[297] 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.

[297] Raghavan et al., "Mitigating Bias in Algorithmic Employment Screening."

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.[298] 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:

[298] Datta, Tschantz, and Datta, "Automated Experiments on Ad Privacy Settings."

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.[299]

There are additional nuances relating to proper randomization and controls, for which we refer the reader to the study.[300] 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.[301] 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.[302] 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.[303] 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 lim-

[299] Ojala and Garriga, "Permutation Tests for Studying Classifier Performance," *Journal of Machine Learning Research* 11, no. Jun (2010): 1833–63.

[300] Datta, Tschantz, and Datta, "Automated Experiments on Ad Privacy Settings."

[301] Venkatadri et al., "Investigating Sources of PII Used in Facebook's Targeted Advertising," *Proceedings on Privacy Enhancing Technologies* 2019, no. 1 (2019): 227–44.

[302] Bashir et al., "Tracing Information Flows Between Ad Exchanges Using Retargeted Ads," in *USENIX Security Symposium 16*, 2016, 481–96.

[303] Singer-Vine, Valentino-DeVries, and Soltani, "How the Journal Tested Prices and Deals Online" (Wall Street Journal. http://blogs.%20wsj.%20com/digits/2012/12/23/how-the-journal-tested-prices-and-deals-online, 2012).

ited by the requirement for each account to have a distinct credit card associated with it.[304] 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.[305]

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

[304] Chen, Mislove, and Wilson, "Peeking Beneath the Hood of Uber," in *Proceedings of the 2015 Internet Measurement Conference* (ACM, 2015), 495–508.

[305] Salganik, *Bit by Bit: Social Research in the Digital Age* (Princeton University Press, 2019).

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.[306] 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.[307] Another insight is the increasing homogeneity of users' consumption patterns over time in recommender systems.[308]

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.[309] 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.[310] 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.

[306] Bennett and Lanning, "The Netflix Prize," in *Proceedings of KDD Cup and Workshop*, vol. 2007 (New York, NY, USA., 2007), 35.

[307] Lum and Isaac, "To Predict and Serve?"; Ensign et al., "Runaway Feedback Loops in Predictive Policing."

[308] Chaney, Stewart, and Engelhardt, "How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility," in *Proceedings of the 12th ACM Conference on Recommender Systems* (ACM, 2018), 224–32.

[309] Obermeyer et al., "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science* 366, no. 6464 (2019): 447–53; Chouldechova et al., "A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions," in *Conference on Fairness, Accountability and Transparency*, 2018, 134–48.

[310] Passi and Barocas, "Problem Formulation and Fairness," in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (ACM, 2019), 39–48.

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.

7
A broader view of discrimination

Machine learning systems don't operate in a vacuum; they are adopted in societies that already have many types of discrimination intertwined with systems of oppression such as racism. This is at the root of fairness concerns in machine learning. In this chapter we'll take a systematic look at discrimination in society. This will give us a more complete picture of the potential harmful impacts of machine learning. We will see that while a wide variety of fairness interventions are possible—and necessary—only a small fraction of them translate to technical fixes.

Case study: the gender earnings gap on Uber

We'll use a paper that analyzes the gender earnings gap on Uber[311] as a way to apply some of the lessons from the previous two chapters while setting up some of the themes of this chapter.[312] The paper starts with the observation that female drivers earn 7% less on Uber per active hour than male drivers do. It concludes that this gap can be explained by three factors: gender differences in drivers' choices of where to drive, men's greater experience on the platform, and men's tendency to drive faster. It finds that customer discrimination and algorithmic discrimination do not contribute to the gap. We'll take the paper's technical claims at face value, but use the critical framework we've introduced to interpret the findings quite differently from the authors.

First, let's understand the findings in more detail.

The paper analyzes observational data on trips in the United States, primarily in Chicago. Above, we've drawn a causal graph showing what we consider to be the core of the causal model studied in the paper (the authors do not draw such a graph and do not pose their questions in a causal framework; we have chosen to do so for pedagogical purposes). A full graph would be much larger than the Figure; for example, we've omitted a number of additional controls,

[311] Cook et al., "The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers" (National Bureau of Economic Research, 2018).

[312] The study was coauthored by current and former Uber employees.

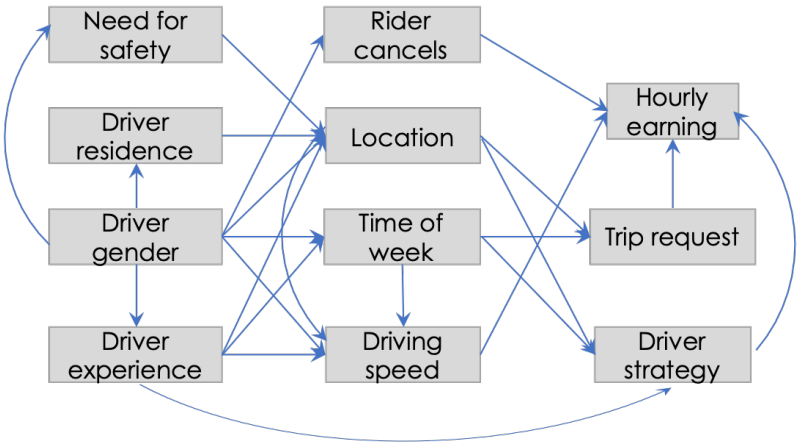


Figure 33: Our understanding of the causal model implicit in the Uber study.

such as race, that are presented in the appendices.

We'll use this graph to describe the findings. At a high level, the graph describes a joint distribution whose samples are trips. To illustrate, different trips corresponding to the same driver will have the same *Residence* (unless the driver moved during their tenure on the platform), but different *Experience* (measured as number of prior trips).

Drivers' hourly earnings are primarily determined by the algorithm that allocates trip requests from riders to drivers. The allocation depends on demand, which in turn varies by location and time of the week (the week-to-week variation is considered noise). Uber's algorithm ignores driver attributes including experience and gender, hence there are no arrows from those nodes to *Trip request*. In addition, a few other factors might affect earnings. Drivers who drive faster complete more trips, drivers may strategically accept or cancel trips, and riders might discriminate by cancelling trips after the driver accepts.

The paper uses a technique called Gelbach decomposition[313] to identify the effect of each of several variables on the hourly earnings. It finds that the earnings gap (i.e. effect of *Driver gender* on *Hourly earning*) can be entirely explained by paths involving *Driver experience*, *Location*, and *Driving speed*. Paths through *Rider cancellation* and *Time of week* don't have significant effects.

The authors further interrogate the effect of gender on location (i.e. the choice of where to drive), and find that women are less likely to drive in less safe areas that also turn out to be more lucrative. They then dig deeper and argue that this effect operates almost entirely by women *residing* in safer areas and choosing to drive based on where they live.

[313] Decomposition is a set of techniques used in economics for quantifying the contribution of various sources to an observed difference in outcomes. Although the authors don't perform causal inference, we will continue to talk about their findings in causal terms for pedagogical purposes. The difference is not salient to the high-level points we wish to make.

The returns to experience could operate in several ways. The authors don't decompose the effect but suggest several possibilities: the choice of where and when to drive and other elements of strategy including which rides to accept. A key finding of the paper is the effect of gender on experience. Men are less likely to leave the platform and drive more hours during each week that they stay on the platform, resulting in a large experience differential. There are no gender differences in *learning* from experience: male and female drivers' behavior changes at the same rate for a given number of trips.

The paper highlights questions that can be studied using observational data but not necessarily with field experiments (audit studies). An audit study of the Uber gender pay gap (along the lines of those discussed in the previous chapter) may have involved varying the driver's name to test the effect on rider cancellation and ratings. Such an experiment would have no way to uncover the numerous other paths by which gender affects earnings.[314]

Causal diagrams in realistic scenarios are more complex than typical textbook examples. We reiterate that the graph above is much simplified compared to the (implicit) graph in the paper. The estimation in the paper proceeds as a series of regressions focusing iteratively on small parts of the graph, rather than an analysis of the entire graph at once. In any messy exercise such as this, there is always the possibility of unobserved confounders.

Despite the number of possible effects considered in the study, it leaves out many others. For example, some drivers may move to take advantage of the earning potential. This would introduce a cycle into our causal graph (*Location –> Residence*). This type of behavior might seem unlikely for an individual driver, which justifies ignoring such effects in the analysis. Over time, however, the introduction of transportation systems has the potential to reshape communities.[315] Today's empirical methods have limitations in understanding these types of long-term phenomena that involve feedback loops.

A more notable omission from the paper is the effect of driver gender on experience. Why do women drop off the platform far more frequently? Could one reason be that they face more harassment from riders? The authors don't seem to consider this question.

This leads to our most salient observation about this study: the narrow definition of discrimination. First, as noted, the study doesn't consider that differential dropout rates might be due to discrimination.[316] This is especially pertinent since the gender gap in hourly earnings is merely 7% whereas the gap in participation rate is a factor of 2.7! One would think that if there is rider discrimination, it would be most apparent in its effect on dropout rates. In contrast, the only

[314] An audit study would be more suited for studying discrimination *by drivers against riders*, in part because drivers in these systems exercise more choice in the matching process than riders do. Indeed, a study found that UberX and Lyft drivers discriminate against Black and female riders. (Ge et al., "Racial and Gender Discrimination in Transportation Network Companies")

[315] Winner, *Do Artifacts Have Politics?* (Routledge, 2017); Bullard, Johnson, and Torres, *Highway Robbery: Transportation Racism & New Routes to Equity* (South End Press, 2004).

[316] For example, the authors say in the abstract: *Our results suggest that, in a "gig" economy setting with no gender discrimination and highly flexible labor markets, women's relatively high opportunity cost of non-paid-work time and gender-based differences in preferences and constraints can sustain a gender pay gap.*

avenue of discrimination considered in the paper involves a (presumably misogynistic) rider who cancels a ride, incurring delays and potentially algorithmic penalties, based solely on the driver's gender.

Further, the authors take an essentialist view of the gender difference in average speed (e.g. "men are more risk tolerant and aggressive than women"). We may question how innate these differences are, given that in contemporary U.S. society, women may face social penalties when they are perceived as aggressive. If this is true of driver-rider interactions, then women who drive as fast as men will receive lower ratings with attendant negative consequences. This is a form of discrimination by riders.[317]

Another possible view of the speed difference, also not considered by the authors, is that male drivers on average provide a lower quality of service due to an increase in accident risk resulting from greater speed (which also creates negative externalities for others on the road). In this view, Uber's matching algorithm discriminates against female drivers by *not* accounting for this difference.[318]

Finally, the paper doesn't consider structural discrimination. It finds that women reside in less lucrative neighborhoods and that their driving behavior is shaped by safety considerations. However, a deeper understanding the reasons for these differences is outside the scope of the paper. In fact, gender differences in safety risks and the affordability of residential neighborhoods can be seen as an example of the greater burden that society places on women. In other words, Uber operates in a society in which women face discrimation and have unequal access to opportunities, and the platform perpetuates those differences in the form of a pay gap.[319]

Let us generalize a bit. There is a large set of studies that seek to explain the reasons for observed disparities in wages or another outcome. Generally these studies find that the direct effect of gender, race, or another sensitive attribute is much smaller than the indirect effect. Frequently this leads to a vigorous debate on whether or not the findings constitute evidence of discrimination or unfairness. There is room for different views on this question. The authors of the Uber study interpreted none of the three paths by which gender impacts earnings—experience, speed, and location—as discrimination; we've argued that all three can plausibly be interpreted as discrimination. Different moral frameworks will lead to different answers. Views on these questions are also politically split. As well, scholars in different fields often tend to answer these questions differently (including, famously, social science and economics).[320]

Certainly these definitional questions are important. However, perhaps the greatest value of studies on mechanisms of discrimination is that they suggest avenues for intervention *without* having to

[317] For a broad discussion of customer ratings as a vehicle for discrimination on Uber, see (Rosenblat et al., "Discriminating Tastes: Customer Ratings as Vehicles for Bias," *Data & Society*, 2016, 1–21).

[318] If riders give lower ratings to drivers who drive faster at the expense of safety, then the matching algorithm does indirectly take safety considerations into account. We think it is unlikely that driver ratings adequately reflect the risks of speeding, due to cognitive biases. After all, that is why we need speed limits instead of leaving it up to drivers.

[319] See (Thebault-Spieker, Terveen, and Hecht, "Toward a Geographic Understanding of the Sharing Economy") for a discussion of many ways in which existing geographic inequalities manifest in sharing economy platforms including Uber.

[320] Small and Pager, "Sociological Perspectives on Racial Discrimination," *Journal of Economic Perspectives* 34, no. 2 (2020): 49–67.

resolve definitional questions. Looking at the Uber study from this lens, several interventions are apparent. Recall that there is a massive gender disparity in the rate at which drivers drop out of the platform. Uber could more actively solicit and listen to feedback from female drivers and use that feedback to inform the design of the app. This may lead to interventions such as making it easier for drivers (and riders) to report harassment and taking stronger action in response to such reports.

As for the speed difference, Uber could warn drivers who exceed the speed limit or whose speed results in a predicted accident risk that crosses some threshold (such a prediction is presumably possible given Uber's access to data). In addition, Uber could use its predictive tools to educate drivers about strategy, decreasing the returns to experience for all drivers. Finally, the findings also give greater urgency to structural efforts to make neighborhoods safe for women. None of these interventions require a consensus on whether or not female drivers on Uber are discriminated against.

Three levels of discrimination

Sociologists organize discrimination into three levels: structural, organizational, and interpersonal.[321] Structural discrimination arises from the ways in which society is organized, both through relatively hard constraints such as discriminatory laws and through softer ones such as norms and customs. Organizational factors operate at the level of organizations or other decision-making units, such as a company making hiring decisions. Interpersonal factors refer to the attitudes and beliefs that result in discriminatory behavior by individuals.

A separate way to classify discrimination is as direct or indirect. By direct discrimination we mean actions or decision processes that make explicit reference to a sensitive attribute. By indirect discrimination we refer to actions or decision processes that make no such reference, yet disadvantage one or more groups. The line between direct and indirect discrimination is hazy and it is better to think of it as a spectrum rather than a binary category.[322]

[321] Pager and Shepherd, "The Sociology of Discrimination"; Small and Pager, "Sociological Perspectives on Racial Discrimination."

[322] For attempts by philosophers to formalize the distinction, see (Altman, "Discrimination," in *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta, Winter 2020 (https://plato.stanford.edu/archives/win2020/entries/discrimination/; Metaphysics Research Lab, Stanford University, 2020)). For a technical treatment of direct vs. indirect effects, refer back to the Causality chapter. See also (Lily Hu, "Direct Effects" (https://phenomenalworld.org/analysis/direct-effects, 2020)); in particular, the point that "any direct effect is really an indirect effect if you zoom further in" on the relevant causal mechanism."

Table 14: Examples of discrimination organized into three levels and on a spectrum of directness

Level	More direct	More indirect
Structural	Laws against same-sex marriage	Better funded schools in wealthier, more segregated areas
Organizational	Lack of disability accommodations	Networked hiring

Level	More direct	More indirect
Interpersonal	Overt animus	Belief in need for innate brilliance (combined with gender stereotypes)

Structural factors

Structural factors refer to ways in which society is organized. A law that overtly limits opportunities for certain groups is an example of a direct structural factor. Due to various rights revolutions around the world, there are fewer of these laws today than there used to be. Yet, discriminatory laws are far from a thing of the past. For example, as of 2021, a mere 29 countries recognize marriage equality.[323] Further, discriminatory laws of the past have created structural effects which persist today.[324]

Indirect structural discrimination is pervasive in virtually every society. Here are two well known examples affecting the United States. Drug laws and drug policies, despite being facially neutral, have the effect of disproportionately affecting minority groups, especially Black people.[325] Schools in high-income neighborhoods tend to be better funded (since public schools are funded primarily through property taxes) and attract more qualified teachers, transmitting an educational advantage to children of higher-income parents.

Other factors are even less tangible yet no less serious in terms of their effects, such as cultural norms and stereotypes. In the case study of gender bias in Berkeley graduate admissions in Chapter 4, we encountered the hypothesis that societal stereotypes influence people's career choices in a way that reproduces gender inequalities in income and status:

> The bias in the aggregated data stems ... apparently from prior screening at earlier levels of the educational system. Women are shunted by their socialization and education toward fields of graduate study that are generally more crowded, less productive of completed degrees, and less well funded, and that frequently offer poorer professional employment prospects.

Organizational factors

Organizational factors operate at the level of organizations or decision-making units: how they are structured, the decision making rules and processes they put in place, and the context in which individual actors operate. Again, these lie on a spectrum between direct and indirect.

The most direct form of discrimination—excluding people from participation explicitly based on group membership—is mostly

[323] Wikipedia contributors, "Same-Sex Marriage — Wikipedia, the Free Encyclopedia," 2021, https://en.wikipedia.org/w/index.php?title=Same-sex_marriage&oldid=1027216965.

[324] Rothstein, *The Color of Law: A Forgotten History of How Our Government Segregated America* (Liveright Publishing, 2017).

[325] Fellner, "Race, Drugs, and Law Enforcement in the United States," *Stan. L. & Pol'y Rev.* 20 (2009): 257.

unlawful in liberal democracies. However, practices such as lack of disability accommodations and failure to combat sexual harassment are rampant. A more indirectly discriminatory policy is the use of employees' social networks in hiring, an extremely common practice. One observational study found that the use of employee referrals in predominantly White firms reduced the probability of a Black hire by nearly 75% relative to the use of newspaper ads.[326] The study controlled for spatial segregation, occupational segregation, city, and firm size.

[326] Mouw, "Are Black Workers Missing the Connection? The Effect of Spatial Distance and Employee Referrals on Interfirm Racial Segregation," *Demography* 39, no. 3 (2002): 507–28.

Organizational discrimination can be revealed and addressed at the level of a single organization, unlike structural factors (e.g. no individual school is responsible for teachers being attracted to schools in high-income neighborhoods).

Interpersonal factors

Interpersonal factors refer to the attitudes and beliefs that result in discriminatory behavior by individuals. Sometimes people may discriminate because of an overt animus for a certain group, in the sense that the discriminator does not attempt to justify it by any appeal to rationality.

More often, the mechanisms involved are relatively indirect. A 2015 study found that academic fields in which achievement is believed to be driven by innate brilliance exhibit a greater gender disparity, i.e., they have fewer women.[327] The authors propose that the disparity is caused by the combination of the belief in the importance of innate brilliance together with stereotypes about lower innate brilliance in women. This combination could then impact women in brilliance-emphasizing disciplines in two ways: either by practitioners of those disciplines exhibiting biases against women, or by women internalizing those stereotypes and self-selecting out of those disciplines (or performing more poorly than they otherwise would). The authors don't design tests to distinguish between these competing mechanisms. However, they do test whether the observed disparities could alternatively be caused by actual innate differences (rather than beliefs in innate differences) in ability or aptitude, or willingness to work long hours. Using various proxies (such as GRE score for innate ability), they argue that such competing explanations cannot account for the observed differences.

[327] Leslie et al., "Expectations of Brilliance Underlie Gender Distributions Across Academic Disciplines," *Science* 347, no. 6219 (2015): 262–65.

One may wonder: can we not test for innate differences more rigorously, such as by examining young children? A follow-up study showed that children as young as six tend to internalize gendered stereotypes about innate brilliance, and these stereotypes influence their selection of activities.[328] These difficulties hint at the under-

[328] Bian, Leslie, and Cimpian, "Gender Stereotypes about Intellectual Ability Emerge Early and Influence Children's Interests," *Science* 355, no. 6323 (2017): 389–91.

lying complexity of the concept of gender, which is produced and reinforced in part through these very stereotypes.[329]

To recap, we've discussed structural, organizational, and interpersonal discrimination, and the fact that these are often indirect and pervasive. The three levels are interconnected: for example, in the Uber case study, structural inequalities don't perpetuate themselves, but rather through organizational decisions; those decisions at Uber are made by individuals whose worldviews are shaped by culture. In other words, even structural discrimination is actively perpetuated, and we collectively have the power to mitigate it and to reverse course. It would be a mistake to resign ourselves to viewing structural discrimination as simply the way the world is.

Notice that adopting statistical decision making is not automatically a way out of any of these factors, which operate for the most part in the background and not at a single, discrete moment of decision making.

Exercise. Review each of the examples of discrimination above and discuss whether and how the adoption of statistical decision making would change the discriminatory effect.

Exercise. In the Uber case study, what are the structural, organizational, and interpersonal factors that produce gender discrimination?

[329] West and Zimmerman, "Doing Gender," *Gender & Society* 1, no. 2 (1987): 125–51.

The persistence and magnitude of inequality

Formal equality under the law primarily addresses direct discrimination and has relatively little effect on indirect discrimination, whether structural, organizational, or interpersonal. This is one reason why inequality can be persistent in societies that seemingly promise equal opportunity. Here are two stark examples of how long inequalities can sustain themselves.

Beginning in 1609, Jesuit missions were established in the Guaraní region of South America that overlaps modern day Argentina, Paraguay, and Brazil. In addition to religious conversion, the missionaries undertook educational efforts among the indigenous people. However, due to political upheaval in Spain and Portugal, the missions abruptly ended in 1767-68 and the missionaries were expelled. How long after this date would we expect the geographic inequalities introduced by Jesuit presence to persist? Perhaps a generation or two? Remarkably, the Jesuit effect on educational attainment has been found to persist 250 years later: areas closer to a former Mission have 10-15% higher literacy rates as well as 10% higher incomes.[330] Another study of the long-run persistence of inequality shows the present-day effects of a system of colonial forced labor in Peru in Bolivia between 1573 and 1812.[331]

[330] See (Valencia Caicedo, "The Mission: Human Capital Transmission, Economic Persistence, and Culture in South America," *The Quarterly Journal of Economics* 134, no. 1 (2019): 507–56). The study makes use of a clever idea to argue that the mission locations were essentially random, making this a natural experiment, and includes various checks to rule out alternative explanations.

[331] Dell, "The Persistent Effects of Peru's Mining Mita," *Econometrica* 78, no. 6 (2010): 1863–903.

More evidence for the long-run persistence of inequality comes from the city of Florence, based on a unique dataset containing tax-related data for all individuals from the year 1427. A working paper finds that surnames associated with wealthier individuals in the dataset are associated with wealthier individuals today, six hundred years later.[332]

While these are just a few examples, research shows that persistence of inequality over generations along social and geographic lines is the norm. Yet it is not widely appreciated. For example, Americans believe that an individual born into the bottom quintile of the income distribution has a 1-in-6 chance of rising to the top quintile but the observed likelihood is 1-in-20.[333] Mobility in the U.S. has decreased since the 1980s, and is lower for Black Americans than White Americans.[334]

These inequalities are significant because of their magnitude in addition to their persistence. Median income of Black Americans is about 65% that of White Americans.[335] Wealth inequality is much more severe: the median wealth of Black households is about 11% that of White households. Turning to gender, full-time, year-round working women earned 80% of what their male counterparts earned.[336] Geographic inequalities also exist. For example, the richest and poorest census tracts in the United States differ in average income by a factor of about 30.[337]

[332] Barone and Mocetti, "Intergenerational Mobility in the Very Long Run: Florence 1427-2011," *Bank of Italy Temi Di Discussione (Working Paper) No* 1060 (2016).

[333] Davidai and Gilovich, "Building a More Mobile America—One Income Quintile at a Time," *Perspectives on Psychological Science* 10, no. 1 (2015): 60–71.

[334] Chetty et al., "Race and Economic Opportunity in the United States: An Intergenerational Perspective," *The Quarterly Journal of Economics* 135, no. 2 (2020): 711–83.

[335] Kochhar and Cilluffo, "Key Findings on the Rise in Income Inequality Within America's Racial and Ethnic Groups," *Pew Research Center*, 2018.

[336] Semega, Fontenot, and Kollar, "Income and Poverty in the United States: 2016," *Current Population Reports*, no. P60–259 (2017).

[337] Pendall and Hedman, "Worlds Apart: Inequality Between America's Most and Least Affluent Neighborhoods," *Urban Institute*, 2015.

Machine learning and structural discrimination

For a book about machine learning, we've covered a lot of ground on discrimination and inequality in society. There's a reason. To understand fairness, it isn't enough to think about the moment of decision making. We also need to ask: what impact does the adoption of machine learning by decision makers have in long-lasting cycles of structural inequality in society? Does it help us make progress toward enabling equality of opportunity, or other normative ideals, over the course of people's lives? Here are some observations that can help answer those questions.

Predictive systems tend to preserve structural advantages and disadvantages

Predictive systems tend to operate within existing institutions. When such institutions perpetuate inequality due to structural factors, predictive systems will only reify those effects, absent explicit intervention. Predictive systems tend to inherit structural discrimination because the objective functions used in predictive models usually

reflect the incentives of the organizations deploying them. As an example, consider a 2019 study found strong racial bias in a system used to identify patients with a high risk of adverse health outcomes, in the sense that Black patients were assigned lower scores compared to equally at-risk White patients.[338] The authors found that this happened because the model was designed to predict healthcare costs instead of needs, and the healthcare system spends less caring for Black patients than White patients even when they have the same health conditions.

[338] Obermeyer et al., "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations."

Suppose a firm makes hiring decisions based on a model that predicts job performance based on educational attainment. Imagine a society where students from higher-income families, on average, have had better educational opportunities that translate to greater job skills. This is not a measurement bias in the data that can be corrected away: education level genuinely predicts job performance. Thus, an accurate predictive system will rank higher-income candidates higher on average.

The structural effect of such systems become clear when we imagine every employer applying similar considerations. Candidates with greater educational opportunities end up with more desirable jobs and higher incomes. In other words, predictive systems have the effect of transferring advantages from one phase of life to the next, and one generation to the next.

This phenomenon shows up in less obvious ways. For instance, online ad targeting is based on the assumption that differences in past behavior between users reflect differences in preferences. But they might also result from differences in structural *circumstances*, and there is no way for targeting engines to tell the difference. This helps explain why ads, including job ads, may be targeted in ways that reinforce stereotypes and structural discrimination.[339]

[339] (Merill, "Does Facebook Still Sell Discriminatory Ads?" (The Markup, 2020)). See, in particular, the quotes from David Brody.

This aspect of predictive systems is amplified by compounding injustice.[340] That is, individuals are subject to a series of decisions over the course of their lives, and the effects of these decisions both accumulate and compound over time. When a person receives (or is denied) one opportunity, they are likely to appear more (or less) qualified at their next encounter with a predictive system.

[340] Gandy, *Coming to Terms with Chance: Engaging Rational Discrimination and Cumulative Disadvantage* (Routledge, 2016); Hellman, "Sex, Causation, and Algorithms: How Equal Protection Prohibits Compounding Prior Injustice," *Washington University Law Review* 98, no. 2 (2020): 481–523.

Machine learning systems may make self-fulfilling predictions

Suppose we find that chess skill is correlated with productivity among software engineers. Here are a few possible explanations: 1. Chess skill makes one a better software engineer. 2. There are underlying cognitive skills that make one better at both. 3. College professors hold stereotypes about chess skills and software engi-

neering, and steered students good at chess into computer science courses. 4. People with more leisure time were both able to pursue chess as a hobby and devote time to improving their software engineering skills.

Standard supervised learning does not distinguish between these causal paths. Regardless of the correct causal explanation, once a large swath of employers start using chess skill as a hiring criterion, they contribute to the perpetuation of the observed correlation. That is because applicants who are better at chess will have better opportunities for software engineering positions in this world, and these opportunities will allow them to develop their software engineering skills.

Machine learning automates the discovery of correlations such as the above. When we deploy those correlations as decision criteria, we alter the very phenomena that we are supposedly measuring. In other words, using non-causal variables as decision criteria may give them causal powers over time. This is not limited to machine learning: sociologists have long recognized that stereotypes that are used to justify discrimination may in fact be produced by that discrimination.[341]

[341] Myrdal, *An American Dilemma: The Negro Problem and Modern Democracy, Volume 2* (Routledge, 2017).

Algorithmic recommendation systems may contribute to segregation

Even small preferences for homogeneous neighborhoods can lead to dramatic large-scale effects. In the Appendix, we discuss a toy model of residential segregation showing such effects. But what about the online world, e.g., online social networks? The phenomenon of people making friends with similar others (online or offline) is called homophily.

In the early days of social media, there was a hope—now seen as naive—that in the online sphere there would be no segregation due to the ease with which people can connect with each other. Instead, we observe similar patterns of homophily and segregation online as offline. This is partly because real-world relationships are reflected online, but in part it is because segregation emerges through our online preferences and behaviors.[342]

[342] Boyd, "White Flight in Networked Publics: How Race and Class Shaped American Teen Engagement with MySpace and Facebook. Nakamura l, Chow-White PA, Eds. Race After the Internet," *Race After the Internet*, 2012, 203–22.

As social media has matured, concerns arising from homophily have expanded from demographic segregation to ideological echo chambers. The causal mechanisms behind polarized online discourse and the role of recommendation algorithms are being researched and debated (see the Testing chapter), but there is no doubt that online media can have structural effects.

Machine learning may lead to homogeneity of decision making

If a company hires only people whose names begin with certain letters of the alphabet, it may seem absurd but not necessarily a cause for alarm. One reason behind this intuition is that we expect that the effect of any such idiosyncratic policies will cancel out, given that job candidates have many firms to apply to. If, on the other hand, every employer adopted such a policy, then the experience of job seekers becomes radically different.

Machine learning results in more homogeneous decision making compared to the vagaries of individual decisions. Studies of human behavior show that human decisions have a lot of "noise."[343] Removing the noise is one of the main attractions of statistical decision making. But there are also risks. If statistical decision making results in similar decisions being made by many decision makers, otherwise idiosyncratic biases could become amplified and reified to the point where they create structural impediments.[344]

Homogeneity can happen in many ways. At a high level, if many machine learning systems use the same training data and the same target variable, they will make roughly the same classifications, even if the learning algorithms are very different. Intuitively, if this *weren't* the case, one could make more accurate classifications by ensembling their predictions. For a stark illustration of homogeneous predictions from the domain of predicting life outcomes, see the Fragile Families Challenge.[345]

Alternatively, many decision makers could use the same underlying system. There are anecdotes of job seekers being repeatedly screened out of jobs on the basis of personality tests, all offered by the same vendor.[346]

Even *individual* algorithmic systems may have such an outsized influence in society that their policies may have structural effects. The most obvious example are systems adopted by the state, such as a predictive policing system that leads to the overpolicing of low-income neighborhoods.

But it is private platforms, especially those with a global scale, where this effect has been most prominent. Take content moderation: a small number of social media companies together determine which types of speech can be a part of mainstream online discourse and which communities are able to mobilize online. Platform companies have faced criticism for allowing content that incites violence and, conversely, for being overzealous in deplatforming individuals or groups.

In some cases, platform policies are shaped by the capabilities and limitations of machine learning.[347] For example, algorithms are

[343] See (Kahneman et al., "Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Makinghttps," *Harvard Business Review*, 2016). The article makes both a descriptive claim about the inconsistency of human decisions as well as a normative claim that inconsistent decision making is poor decision making. The latter claim can be contested along many lines, one of which we pursue here.

[344] Creel and Hellman, "The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems," *Virginia Public Law and Legal Theory Research Paper*, no. 2021–13 (2021).

[345] Salganik et al., "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration," *Proceedings of the National Academy of Sciences* 117, no. 15 (2020): 8398–8403.

[346] O'Neil, "How Algorithms Rule Our Working Lives," *The Guardian* 16 (2016).

[347] Gillespie, "Content Moderation, AI, and the Question of Scale," *Big Data & Society* 7, no. 2 (2020): 2053951720943234.

relatively good at detecting nudity but relatively poor at detecting context. Companies such as Facebook have had broad bans on nudity without much attention to context, often taking down artwork and iconic historical images.

Machine learning shifts power

Like all technologies, machine learning shifts power. To make this more precise, we analyze the adoption of machine learning by a bureaucracy. We don't mean the term bureaucracy in its colloquial, pejorative sense of an inefficient, rule-bound government agency. We rather use the term as social scientists do: a bureaucracy is a public or private entity in which highly-trained workers called bureaucrats, operating in a hierarchical structure, make decisions in a way that is constrained by rules and policies but also requires expert judgment. Firms, universities, hospitals, police forces, and public assistance programs are all bureaucracies to various degrees. Most of the decision making scenarios that motivate this book are situated in bureaucracies.

To understand the effect of adopting machine learning, we consider five types of stakeholders: decision subjects, the people who provide the training data, domain experts, machine learning experts, and policy makers. Our analysis builds on a talk by Kalluri.[348]

[348] Kalluri, "The Values of Machine Learning" (NeurIPS Queer in AI workshop, 2019).

Machine learning as generally implemented today shifts power away from the first three categories. By representing decision subjects as standardized feature vectors, statistical decision making removes their agency and ability to advocate for themselves. In many domains, notably the justice system, this ability is central to the rights of decision subjects. Even in a relatively less consequential domain such as college admissions, the personal statement provides this ability and is a key component of the evaluation.

People who provide training data may have *knowledge* about the task at hand, but provide only their *behavior* as input to the system (think of email recipients clicking the "spam" button). Machine learning instead constructs a form of knowledge in a centralized way. In contrast, domain experts learn in part from the knowledge and lived experience of the individuals they interact with. Admittedly, experts such as physicians are often criticized for devaluing the knowledge and experience of decision subjects (patients). But the fact that such a debate is happening at all is evidence of the fact that patients have at least some power in the traditional system.

The role of domain experts is also more limited compared to traditional decision making where the discretion and judgment of such experts holds sway. In machine learning applications, domain

experts have two main roles: formulating the problem and task, and labeling training examples.

These effects are not always harmful. In government bureaucracies, the power wielded by "street-level bureaucrats" such as police officers and social service caseworkers—the people who translate policy into individual decisions—can be abused, and removing their discretion is often seen as a fairness intervention. Yet the discretion and human intelligence of these decision makers can also be a vital fairness-promoting element due to the existence of extenuating factors or novel circumstances not seen in the training data or covered in existing policies.[349]

Machine learning experts, of course, tend to have a central role. Stakeholders' requirements have to be translated into implementation by these experts; whether intentionally or unintentionally, there are often substantial gaps between the desired policy and the policy that's realized in practice.[350] In every automated system, there is something lost in the translation of policy from human language to computer code. For example, there have been cases where software miscalculated prison inmates' eligibility for early release, with harrowing consequences including being held in prison too long and being returned to prison after being released.[351] But in those classic automated systems, these gaps tend to be mistakes that are generally obvious upon manual inspection (not that it is any comfort to those who are harmed). But when machine learning is involved, the involvement of the expert is often necessary even to recognize that something has gone wrong. This is because the policy tends to be more ambiguous (what does "high risk" mean?) and because deviations from the policy become apparent only in aggregate.

Finally, machine learning empowers policy makers or centralized decision makers. Consider a risk prediction tool used by a child protection agency to screen calls. Depending on the agency's budget and other factors, the decision maker may want to screen in a higher or lower proportion of calls. With a statistical tool, such a policy change can be implemented instantly, and is enormously easier than the alternative of retraining hundreds of case workers to adjust their mental heuristics. This is just one example that illustrates why such tools have proven so attractive to those who make the decision to deploy them.

Structural interventions for fair machine learning

The fact that machine learning may contribute to structural discrimination motivates the need for interventions that are similarly broad in scope. We call these structural interventions: changing the way ma-

[349] Lipsky, *Street-Level Bureaucracy: Dilemmas of the Individual in Public Service* (Russell Sage Foundation, 2010); Alkhatib and Bernstein, "Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, 1–13.

[350] Citron, "Technological Due Process," *Wash. UL Rev.* 85 (2007): 1249.

[351] "US prisoners released early by software bug" (BBC news, 2015).

chine learning gets built and deployed. The changes we have in mind go beyond the purview of any single organization, and require collective action. This could take the form of a broad social movement, or other collectives including communities, workers, researchers, and users.

Reforming the underlying institutions

One approach is to focus on the underlying institution rather than the technology, and change it so that it is less prone to adopt harmful machine learning tools in the first place. For example, shifting the focus of the criminal justice system from incapacitation to rehabilitation could decrease the demand for risk prediction tools.[352] Many scholars and activists distinguish between reform and abolition (sometimes called non-reformist reform), abolition being a more radical and transformative approach.[353] For our purposes, however, they both have the effect of centering the intervention on the institution rather than the technology.

In many domains, the very purposes and aims of our institutions remain contested. For example, what are the goals of policing? Commonly accepted goals include deterrence and prevention of crime, ensuring public safety and minimizing disorder, and bringing offenders to justice; they might also include broader efforts to improve the health and vitality of communities. The relative importance of these goals varies between communities and over time. Thus, formulating police allocation decisions as an optimization problem, as predictive policing systems do, involves taking positions on these deeply contested issues.

History shows us that many institutions that may feel like fixtures of modern society, such as higher education, have in fact repeatedly redefined their goals and purposes to adapt to a changing world. In fact, sometimes the impetus for such shifts was to *more effectively discriminate*. In the early twentieth century, elite American universities morphed from treating size (in terms of enrollment) as a source of prestige to selectivity. A major reason for this change was to curtail the rising proportion of Jewish students without having to introduce explicit quotas; the newfound mission of being selective enabled them to emphasize traits like character and personality in admissions, which in turn allowed much leeway for discretion. In fact, this system that Harvard adopted in 1926 was the origin of the holistic approach to admissions that continues to be contentious today.[354]

Some scholars have gone beyond the position that intervention to address algorithmic harms should focus on the underlying institution, and argued that the adoption of automated decision making

[352] Barabas et al., "Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment," in *Conference on Fairness, Accountability and Transparency* (PMLR, 2018), 62–76.

[353] (Akbar, "An Abolitionist Horizon for Police (Reform)," *California Law Review* 108, no. 6 (2020)). See also Chapter 5 of (Benjamin, *Race After Technology*) which includes a discussion of tech tools for resisting oppressive institutions.

[354] Karabel, *The Chosen: The Hidden History of Admission and Exclusion at Harvard, Yale, and Princeton* (Houghton Mifflin Harcourt, 2005).

actually enables resistant institutions to stave off necessary reform. Virginia Eubanks examines four public assistance programs for poor people in the United States—food assistance, Medicaid, homelessness, and at-risk children.[355] In each case there are eligibility criteria administered automatically, some of which use statistical techniques. The book documents the harmful effects of these systems, including the punitive effects on those deemed ineligible; the disproportionate impact of those burdens on low-income people of color, especially women; the lack of transparency and seeming arbitrariness of the decisions; and the tracking and surveillance of the lives of poor that is necessary for these systems to operate.

[355] Eubanks, *Automating Inequality*.

These problems may be fixable to some extent, but Eubanks has a deeper critique: that these systems distract from the more fundamental goal of eradicating poverty ("We manage the individual poor in order to escape our shared responsibility for eradicating poverty"). In theory, the two approaches may coexist. In practice, Eubanks argues, these systems legitimize the idea that there is something wrong with some people, hide the underlying structural problem, and foster inaction. They also incur a high monetary cost that could otherwise be put toward more fundamental reform.

Community rights

Harmful technologies are often legally justified under a notice-and-consent framework which rests on an individualistic conception of rights and is ill-equipped to address collective harms. For example, police departments obtain footage en masse from residential security cameras with the consent of residents through centralized platforms like Amazon Ring.[356] However, consent is not a meaningful check in this scenario, because the people who stand to be harmed by police abuse of surveillance footage—such as protesters or members of racial minorities who had the police called on them for "acting suspiciously"—are not the ones whose consent is sought or obtained.

[356] Harwell, "Doorbell-Camera Firm Ring Has Partnered with 400 Police Forces, Extending Surveillance Concerns," *Washington Post*, 2019.

This gap is especially salient in machine learning applications: even if a classifier is trained on data provided with consent, it may be applied to nonconsenting decision subjects. An alternative is to allow groups, such as geographic communities, the right to collectively consent to or reject the adoption of technology tools. In response to the police use of facial recognition, civil liberties activists advocated for a community right to reject such tools; the success of this advocacy has led to various local bans and moratoria.[357] In contrast, consider online targeted advertising, another technology that has faced widespread dissent. In this case, there are no analogous collectives who can organize effective resistance, and hence attempts to

[357] Whittaker et al., *AI Now Report 2018* (AI Now Institute at New York University New York, 2018).

reject the technology have been much less successful.[358]

Beyond collective consent, another goal of community action is to obtain a seat at the table in the design of machine learning systems as stakeholders and participants whose expertise and lived experience shapes the conception and implementation of the system rather than mere data providers and decision subjects. Among other benefits, this approach would make it easier to foresee and mitigate representational harms—issues such as demeaning categories in computer vision datasets or image search results that represent offensive stereotypes. But there are also potential risks to participatory design: it may create further burdens for members of underrepresented communities, and it may act as a smokescreen for organizations resisting meaningful change. It is essential that participation be recognized as labor and be fairly compensated.[359]

[358] Barocas and Levy, "Privacy Dependencies," *Wash. L. Rev.* 95 (2020): 555.

[359] Sloane et al., "Participation Is Not a Design Fix for Machine Learning," *arXiv Preprint arXiv:2007.02423*, 2020.

Regulation

Regulation that promotes fair machine learning can take the form of applying existing laws to decision systems that incorporate machine learning, or laws that specifically address the use of technology and its attendant harms. Examples of the latter include the above-mentioned bans on facial recognition, and restrictions on automated decision making under the European Union's General Data Protection Regulation (GDPR). Both flavors of regulation are evolving in response to the rapid adoption of machine learning in decision making systems. Regulation is a major opportunity for structural intervention for fair machine learning. Yet, because of the tendency of law to conceptualize discrimination in narrow terms, its practical effect on curbing harmful machine learning largely remains to be seen.[360]

The gap between the pace of adoption of machine learning and the pace of law's evolution has led to attempts at self-regulation: a 2019 study found 84 AI ethics guidelines around the world.[361] Such documents don't have the force of law but attempt to shape norms for organizations and/or individual practitioners. While self-regulation has been effective in some fields such as medicine, it is doubtful if AI self-regulation can address the thorny problems we have identified in this chapter. Indeed, industry self-regulation generally aims to forestall actual regulation and the structural shifts it may necessitate.[362]

[360] Hoffmann, "Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse," *Information, Communication & Society* 22, no. 7 (2019): 900–915.

[361] Jobin, Ienca, and Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence* 1, no. 9 (2019): 389–99.

[362] For a deeper critique of industry-led statements of principles see (Greene, Hoffmann, and Stark, "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019).

Workforce interventions

Machine learning shifts power to machine learning experts, which makes the ML workforce an important locus of interventions. One set

of efforts is aimed at enabling more people to benefit from valuable job opportunities in the industry[363] and to fight imbalances of power within the workforce—notably, between technology experts and those who perform other roles such as annotation.[364] Another set of efforts seeks to align the uses of ML with ethical values of the ML workforce. The nascent unionization movement in technology companies seems to have both objectives.

While a more diverse workforce is morally valuable for its own sake, it is interesting to ask what effect it has on the fairness of the resulting products. One experimental study of programmers found that the gender or race of programmers did not impact whether they produced biased code.[365] However, this is a lab study and should not be seen as a guide to the effects of structural interventions. For example, one causal path by which workforce diversity could impact products (not captured in the study's design) is that a team with a diversity of perspectives may be more willing to ask critical questions about whether a product should be built or deployed.

Another workforce intervention is education and training. Ethics education for computer science students is on the rise, and a 2018 compilation included over 200 such courses.[366] A long-standing debate is about the relative merits of stand-alone courses and integration of ethics into existing computer science courses.[367] Professional organizations such as the Association for Computing Machinery (ACM) have had codes of ethics for several decades, but it's unclear if these codes have had a meaningful impact on practitioners.

In many professional fields including some engineering fields, ethical responsibilities are enforced in part through licensing of practitioners. Professionals such as doctors and lawyers must master a body of professional knowledge, including ethical codes, are required by law to pass standardized exams before being licensed to practice, and may have that license revoked if they commit ethical transgressions. This is not the case for software engineering. At any rate, the software engineering certification standards that do exist[368] have virtually no overlap with the topics in this book.

[363] Judd, "Activities for Building Understanding: How Ai4all Teaches AI to Diverse High School Students," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 2020, 633–34.

[364] "Announcing the Contract Worker Disparity Project" (Tech Equity Collaborative, 2021).

[365] Cowgill et al., "Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing Ai Ethics," in *Proceedings of the 21st ACM Conference on Economics and Computation*, 2020, 679–81.

[366] Fiesler, Garrett, and Beard, "What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis," in *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 2020, 289–95.

[367] Martin et al., "Implementing a Tenth Strand in the CS Curriculum," *Communications of the ACM* 39, no. 12 (1996): 75–84.

[368] Wikipedia contributors, "Certified Software Development Professional — Wikipedia, the Free Encyclopedia," 2021, https://en.wikipedia.org/w/index.php?title=Certified_Software_Development_Professional&oldid=1021729741.

The research community

The machine learning research community is another important locus for reform and transformation. The most significant push for change has been the ongoing fight for treating research topics such as fairness, ethics, and justice as legitimate and first-rate. Traditionally, a few topics in machine learning such as optimization algorithms have been considered "core" or "real" machine learning, and other topics—even dataset construction—seen as peripheral and less intellectually

serious.

A few other key debates: should all machine learning researchers be required to reflect on the ethics of their research?[369] Is there too much of a focus on fixing bias as opposed to deeper questions about power and justice?[370] How to center the perspectives of people and communities affected by machine learning systems? What is the role of industry research on fair machine learning given the conflicts of interest?

[369] Nanayakkara, Hullman, and Diakopoulos, "Unpacking the Expressed Consequences of AI Research in Broader Impact Statements," *arXiv Preprint arXiv:2105.04760*, 2021.

[370] Kasy and Abebe, "Fairness, Equality, and Power in Algorithmic Decision-Making," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, 576–86.

Organizational interventions for fairer decision making

The structural interventions we've discussed above require social movements or other collective action and have been evolving on a timescale of years to decades. This is not to say that an organization should throw up its hands and wait for structural shifts. A plethora of interventions are available to most types of decision makers. This section is an overview of the most important ones.

As you read, observe that the majority of interventions attempt to improve outcomes for all decision subjects rather than viewing fairness as an inescapable tradeoff. One reason this is possible is that many of them don't operate at the moment of decision. Note, also, that evaluating the effects of interventions—whether with respect to fairness or other metrics—generally requires causal inference. Finally, only a small subset of potential fairness interventions can be implemented in the framework of machine learning. The others focus on organizational or human practices rather than the technical subsystem involved in decision making.

Type	**Intervention**	**Example**
Modifying the outputs	Reallocation	Group-specific decision thresholds
Modifying the decision process	Combatting interpersonal discrimination	Implicit bias training
	Formalization	Adopting statistical decision making
	Procedural protections	Explanation and recourse
Before the decision	Outreach	Sending mailers about scholarships
	Intervening on causal factors	Job training, preventive health
After the decision	Modifying the environment	Helping defendants show up to court

Figure 34: A summary of major types of organizational interventions

Redistribution or reallocation

Redistribution and reallocation are terms that refer to interventions that modify a decision-making process to introduce an explicit preference for one or more groups, usually groups considered to be disadvantaged. When we talk about fairness interventions, this might be the kind that most readily comes to mind.

When applied to selection problems where there is a relatively static number of slots, as is typical in hiring or college admissions, a plethora of algorithmic fairness interventions reduce to different forms of reallocation. This includes techniques such as adding a fairness constraint to the optimization step, or a post-processing adjustment to improve the scores of the members of the disadvantaged groups. This is true regardless of whether the goal is demographic parity or any other statistical criterion.

Reallocation is appealing because it doesn't require a causal understanding of why the disparity arose in the first place. By the same token, reallocation is a crude intervention. It is designed to benefit a group—and it has the advantage of providing a measure of transparency by allowing a quantification of the group benefit—but most reallocation procedures don't incorporate a notion of deservingness of members within that group. Often, reallocation is accomplished by a uniform preference for members of the disadvantaged group. Alternatively, it may be accomplished by tinkering with the optimization objective to incorporate a group preference. In this approach, distributing the fruits of reallocation within the group is delegated to the model, which may end up learning a non-intuitive and unintended allocation (for example, an intersectional subgroup may end up further disadvantaged compared to a no-intervention condition). At best, reallocation methods will aim to ensure that relative ranking within groups is left unchanged.

As crude as reallocation is, another intervention with an even worse tradeoff is to omit features correlated with group identity from consideration. To be clear, if the feature is statistically, causally, or morally irrelevant, that may be a good reason for omitting it (Chapter 3). But what if the feature is in fact relevant to the outcome? For example, suppose that people who contribute to open-source software projects tend to be better software engineers. This effect acts through a morally relevant causal path because programmers obtain useful software-engineering skills through open-source participation. Unfortunately, many open-source communities are hostile and discriminatory to women and minorities (this is perhaps because they lack the formal organizational structures that firms use to keep interpersonal discrimination in check to some degree). Recognizing this bias, a

software company could either explicitly account for it in hiring decisions or simply omit consideration of open-source contributions as a criterion. If it does the latter, it ends up with less qualified hires on average; it also disadvantages the people who braved discrimination to develop their skills, arguably the most deserving group.

Omitting features based on statistical considerations without a moral or causal justification is extremely popular in practice because it is simple to implement, politically palatable, and avoids the legal risk of disparate treatment.

Combatting interpersonal discrimination

Rather than intervene directly on the outputs, organizations can try to improve the process of decision making. In many cases, discriminators are surprisingly candid about their prejudices in surveys and interviews.[371] Can they perhaps be trained out of their implicit or overt biases? This is the idea behind prejudice reduction, often called diversity training.

[371] Neckerman and Kirschenman, "Hiring Strategies, Racial Bias, and Inner-City Workers."

But does diversity training work? Paluck & Green conducted a massive review of nearly a thousand such interventions in 2009.[372] The interventions include promoting contact with members of different groups, recategorization of social identity, explicit instruction, consciousness raising, targeting emotions, targeting value consistency and self-worth, cooperative learning, entertainment (reading, media), discussion and peer influence. Unfortunately, only a small fraction of the published studies reported on field experiments; Paluck & Green are dubious about both observational field studies and lab experiments. Overall, the field experiments don't provide much support for the effectiveness of diversity interventions. That said, there were many promising lab methods that hadn't yet been tested in the field. A more recent review summarizes the research progress from 2007 to 2019.[373]

[372] Paluck and Green, "Prejudice Reduction: What Works? A Review and Assessment of Research and Practice," *Annual Review of Psychology* 60 (2009): 339–67.

[373] Paluck et al., "Prejudice Reduction: Progress and Challenges," *Annual Review of Psychology* 72 (2020).

Minimizing the role of human judgment via formalization

Approaches like implicit bias training seek to improve the judgment of human decision makers, but ultimately defer to that judgment. In contrast, formalization aims to curb judgment and discretion.

The simplest formalization technique is to withhold the decision subject's identity (or other characteristics considered irrelevant) from the decision maker. Although this idea dates to antiquity, in many domains the adoption of anonymous evaluation is a recent phenomenon and has been made easier by technology.[374] Two major limitations of this approach are the ubiquitous availability of proxies and the fact that anonymization is not feasible in many contexts such

[374] Chohlas-Wood et al., "Blind Justice: Algorithmically Masking Race in Charging Decisions" (Technical report, 2020).

as in-person hiring interviews.[375]

A more ambitious approach is rule-based or statistical decision making that removes human discretion entirely. For example, removing lender discretion in loan underwriting was associated with a nearly 30% increase in the approval rates of minority and low-income applicants, while at the same time increasing predictive accuracy (of the risk of default).[376] Human decision makers tend to selectively ignore credit history irregularities of White applicants.[377]

In some ways, machine learning can be seen as a natural progression of the shift from human judgment to rule-based decision making. In machine learning, the discovery of the rule—and not just its application—is deferred to the data and implemented by an automated system. Based on this, one might naively hope that machine learning will be even more effective at minimizing discrimination.

However, there are several counterarguments. First, claims of the superiority of statistical formulas over human judgment, at least in some domains, have been questioned as being based on apples-to-oranges comparisons because the human experts did not view their role as pure prediction. For example, judges making sentencing decisions may consider the wishes of victims, and may treat youth as a morally exculpatory factor deserving of leniency.[378] Second, there has been a recognition of all the ways in which machine learning can be discriminatory, which is of course a central theme of this book. Third, there are numerous potential drawbacks such as a loss of explainability and structural effects that are not captured by the human-machine comparisons.

Perhaps most significantly, incomplete formalization can simply shift the abuse of discretion elsewhere. In Kentucky, the introduction of pretrial risk assessment *increased* racial disparities for defendants with the same predicted risk. The effect appears to be partly because of differential adoption of risk assessment in counties with different racial demographics, and partly because even the same judges are more likely to override the recommended decision for Black defendants compared to White defendants.[379] In Ontario, social service caseworkers described how they manipulate the inputs to the automated system to get the outcomes they want.[380,381] In Los Angeles, police officers used many strategies to resist being managed by predictive policing algorithms.[382]

The most pernicious effect of formalization as a fairness intervention is that it may shift discretion to earlier stages of the process making bias *harder* to mitigate. Examples abound. Mandatory minimum sentencing guidelines for drug possession in the United States in the 1980s were justified in part as a way to combat judges' biases and arbitrariness,[383] but are now widely recognized as overly puni-

[375] Even in these contexts, blinding of attributes that are not readily inferrable can be effective. Indeed, it is frowned upon to inquire about candidates' marital status during job interviews, and such inquiries may be treated as evidence of intent to discriminate ("Pre-Employment Inquiries and Marital Status or Number of Children" (U.S. Equal Employment Opportunity Commission, 2021)).

[376] Gates, Perry, and Zorn, "Automated Underwriting in Mortgage Lending: Good News for the Underserved?" *Housing Policy Debate* 13, no. 2 (2002): 369–91.

[377] Squires and others, *Capital and Communities in Black and White: The Intersections of Race, Class, and Uneven Development* (Suny Press, 1994).

[378] Stevenson and Doleac, "Algorithmic Risk Assessment in the Hands of Humans," *Available at SSRN 3489440*, 2019.

[379] Stevenson, "Assessing Risk Assessment in Action," *Minn. L. Rev.* 103 (2018): 303; Albright, "If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions," *Harvard John M. Olin Fellow's Discussion Paper* 85 (2019) .

[380] Raso, "Displacement as Regulation: New Regulatory Technologies and Front-Line Decision-Making in Ontario Works," *Canadian Journal of Law and Society* 32, no. 1 (2017): 75–95.

[381] Caseworkers report doing so in order to work around the limitations and non-transparency of the automated system to achieve just outcomes for clients. The difficulty of distinguishing between abuse of discretion and working around an overly rigid system further illustrates the double-edged nature of formalization as a fairness intervention.

[382] Brayne, *Predict and Surveil: Data, Discretion, and the Future of Policing* (Oxford University Press, USA, 2020).

[383] Frankel, "Criminal Sentences: Law Without Order," 1973.

tive and structurally racist. One way in which such laws can encode race is the 100-to-1 sentencing disparity between powder and crack cocaine, the popularity of the two forms of the same drug differing by income and socioeconomic status.[384] A very different kind of example comes from Google, which has had a vaunted, highly formalized process for recruiting in order to combat bias and enhance the quality of decisions.[385] But recruiters have argued that this process in fact bakes in bias because it incorporates a ranking of colleges in which Historically Black Colleges and Universities are not ranked at all.[386]

The Harvard admissions lawsuit from Chapter 4 is another case study of formalization versus holistic decision making. Plaintiffs point out that the admissions criteria include subjective assessments of personality traits such as likability, integrity, helpfulness, kindness, and courage. Harvard scored Asian-American applicants on average far lower on these traits than any other racial group. Harvard, on the other hand, argues that evaluating the "whole person" is important to identify those with unique life experiences that would contribute to campus diversity, and that a consideration of subjective traits is a necessary component of this evaluation.

[384] Palamar et al., "Powder Cocaine and Crack Use in the United States: An Examination of Risk for Arrest and Socioeconomic Disparities in Use," *Drug and Alcohol Dependence* 149 (2015): 108–16.

[385] Google re:work team, "Guide: Hire by Committee" (https://rework.withgoogle.com/print/guides/6053596147744768/, 2021).

[386] Natasha Tiku, "Google's Approach to Historically Black Schools Helps Explain Why There Are Few Black Engineers in Big Tech" (Washington Post, 2021).

Procedural protections

Diversity training and formalization are examples of procedural fairness interventions. There are many other procedural protections: notably, making the process transparent, providing explanations of decisions, and allowing decision subjects to contest decisions that may have been made in error. As we discussed above, procedural protections are more important when machine learning is involved than for other types of automated systems.

United States law emphasizes procedural fairness over outcomes. This is one reason for the great popularity of diversity training despite its questionable effectiveness.[387] When the decision maker is the government, the legal conception of fairness is even more focused on procedure. For example, there is no notion of disparate impact under United States constitutional law.

[387] Edelman, "Law at Work: The Endogenous Construction of Civil Rights," in *Handbook of Employment Discrimination Research* (Springer, 2005), 337–52.

While some procedural interventions such as diversity training have been widely adopted, many others remain rare despite their obvious fairness benefits. For example, few employers offer candid explanations for job rejection. Decision makers turning to automated systems are often looking to cut costs, and may hence be especially loath to adopt procedural protections.

There are many examples of fairness concerns with automated systems for which *only* procedural protections can be an effective

remedy (other than scrapping the system altogether). For example, Google's policy is to suspend users across its entire suite of services if they violate its terms of service. There are many anecdotal reports from users who have lost years' worth of personal and professional data, insist that Google's decision was made in error, and that Google's appeal process did not result in a meaningful human review of the decision.

Outreach

The rest of the interventions are not about changing the decision making process (or outcomes). Instead, they change something about the decision subjects or the organizational environment.

A recent study sought to address the puzzling phenomenon that low-income students tend not to attend highly selective colleges, even when their strong academic credentials qualify them for admission and despite the availability of financial aid that would make it *cheaper* to attend a selective institution.[388] The authors designed an intervention in which they sent flyers to low-income high-school students informing them about a new scholarship at the University of Michigan, and found that compared to a control group, these students were more than twice as likely to apply as well as enroll at the University. The effect was entirely due to students who would have otherwise attended less selective colleges or not attended college at all. The targets of outreach were highly qualified students identified based on standardized test scores (ACT and SAT), which allowed the university to guarantee financial aid conditional on admission. It is worth reiterating that this was a purely informational intervention: the scholarship was equally available to students in the control group, who received only postcards listing University of Michigan application deadlines.

[388] Dynarski et al., "Closing the Gap: The Effect of a Targeted, Tuition-Free Promise on College Choices of High-Achieving, Low-Income Students" (National Bureau of Economic Research, 2018).

To the extent that disparities are due to disadvantaged groups lacking knowledge of opportunities, informational interventions should decrease those disparities, but this point doesn't appear to be well-researched. For example, the Michigan study targeted the intervention at low-income students, so it doesn't address the question of whether informing *all* students would close the income gap.

Intervening on causal factors

If we understand the causal factors that lead to underperformance of some individuals or groups, we can intervene to mitigate them. Like informational interventions, this approach seeks to help all individuals rather than simply minimize disparities. This type of intervention

is extremely common. Some examples: job training programs for formerly incarcerated people to improve welfare and decrease the chances of recidivism; efforts to bolster math and science education to address an alleged labor shortage of engineers (a so-called pipeline problem); and essentially all of public health and preventive healthcare. The use of randomized controlled trials to identify and intervene on the causes of poverty has been so influential in development economics that it led to the 2019 Nobel Prize to Duflo, Banerjee, and Kremer.

In a competitive market, such as an employer competing for workers, this intervention may not pay off for an individual decision maker from an economic perspective: job seekers who have benefited from the intervention may choose to join other firms instead. Many approaches have been used to overcome this misalignment of incentives. Firms may act collectively, or the state may fund the intervention. If a firm is large enough, the overall payoffs could be so high relative to the cost of the intervention that the reputational benefit to the firm may be sufficient to justify it.

Modifying the organizational environment

If decision makers have many opportunities to intervene before the point of decision (e.g. hiring), they also have opportunities to intervene after that point to ensure that individuals fulfill their potential. If a firm finds that few minority employees are successful, it may be because the workplace is hostile and discriminatory.

In other cases, some individuals or groups may need additional accommodations to remedy past disadvantages or because of morally irrelevant differences. A few examples: remedial courses for disadvantaged students, a peer group for first-time college students, need-based scholarships, a nursing mother's room in a workplace, and disability accommodations.

Accommodation isn't simply redistribution in disguise: it does not (or need not) involve an explicit preference for the disadvantaged group. Even if the accommodation is made available to everyone, the disadvantaged group will preferentially benefit from it. This is obvious in the case of, say, disability accommodations. In other cases this is less obvious, but no less true. Even if financial aid were available to all students at a university, it would differentially benefit low-income students.

However, the actual effects of accommodations can be hard to predict and must be carefully measured empirically. A notable example comes from a study showing that men benefit from gender-neutral clock-stopping policies.[389] Such policies in universities allow both

[389] Antecol, Bedard, and Stearns, "Equal but Inequitable: Who Benefits from Gender-Neutral Tenure Clock Stopping Policies?" *American Economic Review* 108, no. 9 (2018): 2420–41.

men and women to add time to the tenure clock with the birth of a child. While they are often adopted in the interest of fairness, the study shows that they increase men's tenure rates and lower women's; this is presumably because men are able to be more productive during their extended time due to differences in child-care responsibilities or the impact of the birth itself.[390]

Here's a stark example of how organizational policies can cause people to fail and how easily they can be remedied. In New York City, there are approximately 300,000 cases of low level offenses every year. The defendants are required to appear in court;[391] if they fail to appear, arrest warrants are automatically issued. Historically, a remarkable 40% of defendants fail to appear in court. The resulting negative consequences of Failure to Appear (FTA) are both severe and unequally distributed: for instance, members of groups that are subject to overpolicing are more likely to be arrested. Remarkably, a study found that FTA rates decreased from 41% to 26% simply by redesigning the summons form to be less confusing and sending defendants text messages shortly before their court dates![392]

[390] However, note that the policy has two fairness goals: to mitigate the adverse career impact of childbirth and to decrease gender disparities in said impacts. Presumably the policy still meets the first goal even if it fails the second.

[391] For offenses of the lowest severity, the summons may be resolved by mail.

[392] Fishbane, Ouss, and Shah, "Behavioral Nudges Reduce Failure to Appear for Court," *Science* 370, no. 6517 (2020).

Summary

We looked at seven broad types of fairness interventions that organizations can deploy. The majority of these interventions potentially improve opportunity for all decision subjects as they are motivated by some underlying injustice rather than merely mitigating some disparity. In fact, interventions that aim to address an underlying justice might sometimes increase certain disparities between groups—a possibility that would be morally justified under a non-comparative notion of fairness that calls for treating each subject as they ought to be treated.[393]

Comparative notions of fairness are appealing to focus on because they are easy to quantify, but we shouldn't forget the deeper questions. A domain where this seems to have happened is algorithmic hiring. Tools used in algorithmic hiring utilize situational judgment tests, personality tests, and sometimes much more dubious techniques—increasingly involving machine learning—for screening and selecting candidates. Firms adopt such tools to cut recruitment costs, especially for low-wage positions where the cost of hiring a worker through the traditional process can be seen as significant in relation to a worker's contribution to the firm's revenue over the course of the period of their employment.

[393] Hellman, "Two Concepts of Discrimination," *Va. L. Rev.* 102 (2016): 895.

These tools are problematic for many reasons. While they aim to formalize the hiring process, they often use attributes that are morally and causally irrelevant to job performance. HireVue, for

example, previously relied on facial expressions and intonations in a person's voice as part of its automated assessment. They also fail to take a broad view of discrimination. Focusing narrowly on minimizing disparities in hiring rates across groups leaves unaddressed what kind of environment employees will encounter once hired. If job applicants from certain groups were previously predicted to perform poorly in a certain workplace, the employer should strive to understand the reasons for this difference in success, rather than simply trying to find members of these groups that might be able to succeed under such unfavorable, unwelcoming, or hostile conditions. Parity-promiting interventions change the selection process, but preserve the organizational status quo, endorsing the idea that the candidates that have been selected should be able to deal with these conditions sufficiently well to be as productive as their peers who don't face similar challenges. Other productive—and potentially less harmful—forms of intervention include on-the-job training (which might be understood as a way of intervening on causal factors), meaningful feedback for rejected applicants (which would provide some degree of procedural protection, but also help guide applicants' future investment in their own development), and a strategic approach to sourcing candidates who firms with more accurate tools might now be better able to assess.

The narrow focus on disparities can mean that there is little consideration of the quality of decisions made by the tools. Tools that simply lack validity raise a host of normative concerns. Notably, assessments that achieve approximate demographic parity but continue to suffer from accuracy disparity (also called differential validity) can set members of certain groups up for failure by expecting them to be able to perform better than they would be currently prepared to.[394]

To reiterate, we do not advocate for treating statistical fairness criteria as constraints, at least in the first instance. That approach assumes that reallocation is the only available intervention. Instead, if we treat statistical fairness criteria as diagnostics, we are likely to uncover deeper problems that require remedying. Unfortunately, these deeper remedies are also harder. They require both causal inference and normative depth. That is of course why they are often ignored, and foundational questions remain unaddressed.

A case in point: a 2021 paper analyzes the fairness of pre-trial detention in a non-comparative sense.[395] How risky does a defendant have to be so that the expected benefit to public safety justifies the harm to the defendant from detention? Using the clever approach of asking survey recipients to choose between being detained and becoming victims of certain crimes, the authors conclude that pretrial detention is essentially never justified.

[394] See also the discussion of the limitations of independence as a fairness criterion in Chapter 2.

[395] Stevenson and Mayson, "Pretrial Detention and the Value of Liberty," *Virginia Public Law and Legal Theory Research Paper*, no. 2021–14 (2021).

The study's method is sure to be debated, but the point remains that there have been relatively few principled, quantitative attempts to justify the risk thresholds used in pretrial detention. There have been many other calls to end pretrial detention based on different moral and legal arguments. When such foundational questions continue to be debated, it would be exceedingly premature to declare a risk-based pretrial detention system to be "fair" because it satisfied some statistical criterion.

Appendix: a deeper look at structural factors

Let us briefly discuss two phenomena that help explain the long-run persistence of inequality: segregation and feedback loops.

The role of segregation

A structural factor that exacerbates all of the mechanisms of discrimination we discussed is the segregation of society along the lines of group identity. Segregation arguably enables interpersonal discrimination because increased contact among groups decreases prejudice toward outgroups—the controversial contact hypothesis.[396]

At a structural level, segregation sustains inequality because an individual's opportunities for economically productive activities depend on her social capital, including the home, community, and educational environment. A strand of the economics literature has built mathematical models and simulations to understand how group inequalities—especially racial inequalities—arise and persist indefinitely even in the absence of interpersonal discrimination, and despite no intrinsic differences between groups. In the extreme case, if we imagine two or more groups belonging to non-interacting economies that grow at the same rate, it is intuitively clear that differences can persist indefinitely. If segregation is imperfect, do gaps eventually close? This is sensitive to the assumptions in the model. In Lundberg and Startz's model the gaps close eventually, although extremely slowly.[397] In Bowles et al.'s model, they don't under some conditions;[398] one reason is that the disadvantaged group might face higher costs of labor-market skill acquisition due to lower social capital.[399]

In the United States, after the civil rights legislation of the 1960s and 70s, residential segregation by race has been decreasing, albeit slowly. On the other hand, residential segregation by income appears to be increasing.[400]

[396] Paluck, Green, and Green, "The Contact Hypothesis Re-Evaluated," *Behavioural Public Policy* 3, no. 2 (2019): 129–58.

[397] Lundberg and Startz, "On the Persistence of Racial Inequality," *Journal of Labor Economics* 16, no. 2 (1998): 292–323.

[398] Bowles and Sethi, "Social Segregation and the Dynamics of Group Inequality," 2006.

[399] A pioneering work in this area is (Loury, "A Dynamic Theory of Racial Income Differences" (Discussion paper, 1976)).

[400] Massey, Rothwell, and Domina, "The Changing Bases of Segregation in the United States," *The Annals of the American Academy of Political and Social Science* 626, no. 1 (2009): 74–90.

The role of feedback loops

There is a classic economic model of feedback loops in the context of a labor market.[401] There are two groups of workers and two types of jobs: high and low skilled, with high-skilled jobs requiring certain qualifications to perform effectively. Under suitable assumptions (especially, employers cannot perfectly observe worker qualifications before hiring them, but only after providing costly on-the-job skills training) there exists an economic equilibrium in which the following feedback loop sustains itself:

1. The employer practices wage discrimination between the two groups.
2. As a result, the disadvantaged group achieves lower returns to investment in qualifications.
3. Workers, assumed to be rational, respond to such a differential by investing differently in acquiring qualifications, with one group acquiring more qualifications.
4. The employer—again, under certain rationality assumptions—wage discriminates because of the observed difference in qualifications.

The significance of this model is that it can explain the persistence of inequality (and discrimination) without assuming intrinsic differences between the groups, and without employers discriminating between equally qualified workers. It should be viewed as showing only the possibility of such feedback loops. Like any theoretical model, a claim that such a feedback loop explains some actually observed disparity would require careful empirical validation.

[401] Arrow, "The Theory of Discrimination."

8
Datasets

It's become commonplace to point out that machine learning models are only as good as the data they're trained on. The old slogan "garbage in, garbage out" no doubt applies to machine learning practice, as does the related catchphrase "bias in, bias out." Yet, these proverbs still understate—and somewhat misrepresent—the significance of data for machine learning.

It's not only the output of a learning algorithm that may suffer with poor input data. A dataset serves many other vital functions in the machine learning ecosystem. The dataset itself is an integral part of the problem formulation. It implicitly sorts out and operationalizes what the problem is that practitioners end up solving. Datasets have also shaped the course of entire scientific communities in their capacity to measure and benchmark progress, support competitions, and interface between researchers in academia and practitioners in industry.

If so much hinges on data in machine learning, it might come as a surprise that there is no simple answer to the question of what makes data good for what purpose. The collection of data for machine learning applications has not followed any established theoretical framework, certainly not one that was recognized a priori.

In this chapter, we take a closer look at popular datasets in the field of machine learning and the benchmarks that they support. We will use this to tease apart the different roles datasets play in scientific and engineering contexts. Then we will review the harms associated with data and discuss how they can be mitigated based on the dataset's role. We will conclude with several broad directions for improving data practices.

We limit the scope of this chapter in some important ways. Our focus will be largely on publicly available datasets that support training and testing purposes in machine learning research and applications. Our focus excludes large swaths of industrial data collection, surveillance, and data mining practices. It also excludes

data purposefully collected to test specific scientific hypotheses, such as, experimental data gathered in a medical trial.

A tour of datasets in different domains

The creation of datasets in machine learning does not follow a clear theoretical framework. Datasets aren't collected to test a specific scientific hypothesis. In fact, we will see that there are many different roles data plays in machine learning. As a result, it makes sense to start by looking at a few influential datasets from different domains to get a better feeling for what they are, what motivated their creation, how they organized communities, and what impact they had.

TIMIT

Automatic speech recognition is a machine learning problem of significant commercial interest. Its roots date back to the early 20th century.[402]

Interestingly, speech recognition also features one of the oldest benchmarks data sets, the TIMIT (Texas Instruments/Massachusetts Institute for Technology) data. The creation of the dataset was funded through a 1986 DARPA program on speech recognition. In the mid-eighties, artificial intelligence was in the middle of a "funding winter" where many governmental and industrial agencies were hesitant to sponsor AI research because it often promised more than it could deliver. DARPA program manager Charles Wayne proposed that a way around this problem was establishing more rigorous evaluation methods. Wayne enlisted the National Institute of Standards and Technology to create and curate shared datasets for speech, and he graded success in his program based on performance on recognition tasks on these datasets.

Many now credit Wayne's program with kick starting a revolution of progress in speech recognition.[403,404] [405] According to Kenneth Ward Church,

> It enabled funding to start because the project was glamour-and-deceit-proof, and to continue because funders could measure progress over time. Wayne's idea makes it easy to produce plots which help sell the research program to potential sponsors. A less obvious benefit of Wayne's idea is that it enabled hill climbing. Researchers who had initially objected to being tested twice a year began to evaluate themselves every hour.

A first prototype of the TIMIT dataset was released in December of 1988 on a CD-ROM. An improved release followed in October 1990.

[402] Li and Mills, "Vocal Features: From Voice Identification to Speech Recognition by Machine," *Technology and Culture* 60, no. 2 (2019): S129–60.

[403] Liberman, "Fred Jelinek," *Computational Linguistics* 36, no. 4 (2010): 595–99.

[404] Church, "Emerging Trends: A Tribute to Charles Wayne," *Natural Language Engineering* 24, no. 1 (2018): 155–60.

[405] Liberman and Wayne, "Human Language Technology." *AI Magazine* 41, no. 2 (2020).

TIMIT already featured the training/test split typical for modern machine learning benchmarks. There's a fair bit we know about the creation of the data due to its thorough documentation.[406]

[406] Garofolo et al., "DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1," *NASA STI/Recon Technical Report n* 93 (1993): 27403.

TIMIT features a total of about 5 hours of speech, composed of 6300 utterances, specifically, 10 sentences spoken by each of 630 speakers. The sentences were drawn from a corpus of 2342 sentences such as the following.

```
She had your dark suit in greasy wash water all year. (sa1)
Don't ask me to carry an oily rag like that. (sa2)
This was easy for us. (sx3)
Jane may earn more money by working hard. (sx4)
She is thinner than I am. (sx5)
Bright sunshine shimmers on the ocean. (sx6)
Nothing is as offensive as innocence. (sx7)
```

The TIMIT documentation distinguishes between 8 major dialect regions[407] in the United States. Of the speakers, 70% are male and 30% are female. All native speakers of American English, the subjects were primarily employees of Texas Instruments at the time. Many of them were new to the Dallas area where they worked.

[407] New England, Northern, North Midland, South Midland, Southern, New York City, Western, Army Brat (moved around)

Racial information was supplied with the distribution of the data and coded as "White," "Black," "American Indian," "Spanish-American," "Oriental," and "Unknown." Of the 630 speakers, 578 were identified as White, 26 as Black, 2 as American Indian, 2 as Spanish-American, 3 as Oriental, and 17 as unknown.

Table 15: Demographic information about the TIMIT speakers

	Male	Female	Total (%)
White	402	176	578 (91.7%)
Black	15	11	26 (4.1%)
American Indian	2	0	2 (0.3%)
Spanish-American	2	0	2 (0.3%)
Oriental	3	0	3 (0.5%)
Unknown	12	5	17 (2.6%)

The documentation notes:

> In addition to these 630 speakers, a small number of speakers with foreign accents or other extreme speech and/or hearing abnormalities were recorded as "auxiliary" subjects, but they are not included on the CD-ROM.

It comes to no surprise that early speech recognition models had significant demographic and racial biases in their performance.

Today, several major companies, including Amazon, Apple, Google, and Microsoft, all use speech recognition models in a variety of products from cell phone apps to voice assistants. There is no longer a major open benchmark that would support training models competitive with the industrial counterparts. Industrial speech recognition pipelines are generally complex and use proprietary data sources that we don't know a lot about. Nevertheless, today's speech recognition systems continue to exhibit performance disparities along racial lines.[408]

[408] Koenecke et al., "Racial Disparities in Automated Speech Recognition," *Proceedings of the National Academy of Sciences* 117, no. 14 (2020): 7684–89.

UCI Machine Learning Repository

The UCI Machine Learning Repository currently hosts more than 500 datasets, mostly for different classification and regression tasks. Most datasets are relatively small, consisting of a few hundred or a few thousand instances. The majority are structured tabular data sets with a handful or a few tens of attributes.

The UCI Machine Learning Repository contributed to the adoption of the train-test paradigm in machine learning in the late 1980s. Pat Langley recalls:

> The experimental movement was aided by another development. David Aha, then a PhD student at UCI, began to collect data sets for use in empirical studies of machine learning. This grew into the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/), which he made available to the community by FTP in 1987. This was rapidly adopted by many researchers because it was easy to use and because it let them compare their results to previous findings on the same tasks.[409]

[409] Langley, "The Changing Science of Machine Learning" (Springer, 2011).

The most popular dataset in the repository is the Iris Data Set containing taxonomic measurements of 150 iris flowers, 50 from each of 3 species. The task is to classify the species given the measurements.

The second most popular dataset in the UCI repository[410] is the *Adult* dataset. Extracted from the 1994 Census database, it features nearly 50,000 instances describing individuals in the United States, each having 14 attributes. The task is to classify whether an individual earns more than 50,000 US dollars or less.

[410] As of October, 2020.

The Adult dataset remains popular in the algorithmic fairness community, largely because it is one of the few publicly available datasets that features demographic information including *gender* (coded in binary as male/female), as well as *race* (coded as Amer-Indian-Eskimo, Asian-Pac-Islander, Black, Other, and White).

Unfortunately, the data has some idiosyncrasies that make it less than ideal for understanding biases in machine learning models. Due to the age of the data, and the income cutoff at $50,000, almost all instances labeled *Black* are below the cutoff, as are almost all instances

labeled *female*. Indeed, a standard logistic regression model trained on the data achieves about 85% accuracy overall, while the same model achieves 91% accuracy on Black instances, and nearly 93% accuracy on female instances. Likewise, the ROC curves for the latter two groups enclose actually more area than the ROC curve for male instances. This is an atypical situation: more often, machine learning models perform worse on historically disadvantaged groups.

MNIST

The MNIST dataset contains images of handwritten digits. Its most common version has 60,000 training images and 10,000 test images, each having 28x28 black and white pixels.

Figure 35: A sample of MNIST digits

MNIST was created by researchers Burges, Cortes, and Lecun from an earlier dataset released by the National Institute of Standards and Technology (NIST). The dataset was introduced in a research paper in 1998 to showcase the use of gradient-based deep learning methods for document recognition tasks.[411] Since then cited over 30,000 times, MNIST became a highly influential benchmark in the computer vision community. Two decades later, researchers continue to use the data actively.

The original NIST data had the property that training and test data came from two different populations. The former featured the handwriting of two thousand American Census Bureau employees, whereas the latter came from five hundred American high school students.[412] The creators of MNIST reshuffled these two data sources and split them into training and test set. Moreover, they scaled and centered the digits. The exact procedure to derive MNIST from NIST

[411] LeCun et al., "Gradient-Based Learning Applied to Document Recognition," *Proceedings of the IEEE* 86, no. 11 (1998): 2278–2324.

[412] Grother, "NIST Special Database 19," *Handprinted Forms and Characters Database, National Institute of Standards and Technology*, 1995, 10.

was lost, but recently reconstructed by matching images from both data sources.[413]

The original MNIST test set was of the same size as the training set, but the smaller test set became standard in research use. The 50,000 digits in the original test set that didn't make it into the smaller test set were later identified and dubbed *the lost digits*.[414]

From the beginning, MNIST was intended to be a benchmark used to compare the strengths of different methods. For several years, LeCun maintained an informal leaderboard on a personal website that listed the best accuracy numbers that different learning algorithms achieved on MNIST.

[413] Yadav and Bottou, "Cold Case: The Lost Mnist Digits," *arXiv Preprint arXiv:1905.10498*, 2019.

[414] Yadav and Bottou.

Table 16: A snapshot of the original MNIST leaderboard from February 2, 1999. Source: Internet Archive (Retrieved: December 4, 2020)

Method	Test error (%)
linear classifier (1-layer NN)	12.0
linear classifier (1-layer NN) [deskewing]	8.4
pairwise linear classifier	7.6
K-nearest-neighbors, Euclidean	5.0
K-nearest-neighbors, Euclidean, deskewed	2.4
40 PCA + quadratic classifier	3.3
1000 RBF + linear classifier	3.6
K-NN, Tangent Distance, 16x16	1.1
SVM deg 4 polynomial	1.1
Reduced Set SVM deg 5 polynomial	1.0
Virtual SVM deg 9 poly [distortions]	0.8
2-layer NN, 300 hidden units	4.7
2-layer NN, 300 HU, [distortions]	3.6
2-layer NN, 300 HU, [deskewing]	1.6
2-layer NN, 1000 hidden units	4.5
2-layer NN, 1000 HU, [distortions]	3.8
3-layer NN, 300+100 hidden units	3.05
3-layer NN, 300+100 HU [distortions]	2.5
3-layer NN, 500+150 hidden units	2.95
3-layer NN, 500+150 HU [distortions]	2.45
LeNet-1 [with 16x16 input]	1.7
LeNet-4	1.1
LeNet-4 with K-NN instead of last layer	1.1
LeNet-4 with local learning instead of ll	1.1
LeNet-5, [no distortions]	0.95
LeNet-5, [huge distortions]	0.85
LeNet-5, [distortions]	0.8

Method	Test error (%)
Boosted LeNet-4, [distortions]	0.7

In its capacity as a benchmark, it became a showcase for the emerging kernel methods of the early 2000s that temporarily achieved top performance on MNIST.[415] Today, it is not difficult to achieve less than 0.5% classification error with a wide range of convolutional neural network architectures. The best models classify all but a few pathological test instances correctly. As a result, MNIST is widely considered too easy for today's research tasks.

[415] DeCoste and Schölkopf, "Training Invariant Support Vector Machines," *Machine Learning* 46, no. 1 (2002): 161–90.

MNIST wasn't the first dataset of handwritten digits in use for machine learning research. Earlier, the US Postal Service (USPS) had released a dataset of 9298 images (7291 for training, and 2007 for testing). The USPS data was actually a fair bit harder to classify than MNIST. A non-negligible fraction of the USPS digits look unrecognizable to humans,[416] whereas humans recognize essentially all digits in MNIST.

[416] Bromley and Sackinger, "Neural-Network and k-Nearest-Neighbor Classifiers," *Rapport Technique*, 1991, 11359–910819.

ImageNet

ImageNet is a large repository of labeled images that has been highly influential in computer vision research over the last decade. The image labels correspond to nouns from the WordNet lexical database of the English language.[417] WordNet groups nouns into cognitive synonyms, called *synsets*. The words *car* and *automobile*, for example, would fall into the same synset. On top of these categories WordNet provides a hierarchical tree structure according to a super-subordinate relationship between synsets. The synset for *chair*, for example, is a child of the synset for *furniture* in the wordnet hierarchy. WordNet existed before ImageNet and in part inspired the creation of Imagenet.

[417] Miller, *WordNet: An Electronic Lexical Database* (MIT press, 1998).

The initial release of ImageNet included about 5000 image categories, each corresponding to a synset in WordNet. These ImageNet categories averaged about 600 images per category.[418] ImageNet grew over time and its Fall 2011 release had reached about 32,000 categories.

[418] Deng et al., "Imagenet: A Large-Scale Hierarchical Image Database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Ieee, 2009), 248–55.

The construction of ImageNet required two essential steps: retrieving candidate images for each synset, and labeling the retrieved images. This first step utilized online search engines and photo sharing platforms with a search interface, specifically, Flickr. Candidate images were taken from the image search results associated with the synset nouns for each category.

For the second labeling step, the creators of ImageNet turned

to Amazon's Mechanical Turk platform (MTurk). MTurk is an online labor market that allows individuals and corporations to hire on-demand workers to perform simple tasks. In this case, MTurk workers were presented with candidate images and had to decide whether or not the candidate image was indeed an image corresponding to the category that it was putatively associated with.

It is important to distinguish between this ImageNet database and a popular machine learning benchmark and competition, called ImageNet Large Scale Visual Recognition Challenge (ILSVRC), that was derived from it.[419] The competition was organized yearly from 2010 until 2017, reaching significant notoriety in both industry and academia, especially as a benchmark for emerging deep learning models.

[419] Russakovsky et al., "Imagenet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision* 115, no. 3 (2015): 211–52.

When machine learning practitioners say "ImageNet" they typically refer to the data used for the image classification task in the 2012 ILSVRC benchmark. The competition included other tasks, such as object recognition, but image classification has become the most popular task for the dataset. Expressions such as "a model trained on ImageNet" typically refer to training an image classification model on the benchmark data set from 2012.

Another common practice involving the ILSVRC data is *pre-training*. Often a practitioner has a specific classification problem in mind whose label set differs from the 1000 classes present in the data. It's possible nonetheless to use the data to create useful features that can then be used in the target classification problem. Where ILSVRC enters real-world applications it's often to support pre-training.

This colloquial use of the word ImageNet can lead to some confusion, not least because the ILSVRC-2012 dataset differs significantly from the broader database. It only includes a subset of 1000 categories. Moreover, these categories are a rather skewed subset of the broader ImageNet hierarchy. For example, of these 1000 categories only three are in the *person* branch of the WordNet hierarchy, specifically, *groom*, *baseball player*, and *scuba diver*. Yet, more than 100 of the 1000 categories correspond to different dog breeds.[420]

[420] The number is 118, to be exact, not counting wolves, foxes, and wild dogs that are also present among the 1000 categories.

What motivated the exact choice of these 1000 categories is not entirely clear. The apparent canine inclination, however, isn't just a quirk either. At the time, there was an interest in the computer vision community in making progress on prediction with many classes, some of which are very similar. This reflects a broader pattern in the machine learning community. The creation of datasets is often driven by an intuitive sense of what the technical challenges are for the field. In the case of ImageNet, another important consideration was scale, both in terms of the number of images and the number of classes.

The large scale annotation and labeling that went into Imagenet

falls into a category of labor that anthropologist Gray and computer scientist Suri call *ghost work* in their book of the same name.[421] They point out:

> MTurk workers are the AI revolution's unsung heroes.

Indeed, ImageNet was labeled by about 49,000 MTurk workers from 167 countries over the course of multiple years.

[421] Gray and Suri, *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass* (Eamon Dolan Books, 2019).

The Netflix Prize

The Netflix Prize was one of the most famous machine learning competitions. Starting on October 2, 2006, the competition ran for nearly three years ending with a grand prize of $1M, announced on September 18, 2009. Over the years, the competition saw 44,014 submissions from 5169 teams.

The Netflix training data contained roughly 100 million movie ratings from nearly 500 thousand Netflix subscribers on a set of 17770 movies. Each data point corresponds to a tuple `<user, movie, date of rating, rating>`. At about 650 megabytes in size, the dataset was just small enough to fit on a CD-ROM, but large enough to be pose a challenge at the time.

The Netflix data can be thought of as a matrix with $n = 480189$ rows and $m = 17770$ columns. Each row corresponds to a Netflix subscriber and each column to a movie. The only entries present in the matrix are those for which a given subscriber rated a given movie with rating in $\{1, 2, 3, 4, 5\}$. All other entries—that is, the vast majority—are missing. The objective of the participants was to predict the missing entries of the matrix, a problem known as matrix completion, or collaborative filtering somewhat more broadly. In fact, the Netflix challenge did so much to popularize this problem that it is sometimes called the Netflix problem. The idea is that if we could predict missing entries, we'd be able to recommend unseen movies to users accordingly.

The hold out data that Netflix kept secret consisted of about three million ratings. Half of them were used to compute a running leaderboard throughout the competition. The other half determined the final winner.

The Netflix competition was hugely influential. Not only did it attract significant participation, it also fueled much academic interest in collaborative filtering for years to come. Moreover, it popularized the competition format as an appealing way for companies to engage with the machine learning community. A startup called Kaggle, founded in April 2010, organized hundreds of machine learning competitions for various companies and organizations before its acquisition by Google in 2017.

But the Netflix competition became infamous for another reason. Although Netflix had replaced usernames by pseudonymous numbers, researchers Narayanan and Shmatikov were able to re-identify some of the Netflix subscribers whose movie ratings were in the dataset[422] by linking those ratings with publicly available movie ratings on IMDB, an online movie database. Some Netflix subscribers had also publicly rated an overlapping set of movies on IMDB under their real identities. In the privacy literature, this is called a *linkage attack* and it's one of the ways that seemingly anonymized data can be de-anonymized.[423]

What followed were multiple class action lawsuits against Netflix, as well as an inquiry by the Federal Trade Commission over privacy concerns. As a consequence, Netflix canceled plans for a second competition, which it had announced on August 6, 2009.

To this day, privacy concerns are a legitimate obstacle to public data release and dataset creation. Deanonymization techniques are mature and efficient. There provably is no algorithm that could take a dataset and provide a rigorous privacy guarantee to all participants, while being useful for all analyses and machine learning purposes. Dwork and Roth call this the Fundamental Law of Information Recovery: *"overly accurate answers to too many questions will destroy privacy in a spectacular way."*[424]

[422] Narayanan and Shmatikov, "Robust de-Anonymization of Large Sparse Datasets," in *2008 IEEE Symposium on Security and Privacy (Sp 2008)* (IEEE, 2008), 111–25.

[423] Dwork et al., "Exposed! A Survey of Attacks on Private Data," *Annual Review of Statistics and Its Application* 4 (2017): 61–84.

[424] Dwork, Roth, and others, "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends in Theoretical Computer Science* 9, no. 3–4 (2014): 211–407.

Roles datasets play

In machine learning research and engineering, datasets play a different and more prominent set of roles than they do in most other fields. We have mentioned several of these above but let us now examine them in more detail. Understanding these is critical to figuring out which technical and cultural aspects of benchmarks are essential, how harms arise, and how to mitigate them.

A source of real data

Edgar Anderson was a botanist and horticulturist who spent much of the 1920s and '30s collecting and analyzing data on Irises to study biological and taxonomic questions. The Iris dataset in the UCI machine learning repository mentioned above is the result of Anderson's labors — or a tiny sliver of them, as most of the observations in the dataset came from a single day of field work. The dataset contains 50 observations each of 3 iris plants; the task is to distinguish the species based on 4 physical attributes (sepal length and width; petal length and width). Most of the tens of thousands of researchers who have used this dataset are not interested in taxonomy, let alone

irises. What, then, are they using the dataset for?

Although the data was collected by Anderson, it was actually published in the paper "The use of multiple measurements in taxonomic problems" by Ronald Fisher, who was a founder of modern statistics as well as a eugenicist.[425,426] Fisher was Anderson's collaborator. Although Fisher had some interest in taxonomy, he was primarily interested in using the data to develop statistical techniques (with an eye toward applications for eugenics). In the 1936 paper, Fisher introduces Linear Discriminant Analysis (LDA) and shows that it performs well on this task.

The reason the Iris dataset proved to be a good application of LDA is that there exists a linear projection of the four features which seems to result in a mixture of Gaussians (one for each of the three species), and the means of the three distributions are relatively far apart; one of the species is in fact perfectly separable from the other two. Every learning algorithm implicitly makes assumptions about the data-generating process: without assumptions, there is no basis for making predictions on unseen points.[427] If we could perfectly mathematically describe the data generating process behind the physical characteristics of irises (or any other population), we wouldn't need a dataset — we could mathematically work out how well an algorithm would perform. In practice, for complex phenomena, such perfect mathematical descriptions rarely exist.[428] Thus, the usefulness of a learning algorithm is established by testing it on real datasets.

The reliance on benchmark datasets as a source of real data was a gradual development in machine learning research. For example, Rosenblatt's perceptron experiments in the 1950s used two artificial stimuli (the characters E and X), with numerous variants of each created by rotation and other transformations.[429] The controlled input was considered useful to understand the behavior of the system. Writing in 1988, Pat Langley advocates for a hybrid approach, pointing out that "successful runs on a number of different natural domains provide evidence of generality" but also highlighting the use of artificial data for better understanding.[430] Especially after the establishment of the UCI repository around this time, it has become common to evaluate new algorithms on widely-used benchmark datasets as a way of establishing that the researcher is not "cheating" by picking contrived inputs.

To summarize, when a researcher seeks to present evidence that an algorithmic innovation is useful, the use of real dataset as opposed to artificial data ensures that the researcher didn't make up data to suit the algorithm. Further, the use of prominent benchmark datasets wards off skepticism that the researcher may have cherry picked a

[425] The eugenics connection is not accidental: other central figures in the development of modern statistics such as Francis Galton and Karl Pearson were algo eugenicists. (Evans, "RA Fisher and the Science of Hatred," *The New Statesman*, 2020; Louçã, "Emancipation Through Interaction–How Eugenics and Statistics Converged and Diverged," *Journal of the History of Biology* 42, no. 4 (2009): 649–84)

[426] Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics* 7, no. 2 (1936): 179–88.

[427] Mitchell, *The Need for Biases in Learning Generalizations* (Department of Computer Science, Laboratory for Computer Science Research . . . , 1980).

[428] Different communities place different value on attempting to discover the true data generating process. Machine learning places relatively little emphasis on this goal. (Breiman and others, "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)," *Statistical Science* 16, no. 3 (2001): 199–231)

[429] Rosenblatt, "Perceptron Simulation Experiments," *Proceedings of the IRE* 48, no. 3 (1960): 301–9.

[430] Langley, "Machine Learning as an Experimental Science" (Springer, 1988).

dataset with specific properties that makes the algorithm effective. Finally, the use of multiple benchmark datasets from different domains suggests that the algorithm is highly general.

Perversely, domain ignorance is treated almost as a virtue rather than a drawback. For example, researchers who achieve state-of-the-art performance on (say) Chinese-to-English translation may point out that none of them speak Chinese. The subtext is that they couldn't have knowingly or unknowingly picked a model that works well only when the source language is linguistically similar to Chinese.

A catalyst and measure of domain-specific progress

Algorithmic innovations that are highly portable across domains, while important, are rare. Much of the progress in machine learning is instead tailored to specific domains and problems. The most common way to demonstrate such progress is to show that the innovation in question can be used to achieve "state of the art" performance on a benchmark dataset for that task.

The idea that datasets spur algorithmic innovation bears some explanation. For example, the Netflix Prize is commonly credited as responsible for the discovery of the effectiveness of matrix factorization in recommender systems (often attributed to Simon Funk, a pseudonymous contestant).[431] Yet, the technique had been proposed in the context of movie recommendation as early as 1998[432] and for search as early as 1990.[433] However, it was not previously apparent that it outperformed neighborhood-based methods and that it could discover meaningful latent factors. The clarity of the Netflix leaderboard and the credibility of the dataset helped establish the significance of matrix factorization.[434]

Somewhat separately from the role of spurring algorithmic innovation, benchmark datasets also offer a convenient way to measure its results (hence the term benchmark). The progression of state-of-the-art accuracy on a benchmark dataset and task can be a useful indicator. A relatively flat curve of accuracy over time may indicate that progress has stalled, while a discontinuous jump may indicate a breakthrough. Reaching an error rate that is close to zero or at least lower than the "human error" for perception tasks is often considered a sign that the task is "solved" and that it is time for the community to move on to a harder challenge.

While these are appealing heuristics, there are also pitfalls. In particular, a statement such as "the state of the art accuracy for image classification is 95%" is not a scientifically meaningful claim that can be assigned a truth value, because the number is highly sensitive to

[431] Funk, "Try This at Home," *Http://Sifter.org/~Simon/Journal/2006*, 2006.

[432] Billsus, Pazzani, and others, "Learning Collaborative Information Filters." in *Icml*, vol. 98, 1998, 46–54.

[433] Deerwester et al., "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science* 41, no. 6 (1990): 391–407.

[434] Koren, Bell, and Volinsky, "Matrix Factorization Techniques for Recommender Systems," *Computer* 42, no. 8 (2009): 30–37.

the data distribution.

A notable illustration of this phenomenon comes from a paper by researchers Recht, Roelofs, Schmidt, and Shankar. They carefully recreated new test sets for the CIFAR-10 and ImageNet classification benchmarks according to the very same procedure as the original test sets.[435] They then took a large collection of representative models proposed over the years and evaluated all of them on the new test sets. All models suffered a significant drop in performance on the new test set, corresponding to about 5 years of progress in image classification. They found that this is because the new test set represents a slightly different distribution. This is despite the researchers' careful efforts to replicate the data collection procedure; we should expect that test sets created by different procedures should result in much greater performance differences.

[435] Recht et al., "Do ImageNet Classifiers Generalize to ImageNet?" in *Proc. 36th ICML*, 2019.

The same graphs also provide a striking illustration of why benchmark datasets are a practical necessity for performance comparison in machine learning. Consider a hypothetical alternative approach analogous to the norm in many other branches of science: a researcher evaluating a claim (algorithm) describes in detail their procedure for sampling the data; other researchers working on the same problem sample their own datasets based on the published procedure. Some reuse of datasets occurs, but there is no standardization. The graphs show that even extremely careful efforts to sample a new dataset from the same distribution would shift the distribution sufficiently to make performance comparison hopeless.

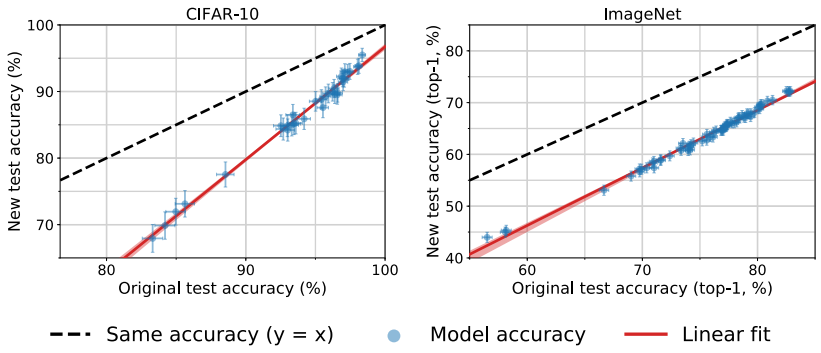


Figure 36: Model accuracy on the original test sets vs. new test sets for CIFAR-10 and ImageNet. Each data point corresponds to one model in a test bed of representative models (shown with 95% Clopper-Pearson confidence intervals). The plots reveal two main phenomena: (i) There is generally a significant drop in accuracy from the original to the new test sets. (ii) The model accuracies closely follow a linear function, meaning that models that perform well on the old test set also tend to perform well on the new test set. The narrow shaded region is a 95% confidence region for the linear fit.

In other words, reported accuracy figures from benchmark datasets do not constitute generalizable scientific knowledge, because they don't have external validity beyond the specific dataset. While the Recht et al. paper is limited to image classification, it seems scientifically prudent to assume a lack of external validity for other machine learning tasks as well, unless there is evidence to the con-

trary. Yet the two graphs above hint at a different type of knowledge that seems to transfer almost perfectly to the new test set: the *relative* performance of models.

The relative performance of models for a given task is a very useful type of practitioner-oriented knowledge that can be gained from benchmark leaderboards. A question that practitioners often face is, "which class of models should I use for [given task] and how should I optimize it?" A benchmark dataset (together with the associated task definition) can be seen as a proxy for answering this question in a constrained setting, analogous to laboratory studies in other branches of science. The hope is that algorithms (and model classes or architectures) identified as state of the art based on benchmark evaluation are also the ones that will be effective on the practitioner's test set. In other words, practitioners can outsource the laborious task of model selection to the benchmark leaderboard.

To be clear, this is an oversimplification. Practitioners have many concerns in addition to accuracy such as the computational cost (of both training and prediction), interpretability, and, increasingly, fairness and environmental cost. Thus, benchmark performance is useful to practitioners but far from the only consideration for model selection.

We can imagine a spectrum of how similar the new test set is to the benchmark set. At the one extreme, if the new test set is truly a new sample from the exact same distribution, then the ranking of model classes should be the same for the two sets. At the other extreme, the distributions may be so different that they constitute essentially different tasks, so that performance on one is not a useful guide to performance on the other. In between these extremes is a big grey area that is not well understood, and it is currently more art than science.

The lack of clarity on how much we can generalize from one or a few benchmarks is associated with well known controversies. For example, support vector machines were competitive with neural networks on earlier-generation benchmarks such as NIST digit recognition,[436] which was one reason why interest in neural networks dwindled in the 1990s. The clear superiority of neural networks on newer benchmarks such as ImageNet was only belatedly recognized.

[436] Cortes and Vapnik, "Support-Vector Networks," *Machine Learning* 20, no. 3 (1995): 273–97.

A source of (pre-)training data

Above, we have envisioned that practitioners use the benchmark leaderboard as a guide to model selection but then train the selected models from scratch on their own (often proprietary) data sources. But practitioners often can and do go further.

In some cases, it may be possible to train on a benchmark dataset and directly use the resulting model in one's application. This depends on the domain and the task, and is more suitable when the distribution shift is minimal and the set of class labels is stable. For example, it is reasonable to deploy a digit recognizer pretrained on MNIST, but not so much an image classifier pretrained on ILSVRC (without some type of adaptation to the target domain). Indeed, ILSVRC consists of a rather arbitrary subset of 1,000 classes of ImageNet, and a pretrained model is correspondingly limited in the set of labels it is able to output.[437] Finally, consider a recommendation system benchmark dataset. There is no way to even attempt to use it directly as training data because the users about whom one wants to make predictions are highly unlikely to be present in the training set.

In most cases, the creators of benchmark datasets do not intend them to be used as a source of training data, although benchmark datasets are often misused for this purpose. A rare exception is The Pile: a large (800 GB) English text corpus that is explicitly targeted at training language models. To improve the generalization capabilities of models trained on this corpus, the authors included diverse text from 22 different sources.[438]

Even when benchmark datasets are not useful as training data for the above-mentioned reasons, they can be useful as pre-training data for transfer learning. Transfer learning refers to using an existing model as a starting point for building a new model. A new model may be needed because the data distribution has shifted compared to what the existing model was optimized for, or because it aims to solve a different task altogether. For example, a model pre-trained on ImageNet (or ILSVRC) may be adapted via further training for recognizing different species (distribution shift) or as part of an image captioning model (a different task).

There are different intuitions to explain why transfer learning is often effective. One is that the final layers of a neural network correspond to semantically high-level representations of the input. Pre-training is a way of learning these representations that tend to be useful for many tasks. Another intuition is that pre-training is a way of initializing weights that offers an improvement over random initialization in that it requires fewer samples from the target domain for convergence.

Pretraining offers the practical benefit of being able to share the knowledge contained in a dataset without releasing the raw data. Many datasets, especially those created by companies using customer data, cannot be published due to privacy or confidentiality concerns. The release of pretrained models is thus an important avenue of knowledge sharing from industry to academia.[439] Sharing pretrained

[437] The ImageNet Roulette project was a telling demonstration of what happens when a model trained on the (full) ImageNet dataset is applied to a different test distribution, one consisting primarily of images of people. The results were grotesque. The demonstration has been discontinued, but many archived results may be found in articles about the project. See also: (Crawford and Paglen, "Excavating AI: The Politics of Training Sets for Machine Learning," *Excavating AI (Www.excavating.ai)*, 2019)

[438] Gao et al., "The Pile: An 800gb Dataset of Diverse Text for Language Modeling," *arXiv Preprint arXiv:2101.00027*, 2020.

[439] However, the privacy concern has resurfaced due to the possibility of model inversion attacks. The term refers to algorithms that aim to reconstruct some of the training data by interacting with a model trained on it. (Fredrikson, Jha, and Ristenpart, "Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, 1322–33; Carlini et al., "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks," in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, 2019, 267–84)

models is also helpful to users for whom training from scratch is cost prohibitive.

Let's wrap up our analysis of the roles of benchmark datasets. We identified six distinct roles: (1) providing data sampled from real-world occurring distributions that enables largely domain-agnostic investigations of learning algorithms; (2) enabling domain-specific progress by providing datasets that are representative of real-world tasks in that domain yet abstract away unnecessary detail; (3) providing a convenient albeit crude numerical way to track scientific progress on a problem; (4) enabling model comparison and allowing practitioners to outsource model selection to public leaderboards; (5) providing a source of pre-training data for representation learning, weight initialization, etc; (6) providing a source of training data. The progression of these six roles is generally toward increasing domain- and task-specificity, and from science-oriented to practice-oriented.

The scientific basis of machine learning benchmarks

Now we examine a seeming mystery: whether and why the benchmark approach works despite the practice of repeated testing on the same data.

Methodologically, much of modern machine learning practice rests on a variant of trial and error, which we call the train-test paradigm. Practitioners repeatedly build models using any number of heuristics and test their performance to see what works. Anything goes as far as training is concerned, subject only to computational constraints, so long as the performance looks good in testing. Trial and error is sound so long as the testing protocol is robust enough to absorb the pressure placed on it. We will examine to what extent this is the case in machine learning.

From a theoretical perspective, the best way to test the performance of a classifier is to collect a sufficiently large fresh dataset and to compute the average error on that test set. Data collection, however, is a difficult and costly task. In most applications, practitioners cannot sample fresh data for each model they would like to try out. A different practice has therefore become the de-facto standard. Practitioners split their dataset into typically two parts, a *training set* used for training a model, and a *test set* used for evaluating its performance.[440] Often the split is determined when the dataset is created. Datasets used for benchmarks in particular have one fixed split persistent throughout time. A number of variations on this theme go under the name *holdout method*.

[440] Sometimes practitioners divide their data into multiple splits, e.g., training, validation, and test sets. However, for our discussion here that won't be necessary.

Machine learning competitions have adopted the same format. The company Kaggle, for example, has organized hundreds of compe-

titions since it was founded. In a competition, a holdout set is kept secret and is used to rank participants on a public leaderboard as the competition unfolds. In the end, the final winner is whoever scores highest on a separate secret test set not used to that point.

In all applications of the holdout method the hope is that the test set will serve as a fresh sample that provides good performance estimates for all the models. The central problem is that practitioners don't just use the test data once only to retire it immediately thereafter. The test data are used incrementally for building one model at a time while incorporating feedback received previously from the test data. This leads to the fear that eventually models begin to *overfit* to the test data. This type of overfitting is sometimes called *adaptive overfitting* or *human-in-the-loop overfitting*.

Duda, Hart, and Stork summarize the problem aptly in their 1973 textbook:[441]

> In the early work on pattern recognition, when experiments were often done with very small numbers of samples, the same data were often used for designing and testing the classifier. This mistake is frequently referred to as "testing on the training data." A related but less obvious problem arises when a classifier undergoes a long series of refinements guided by the results of repeated testing on the same data. This form of "training on the testing data" often escapes attention until new test samples are obtained.

[441] Duda, Hart, and Stork, *Pattern Classification and Scene Analysis*, vol. 3 (Wiley New York, 1973).

Nearly half a century later, Hastie, Tibshirani, and Friedman still caution in the 2017 edition of their influential textbook:[442]

> Ideally, the test set should be kept in a "vault," and be brought out only at the end of the data analysis. Suppose instead that we use the test-set repeatedly, choosing the model with smallest test-set error. Then the test set error of the final chosen model will underestimate the true test error, sometimes substantially.

[442] Hastie, Tibshirani, and Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, 2017).

While the suggestion to keep the test data in a "vault" is safe, it couldn't be further from the reality of modern practice. Popular test datasets often see tens of thousands of evaluations.

Yet adaptive overfitting doesn't seem to be happening. Recall the scatter plots by Recht et al. above: the plots admit a clean linear fit with positive slope. In other words, the better a model is on the old test set, the better it is on the new test set. But notice that newer models, i.e., those with higher performance on the original test set, had *more* time to adapt to the test set and to incorporate more information about it. Nonetheless, the better a model performed on the old test set the better it performs on the new set. Moreover, on CIFAR-10 we even see clearly that the absolute performance drops diminishes with increasing accuracy on the old test set. In particular,

if our goal was to do well on the new test set, seemingly our best strategy is to continue to inch forward on the old test set.

The theoretical understanding of why machine learning practice has not resulted in overfitting is still catching up. Here, we highlight one of many potential explanations, called the leaderboard principle. It is a subtle effect in which publication biases force researchers to chase state-of-the-art results, and they only publish models if they see significant improvements over prior models. This cultural practice can be formalized by the *Ladder algorithm*. For each given classifier, it compares the classifier's holdout error to the previously smallest holdout error achieved by any classifier encountered so far. If the error is below the previous best by some margin, it announces the holdout error of the current classifier and notes it as the best seen so far. Importantly, if the error is not smaller by a margin, the algorithm releases the previous best (rather than the new error). It can be proven that the Ladder algorithm avoids overfitting in the sense that it accurately measures the error of the best performing classifier among those encountered.[443]

[443] Blum and Hardt, "The Ladder: A Reliable Leaderboard for Machine Learning Competitions," in *International Conference on Machine Learning* (PMLR, 2015), 1006–14.

Benchmark praxis and culture

The above discussion hints at the importance of cultural practices for a full understanding of benchmark datasets. Let us now discuss these in more detail, highlighting both dataset creators and users. These practices have helped make the benchmark-oriented approach successful but also impact the harms associated with data. Let's start with creators.

Benchmark creators define the task. This involves, among other things, selecting the high-level problem, defining the target variable, the procedure for sampling the data, and the scoring function. If manual annotation of the data is necessary, the dataset creator must develop a codebook or rubric for doing so and orchestrate crowdwork if needed. Data cleaning to ensure high-quality labels is usually required.

In defining the task, benchmark developers navigate a tricky balance: a task that is seen as too easy using existing techniques will not spur innovation while a task that is seen as too hard may be demotivating. Finding the sweet spot requires expertise, judgment, and some luck. If the right balance is achieved, the benchmark drives progress on the problem. In this way, benchmark creators play an outsized role in defining the vision and agenda for machine learning communities.

For example, as discussed above, MNIST was derived from a previous dataset released by NIST in which the training and test set

were drawn from different sources. The MNIST creators argued that this was necessary because

> Drawing sensible conclusions from learning experiments requires that the result be independent of the choice of training set and test among the complete set of samples.

In other words, if an algorithm performs well on NIST it is unclear how much of this due to its ability to learn the training distribution and how much of it is due to its ability to ignore the differences between the train and test distributions. MNIST allows researchers to focus selectively on the former question. This was a fruitful approach in 1995. Decades later, when problems like MNIST classification are effectively solved, the attention of benchmark dataset creators has turned towards methods for handling distribution shift that LeCun et al. justifiably chose to ignore.[444]

[444] Koh et al., "Wilds: A Benchmark of in-the-Wild Distribution Shifts," *arXiv Preprint arXiv:2012.07421*, 2020.

Another tricky balance is between abstracting away domain details so that the task is approachable for a broad swath of machine learning experts, and preserving enough details so that the methods that work in the benchmark setting will translate to production settings. One reason the Netflix Prize was so popular is because the data is just a matrix, and it is possible to achieve good performance (in the sense of beating Netflix's baseline) without really thinking about what the data means. No understanding of film or user psychology was necessary — or helpful, as it turned out. It is possible that domain expertise would have proved essential if the problem had been formulated differently — say, to require explainability or emphasize good performance even for users with very few previous ratings.

Another challenge for dataset creators is to avoid leakage. In an apocryphal story from the early days of computer vision, a classifier was trained to discriminate between images of Russian and American tanks with seemingly high accuracy, but it turned out that this was only because the Russian tanks had been photographed on a cloudy day and the American ones on a sunny day.[445] Data leakage refers to a spurious relationship between the feature vector and the target variable that is an artifact of the data collection or sampling strategy. Since the spurious relationship won't be present when the model is deployed, leakage usually leads to inflated estimates of model performance. Kaufman et al. present an overview of leakage in machine learning.[446]

[445] Branwen, "The Neural Net Tank Urban Legend," 2011.

[446] Kaufman et al., "Leakage in Data Mining: Formulation, Detection, and Avoidance," *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6, no. 4 (2012): 1–21.

Yet another critical responsibility of benchmark dataset creators is to implement a train-test framework. Most contests have various restrictions in place in an attempt to prevent both accidental overfitting to the leaderboard test set and intentional reverse engineering. Although, as we described above, benchmark praxis differs from

the textbook version of the holdout method, practitioners have arrived at a set of techniques that have worked in practice, even if our theoretical understanding of why they work is still catching up.

Taking a step back, in any scientific endeavor there are the difficult tasks of framing the problem, ensuring that the methods have internal and external validity, and interpreting the results. Benchmark dataset creators handle as many of these hard tasks as possible, simplifying the goal of dataset users to the point where if a researcher beats the state-of-the-art performance, there is a good chance that there is a scientific insight somewhere in the methods, although extracting what this insight is may still require work. Further simplifying things for dataset users is the fact that there are no restrictions other than computational constraints on how the researcher uses the training data, as long as performance on the test set looks good.

To be clear, this approach has many pitfalls. Researchers rarely perform the statistical hypothesis tests needed to have confidence in the claim that one model performs better than another. Our understanding of how to account for the numerous sources of variance in these performance measurements is still evolving; a 2021 paper that aims to do so argues that many of the claims of State-of-the-Art performance in natural language performance and computer vision don't hold up when subjected to such tests.[447]

There have long been articles noting the limitations of what researchers and practitioners can learn from benchmark performance evaluation.[448] David Aha, co-creator of the UCI repository, recalls that these limitations were well understood as early as 1995, just a few years after the repository was established.[449]

While it is important to acknowledge the limitations, it is also worth highlighting that this approach works at all. One reason for this success is that the scientific questions are primarily about algorithms and not the populations that the datasets are sampled from.

Indeed, there is a case to be made that other scientific communities should adopt the machine learning community's approach, sometimes called the Common Task Method.[450] Diverse scientific fields including economics, political science, psychology, genetics, and many others have seen an infusion of machine learning methods alongside a new focus on maximizing predictive accuracy as a research objective. These shifts have been accompanied by a rash of reproducibility failures, with large fractions of published papers falling prey to pitfalls such as data leakage and incorrect train-test split.[451] Use of the benchmark dataset approach could have avoided most of these pitfalls.[452]

Now let us transition to dataset users. Benchmark users have

[447] Bouthillier et al., "Accounting for Variance in Machine Learning Benchmarks," *Proceedings of Machine Learning and Systems* 3 (2021).

[448] Saitta and Neri, "Learning in the 'Real World'," *Machine Learning* 30, no. 2 (1998): 133–63; Salzberg, "On Comparing Classifiers: A Critique of Current Research and Methods," *Data Mining and Knowledge Discovery* 1, no. 1 (1999): 1–12.

[449] Radin, "'Digital Natives': How Medical and Indigenous Histories Matter for Big Data," *Osiris* 32, no. 1 (2017): 43–64.

[450] Liberman, "Fred Jelinek."

[451] Ivanescu et al., "The Importance of Prediction Model Validation and Assessment in Obesity and Nutrition Research," *International Journal of Obesity* 40, no. 6 (June 2016): 887–94, https://doi.org/10.1038/ijo.2015.214; Whelan and Garavan, "When Optimism Hurts: Inflated Predictions in Psychiatric Neuroimaging," *Biological Psychiatry*, Mechanisms of Aging and Cognition, 75, no. 9 (May 2014): 746–48, https://doi.org/10.1016/j.biopsych.2013.05.014; DeMasi, Kording, and Recht, "Meaningless Comparisons Lead to False Optimism in Medical Machine Learning," *PLoS ONE* 12, no. 9 (September 2017), https://doi.org/10.1371/journal.pone.0184604; Blagus and Lusa, "Joint Use of over- and Under-Sampling Techniques and Cross-Validation for the Development and Assessment of Prediction Models," *BMC Bioinformatics* 16, no. 1 (November 2015): 363, https://doi.org/10.1186/s12859-015-0784-9; Vandewiele et al., "Overly Optimistic Prediction Results on Imbalanced Data: A Case Study of Flaws and Benefits When Applying over-Sampling," *Artificial Intelligence in Medicine* 111 (January 2021): 101987, https://doi.org/10.1016/j.artmed.2020.101987; Christodoulou et al., "A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models," *Journal of Clinical Epidemiology* 110 (June 2019): 12–22, https://doi.org/10.1016/j.jclinepi.2019.02.004; Arp et al., "Dos and Don'ts of Machine Learning in Computer Security," October 2020, https://arxiv.org/abs/2010.09470v1.

[452] We are grateful to Sayash Kapoor for this observation.

embraced the freedom afforded by the approach. As a result, the community of users is large — for example, the data science platform Kaggle has over 5 million registered users of whom over 130,000 have participated in a competition. There is less gatekeeping in machine learning research than in other disciplines. Many prominent findings bypass peer review. If a technique performs well on the leaderboard, that is considered to speak for itself. Many people who contribute these findings are not formally affiliated with research institutions.

Overall, the culture of progress in machine learning combines the culture of academic scholarship, engineering, and even gaming, with a community of hobbyists and practitioners sharing tips and tricks on forums and engaging in friendly competition. This freewheeling culture may seem jarring to some observers, especially given the sensitivity of some of the datasets involved. The lack of gatekeeping means fewer opportunities for ethical training.

There is another aspect of benchmark culture that amplifies the harms associated with data: collecting data without informed consent and distributing it widely without adequate context. Many modern datasets, especially in computer vision and natural language processing, are scraped from the web. In such cases, it is infeasible to obtain informed consent from the individual authors of the content. What about a dataset such as the Netflix Prize where a company releases data from its own platform? Even if companies disclose in their terms of service that data might be used for research, it is doubtful that informed consent has been obtained since few users read and understand Terms of Service documents and because of the complexity of the issues involved.

When an individual's data becomes part of a benchmark dataset, it gets distributed widely. Popular benchmark datasets are downloaded by thousands of researchers, students, developers, and hobbyists. Scientific norms also call for the data to be preserved indefinitely in the interest of transparency and reproducibility. Thus, not only might individual pieces of data in these datasets be distributed and viewed widely, they are viewed in a form that strips them of their original context. A joke in bad taste written on social media and later deleted may be captured alongside documents from the library of congress.

Harms associated with data

Now we will discuss a few important types of harms associated with benchmark datasets and how to mitigate them. We don't mean to imply that all of these harms are the "fault" of dataset creators, but understanding how data plays into these harms will bring clarity on how to intervene.

Downstream and representational harms

A dataset's downstream harms are those that arise from the models trained on it. This is a type of harm that readily comes to mind: bad data may lead to bad models which can cause harm to the people they purportedly serve. Errors in face recognition have led to arrests of the wrong person[453] and biased criminal risk prediction systems disproportionately harm Black, minority, and overpoliced populations among others.

Properties of datasets that sometimes (but not always, and not in easily predictable ways) propagate downstream include imbalance, biases, stereotypes, and categorization. By imbalance we mean unequal representation of different groups. For example, Buolamwini and Gebru pointed out that two facial analysis benchmarks, IJB-A and Adience, overwhelmingly featured lighter-skinned subjects.[454] By dataset biases we mean incorrect associations, especially those corresponding to social and historical prejudices. For example, a dataset that measures arrests as a proxy for crime may reflect the biases of policing and discriminatory laws. By stereotypes we mean associations that accurately reflect a property of the world (or a specific culture at a specific point in time) that is thought to be the result of social and historical prejudice. For example, gender-occupation associations can be called stereotypes. By categorization we mean assigning discrete (often binary) labels to complex aspects of identity such as gender and race.[455]

Representational harms occur when systems reinforce the subordination of some groups along the lines of identity. Representational harms could be downstream harms — such as when models apply offensive labels to people from some groups — but they could be inherent in the dataset. For example, ImageNet contains numerous slurs and offensive labels inherited from WordNet and pornographic images of people who did not consent to their inclusion in the dataset.[456]

While downstream and representational harms are two categories that have drawn a lot of attention and criticism, there are many other harms that often arise including the environmental cost of training models on unnecessarily large datasets[457] and the erasure of the labor of subjects who contributed the data[458] or the annotators who labeled it.[459] For an overview of ethical concerns associated with datasets, see the survey by Paullada et al.[460]

Mitigating harms: an overview

Approaches for mitigating the harms associated with data are quickly developing. Here we review a few selected ideas.

[453] Hill, "Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match," *New York Times*, 2020.

[454] Buolamwini and Gebru, "Gender Shades."

[455] For a classic text on why categorization is a morally laden activity, see (Bowker and Leigh Star, *Sorting Things Out*). For a discussion of the harms of category systems embedded in machine learning datasets, see (Crawford, *The Atlas of AI* (Yale University Press, 2021)).

[456] Prabhu and Birhane, "Large Image Datasets: A Pyrrhic Win for Computer Vision?" *arXiv Preprint arXiv:2006.16923*, 2020; Yang et al., "Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the Imagenet Hierarchy," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, 547–58.

[457] Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, 610–23.

[458] Radin, "'Digital Natives'."

[459] Gray and Suri, *Ghost Work*.

[460] Paullada et al., "Data and Its (Dis) Contents: A Survey of Dataset Development and Use in Machine Learning Research," *arXiv Preprint arXiv:2012.05345*, 2020.

One approach targets the fact that many machine learning datasets are poorly documented, and details about their creation are often missing. This leads to a range of issues from lack of reproducibility and concerns of scientific validity to misuse and ethical concerns. In response, *datasheets for datasets* is a template and initiative to promote more detailed and systematic annotation for datasets.[461] A datasheet requires the creator of a dataset to answer questions relating to several areas of interest: Motivation, composition, collection process, preprocessing/cleaning/labeling, uses, distribution, maintenance. One goal is that process of creating a datasheet will help anticipate ethical issues with the dataset. But datasheets also aim to make data practices more reproducible, and help practitioners select more adequate data sources.

[461] Gebru et al., "Datasheets for Datasets," *arXiv Preprint arXiv:1803.09010*, 2018.

Going a step beyond datasheets, researchers Jo and Gebru[462] draw lessons from archival and library sciences for the construction and documentation of machine learning datasets. These lessons draw attention to issues of consent, inclusivity, power, transparency, ethics and privacy.

[462] Jo and Gebru, "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning," in *Proc. 3Rd Conference on Fairness, Accountability, and Transparency (FAccT)*, 2020, 306–16.

Other approaches stay within the paradigm of minimally curated data collection but aim to modify or sanitize content deemed problematic in datasets. The ImageNet creators have made efforts to remove slurs and harmful terms as well as categories considered non-imageable, or unable to be characterized using images. "Vegetarian" and "philanthropist" are two such categories that were removed.[463] The REVISE tool aims to partially automate the process of identifying various kinds of biases in visual datasets.[464]

[463] Yang et al., "Towards Fairer Datasets."

[464] Wang, Narayanan, and Russakovsky, "REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets," in *European Conference on Computer Vision* (Springer, 2020), 733–51.

Mitigating harms by separating the roles of datasets

Our analysis of the different roles datasets play allows greater clarity in mitigating harms while preserving benefits. This analysis is not intended as an alternative to the many approaches that have already been proposed for mitigating harms. Rather, it can sharpen our thinking and strengthen other harm-mitigation strategies.

Our main observation is that the reuse of scientific benchmark datasets in engineering pipelines complicates efforts to address biases and harms. Attempts to address harms in such dual-use datasets leaves creators with a conundrum. On the one hand, benchmark datasets need to be long-lived: many benchmark datasets created decades ago continue to be useful and widely used today. Thus, modifying a dataset down the line when new harms become known will compromise its scientific utility, as performance on the modified dataset may not be meaningfully comparable to performance on the older dataset.

On the other hand, attempting to anticipate all possible harms during dataset creation is infeasible if the dataset is going to be used as training or pre-training data. Experience shows that datasets turn out to be useful for an ever-expanding set of downstream tasks, some of which were not even conceived of at the time of dataset creation.

Better tradeoffs are possible if there is a clear separation between scientific benchmarks and production-oriented datasets. In cases where the same dataset can be potentially useful for both purposes, creators should consider making two versions or forks of the data, because many of the harm mitigation strategies that apply to one don't apply to the other, and vice versa.

To enforce this separation, benchmark dataset creators should consider avoiding the use of the dataset in production pipelines by explicitly prohibiting it in the terms of use. Currently the licenses of many benchmark datasets prohibit commercial uses. This restriction has a similar effect, but it is not the best way to make this distinction. After all, production models may be noncommercial: they may be built by researchers or governments, with the latter category having an especially high potential for harm. At the same time, prohibiting commercial uses is arguably too strict, as it prohibits the use of the dataset as a guide to model selection, a use that does not raise the same risks of downstream harm.

One reason why there are fairness interventions applicable to scientific benchmark datasets but not production datasets is that, as we've argued, most of the scientific utility of benchmarks is captured by the *relative* performance of models. The fact that interventions that hurt absolute performance may be acceptable gives greater leeway for harm mitigation efforts. Consider image classification benchmarks. We hypothesize that the relative ranking of models will be only minimally affected if the dataset is modified to remove all images containing people (keeping high-level properties including the number of classes and images the same). Such an intervention would avoid a wide swath of the harms associated with datasets while preserving much of its scientific utility.

Conversely, one reason why there are fairness interventions applicable to production datasets but not scientific benchmarks is that interventions for production datasets can be strongly guided by an understanding of their downstream impacts in specific applications. Language and images, in particular, capture such a variety of cultural stereotypes that sanitizing all of them has proved infeasible.[465] It is much easier to design interventions once we fix an application and the cultural context(s) in which it will be deployed. Different interventions may be applicable to the same dataset used in different applications. Unlike scientific benchmarks, dataset standardization is

[465] Gonen and Goldberg, "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings but Do Not Remove Them," *arXiv Preprint arXiv:1903.03862*, 2019.

not necessary in engineering settings.

In fact, the best locus of intervention even for dataset biases may be downstream of the data. For example, it has been observed for many years that online translation systems perpetuate gender stereotypes when translating gender-neutral pronouns. The text "O bir doctor. O bir hemşire." may be translated from Turkish to English as "He is a doctor. She is a nurse." Google Translate mitigated this by showing multiple translations in such cases.[466] Compared to data interventions, this has the benefit of making the potential bias (or, in some cases, erroneous translation) more visible to the user.

[466] Kuczmarski, "Reducing Gender Bias in Google Translate," *Google Blog* 6 (2018); Johnson, "A Scalable Approach to Reducing Gender Bias in Google Translate," *Google Blog*, 2020.

Our analysis points to many areas where further research could help clarify ethical implications. In particular, the pre-training role of benchmark datasets occupies a grey area where it is not clear when and to what extent data biases propagate to the target task/domain. Research on this area is nascent;[467] this research is vital because the (mis)use of scientific benchmarks for pre-training in production pipelines is common today and unlikely to cease in the near future.

[467] Steed and Caliskan, "Image Representations Learned with Unsupervised Pre-Training Contain Human-Like Biases," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, 701–13.

Datasets should not be seen as static, neutral technical artifacts. The harms that could arise from a dataset depend not just on its contents but also the rules, norms, and culture surrounding its usage. Thus, modifying these cultural practices is one potential way to mitigate harms. As we discussed above, lack of domain knowledge by dataset users has come to be seen almost as a virtue in machine learning. This attitude should be reconsidered as it has a tendency to accentuate ethical blind spots.

Datasets require stewardship, whether by the dataset creator or by another designated entity or set of entities. Consider the problem of derivatives: popular benchmark datasets are often extended by other researchers with additional features, and these derived datasets can introduce the possibility of harms not present in the original (to the same extent). For example, the Labeled Faces in the Wild (LFW) dataset of faces was annotated by other researchers with characteristics as race, gender, and attractiveness.[468] Regardless of the ethics of LFW itself, the derived dataset enables new applications that classify people by appearance in harmful ways.[469] Of course, not all derivatives are ethically problematic. Adjudicating and enforcing such ethical distinctions is only possible if there is a governance mechanism in place.

[468] Huang et al., "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments" (University of Massachusetts, Amherst, 2007); Kumar et al., "Describable Visual Attributes for Face Verification and Image Search," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, no. 10 (2011): 1962–77.

[469] The intended purpose of the derived dataset is to enable searching corpora of face images by describable attributes.

Beyond datasets

In this final section, we discuss important scientific and ethical questions that are relevant to datasets but also go beyond datasets, pervading machine learning: validity, problem framing, and limits to

prediction.

Lessons from measurement

Measurement theory is an established science with ancient roots. In short, measurement is about assigning numbers to objects in the real world in a way that reflects relationships between these objects. Measurement draws an important distinction between a *construct* that we wish to measure and the measurement procedure that we used to create a numerical representation of the construct.

For example, we can think of a well-designed math exam as measuring the mathematical abilities of a student. A student with greater mathematical ability than another is expected to score higher on the exam. Viewed this way, an exam is a *measurement procedure* that assigns numbers to students. The *mathematical ability* of a student is the construct we hope to measure. We desire that the ordering of these numbers reflects the sorting of students by their mathematical abilities. A measurement procedure operationalizes a construct.

Every prediction problem has a target variable, the thing we're trying to predict.[470] By viewing the target variable as a construct, we can apply measurement theory to understand what makes a good target variable.

[470] Recall that in a prediction problem we have covariates X from which we're trying to predict a variable Y. This variable Y is what we call the *target variable* in our prediction problem.

The choice of a poor target variable cannot be ironed out with additional data. In fact, the more data we feed into our model, the better it gets at capturing the flawed target variable. Improved data quality or diversity are no cure either.

All formal fairness criteria that involve the target variable, separation and sufficiency being two prominent examples[471], are either meaningless or downright misleading when the target variable itself is the locus of discrimination.

[471] Recall from Chapter 2 that separation requires the protected attribute to be independent of the prediction conditional on the target variable. Sufficiency requires the target variable to be independent of the protected attribute given the prediction.

But what makes a target variable good or bad? Let's get a better grasp on this question by considering a few examples.

1. Predicting the value of the Standard and Poor 500 Index (S&P 500) at the close of the New York Stock Exchange tomorrow.
2. Predicting whether an individual is going to default on a loan.
3. Predicting whether an individual is going to commit a crime.

The first example is rather innocuous. It references a fairly robust target variable, even though it relies on a number of social facts.

The second example is a common application of statistical modeling that underlies much of modern credit scoring in the United States. At first sight a default event seems like a clean cut target variable. But the reality is different. In a public dataset by the Federal Reserve[472] default events are coded by a so-called *performance* variable

[472] The Federal Reserve Board, "Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit."

that measures a *serious delinquency in at least one credit line of a certain time period*. More specifically, the report states that the

> measure is based on the performance of new or existing accounts and measures whether individuals have been late 90 days or more on one or more of their accounts or had a public record item or a new collection agency account during the performance period.[473]

[473] Quote from the Federal Reserve report.

Our third example runs into the most concerning measurement problem. How do we determine if an individual committed a crime? What we can determine with certainty is whether or not an individual was arrested and found guilty of a crime. But this depends crucially on who is likely to be policed in the first place and who is able to maneuver the criminal justice system successfully following an arrest.

Sorting out what a good target variable is, in full generality, can involve the whole apparatus of measurement theory. The scope of measurement theory, however, goes beyond defining reliable and valid target variables for prediction. Measurement comes in whenever we create features for a machine learning problem and should therefore be an essential part of the data creation process.[474]

[474] Jacobs and Wallach, "Measurement and Fairness," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, 375–85.

Judging the quality of a measurement procedure is a difficult task. Measurement theory has two important conceptual frameworks for arguing about what makes measurement *good*. One is *reliability*. The other is *validity*.

Reliability describes the differences observed in multiple measurements of the same object under identical conditions. Thinking of the measurement variable as a random variable, reliability is about the variance between independent identically distributed measurements. As such, reliability can be analogized with the statistical notion of variance.

Validity is concerned with how well the measurement procedure in principle captures the concept that we try to measure. If reliability is analogous to variance, it is tempting to see validity as analogous to bias. But the situation is a bit more complicated. There is no simple formal criterion that we could use to establish validity. In practice, validity is based to a large extent on human expertise and subjective judgments.

One approach to formalize validity is to ask how well a score predicts some external criterion. This is called *external validity*. For example, we could judge a measure of creditworthiness by how well it predicts default in a lending scenario. While external validity leads to concrete technical criteria, it essentially identifies good measurement with predictive accuracy. However, that's certainly not all there is to validity.

Construct validity is a framework for discussing validity that includes numerous different types of evidence. Messick highlights six aspects of construct validity:

- Content: How well does the content of the measurement instrument, such as the items on a questionnaire, measure the construct of interest?
- Substantive: Is the construct supported by a sound theoretical foundation?
- Structural: Does the score express relationships in the construct domain?
- Generalizability: Does the score generalize across different populations, settings, and tasks?
- External: Does the score successfully predict external criteria?
- Consequential: What are the potential risks of using the score with regards to bias, fairness, and distributive justice?

Of these different criteria, external validity is the one most familiar to the machine learning practitioner. But machine learning practice would do well to embrace the other, more qualitative, criteria as well. Ultimately, measurement forces us to grapple with the often surprisingly uncomfortable question: What are we even trying to do when we predict something?

Problem framing: comparisons with humans

A long-standing ambition of artificial intelligence research is to match or exceed human cognitive abilities by an algorithm. This desire often leads to comparisons between humans and machines on various tasks. Judgments about human accuracy often also enter the debate around when to use statistical models in high stakes decision making settings.

The comparison between human decision makers and statistical models is by no means new. For decades, researchers have compared the accuracy of human judgments with that of statistical models.[475]

Even within machine learning, the debate dates way back. A 1991 paper by Bromley and Säckinger explicitly compared the performance of artificial neural networks to a measure of human accuracy on the USPS digits dataset that predates the famous MNIST data.[476] A first experiment put the human accuracy at 2.5%, a second experiment found the number 1.51%, while a third reported the number 2.37%.[477]

Comparison with so-called human baselines has since become widely accepted in the machine learning community. The Electronic Frontier Foundation (EFF), for example, hosts a major repository of

[475] Dawes, Faust, and Meehl, "Clinical Versus Actuarial Judgment."

[476] Bromley and Sackinger, "Neural-Network and k-Nearest-Neighbor Classifiers."

[477] Chaaban and Scheessele, "Human Performance on the USPS Database," *Report, Indiana University South Bend*, 2007.

AI progress measures that compares the performance of machine learning models to reported human accuracies on numerous benchmarks.[478]

For the ILSVRC 2012 data, the reported human accuracy is 5.1%.[479] This often quoted number corresponds to the performance of a single human annotator who was "trained on 500 images and annotated 1500 test images".[480] A second annotator who was "trained on 100 images and then annotated 258 test images" achieved an accuracy of 12%. Based on this number of 5.1%, researchers announced in 2015 that their model was "the first to surpass human-level performance".[481] Not surprisingly, this claim received significant attention throughout the media.

However, a later more careful investigation into "human accuracy" on ImageNet revealed a very different picture.[482] The researchers found that only models from 2020 are actually on par with the strongest human labeler. Moreover, when restricting the data to 590 object classes out of 1000 classes in total, the best human labeler performed much better at less than 1% error than even the best predictive models. Recall, that the ILSVRC 2012 data featured 118 different dog breeds alone, some of which are extremely hard to distinguish for anyone who is not a trained dog expert. In fact, the researchers had to consult with experts from the American Kennel Club (AKC) to disambiguate challenging cases of different dog breeds. Simply removing dog classes alone increases the performance of the best human labeler to less than 1.3% error.

There is another troubling fact. Small variations in the data collection protocol turn out to have a significant effect on the performance of machine classifiers: "the accuracy scores of even the best image classifiers are still highly sensitive to minutiae of the data cleaning process."[483]

These results cast doubt not only on how me measure human accuracy, but also on the validity of the presumed theoretical construct of "human accuracy" itself. However, the machine learning community has adopted a rather casual approach to measuring human accuracy. Many researchers assume that the construct of *human accuracy* exists unambiguously and it is whatever number comes out of some ad-hoc testing protocol for some set of human beings. These ad-hoc protocols often result in anecdotal comparisons of questionable scientific value.

Invalid judgments about human performance relative to machines are not just a scientific error, they also have the potential to create narratives that support poor policy choices in high stakes policy questions around the use of predictive models in consequential decisions. For example, criminal justice policy is being driven by claims

[478] See https://www.eff.org/ai/metrics.

[479] To be precise, this number is referring to the fraction of times that the correct image label was not contained in the top 5 predicted labels of the model or human.

[480] Russakovsky et al., "Imagenet Large Scale Visual Recognition Challenge."

[481] He et al., "Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, 1026–34.

[482] Shankar et al., "Evaluating Machine Accuracy on Imagenet," in *International Conference on Machine Learning* (PMLR, 2020), 8634–44.

[483] Recht et al., "Do ImageNet Classifiers Generalize to ImageNet?"

that statistical methods are superior to judges at predicting risk of recidivism or failure to appear in court. However, these comparisons are dubious because judges are not solving pure prediction problems but rather incorporate other factors such as leniency towards younger defendants.[484]

[484] Stevenson and Doleac, "Algorithmic Risk Assessment in the Hands of Humans."

Problem framing: focusing on a single optimization objective

Real-life problems rarely involve optimizing a single objective and more commonly involve some kind of tradeoff between multiple objectives. How best to formulate this as a statistical optimization problem is both an art and a science. However, benchmark tasks, especially those with leaderboards, tend to pick a single objective. For high-profile benchmarks, the resulting "overfitting to the problem formulation" may result in scientific blind spots and limit the applicability of published findings to practical settings.

For example, it was well known at the time Netflix launched its Prize that recommendation is not just a matter of maximizing predictive accuracy and, even to the extent that it is, there isn't one single measure that's always appropriate.[485] Yet the contest focused purely on prediction accuracy evaluated by a single metric. A few years after the contest ended, Netflix revealed that most of the work that went into the leaderboard had not translated to production models. Part of the reason was that the contest did not capture the range of Netflix's objectives and constraints: the tight dependence of recommendations on the user interface; the fact that "users" are typically households made of members with differing tastes; explainability; freshness, and many more.[486]

[485] Herlocker et al., "Evaluating Collaborative Filtering Recommender Systems," *ACM Transactions on Information Systems (TOIS)* 22, no. 1 (2004): 5–53.

[486] Amatriain and Basilico, "Netflix Recommendations: Beyond the 5 Stars (Part 1)," *Netflix Tech Blog* 6 (2012).

If many of the insights from the leaderboard did not even generalize to Netflix's own production setting, the gap between Netflix and other recommendation-oriented platforms is far greater. Notably, as a movie platform, Netflix is unusual in that it has a relatively static inventory compared to those with user-generated content such as YouTube or Facebook. When the content pool is dynamic, a different class of algorithms is needed. The pull that the Netflix Prize exerted on recommender systems research may have diverted attention away from the latter type of algorithm for many years, although it is hard to know for sure because the counterfactual is unobservable.

Formal machine learning competitions, even if they cause blind spots due to the need to pick a single optimization objective, are at least carefully structured to promote scientific progress in some narrow sense. Arguably more damaging are the informal competitions that seems to inevitably emerge in the presence of a prominent benchmark dataset, resulting in unfortunate outcomes such as insightful

papers being rejected because they failed to beat the state of the art, or unoriginal papers being published because they did beat the state of the art by (scientifically insignificant) application of greater computing power.

Another downside to a field oriented around one-dimensional, competitive pursuit is that it becomes structurally difficult to address biases in models and classifiers. If a contestant takes steps to prevent dataset bias from propagating to their models, there will be an accuracy drop (because accuracy is judged on a biased dataset) and fewer people will pay attention to the work.

As fairness issues in machine learning have gained prominence, fairness-focused benchmarks datasets have proliferated, such as the Pilot Parliamentarians Benchmark for facial analysis[487] and the Equity Evaluation Corpus for sentiment analysis.[488] An advantage of this approach is that the scientific and cultural machinery of benchmark-oriented innovation can be repurposed for fairness research. A potential danger is Goodhart's law, which states, in its broad form, "When a measure becomes a target, it ceases to be a good measure." As we've emphasized in this book, fairness is multifaceted, and benchmarks can capture only narrow notions of fairness. While these can be useful diagnostics, if they are misconstrued as targets in their own right, then research that is focused on optimizing for these benchmarks may not result in fairness in a more substantive sense.

In addition to creating fairness-focused benchmarks, the algorithmic fairness community has also repurposed earlier benchmarks toward the study of fairness questions. Consider the Census dataset from the UCI repository discussed earlier. It originally gained popularity as a source of real-world data. Its use is acceptable for studying algorithmic questions such as, say, the relative strengths of decision trees and logistic regression. We expect the answers to be insensitive to issues like the cultural context of the data. But now it is being used for studying fairness questions such as how classification accuracy tends to vary by race or gender. For such questions, the answers are sensitive to the details of the subpopulations. Further, the classification task associated with the benchmark (prediction of income treated as a binary variable) is artificial and does not correspond to any real-life application. Thus, accuracy disparities (and other fairness-related measurements) may look different for a different task, or if the data had been sampled differently, or if it came from a different time or place. Using benchmark datasets to make generalizable claims about fairness requires careful attention to issues of context, sampling, and validity.

[487] Buolamwini and Gebru, "Gender Shades."

[488] Kiritchenko and Mohammad, "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems," in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics* (New Orleans, Louisiana: Association for Computational Linguistics, 2018), 43–53, https://doi.org/10.18653/v1/S18-2005.

Limits of data and prediction

Machine learning fails in many scenarios and it's important to understand the failure cases as much as the success stories.

The Fragile Families Challenge was a machine learning competition based on the Fragile Families and Child Wellbeing study (FFCWS).[489] Starting from a random sample of hospital births between 1998 and 2000, the FFCWS followed thousand of American families over the course of 15 years, collecting detailed information, about the families' children, their parents, educational outcomes, and the larger social environment. Once a family agreed to participate in the study, data were collected when the child was born, and then at ages 1, 3, 5, 9, and 15.

[489] Reichman et al., "Fragile Families: Sample and Design," *Children and Youth Services Review* 23, no. 4–5 (2001): 303–26.

The Fragile Families Challenge concluded in 2017. The underlying dataset for the competition contains 4242 rows, one for each family, and 12943 columns, one for each variable plus an ID number of each family. Of the 12942 variables, 2358 are constant (i.e., had the same value for all rows), mostly due to redactions for privacy and ethics concerns. Of the approximately 55 million (4242 x 12942) entries in the dataset, about 73% do not have a value. Missing values have many possible reasons, including non-response of surveyed families, drop out of study participants, as well as logical relationships between features that imply certain fields are missing depending on how others are set. There are six outcome variables, measured at age 15: *1) child grade point average (GPA), 2) child grit, 3) household eviction, 4) household material hardship, 5) caregiver layoff, and 6) caregiver participation in job training.*

The goal of the competition was to predict the value of the outcome variables at age 15 given the data from age 1 through 9. As is common for competitions, the challenge featured a three-way data split: training, leaderboard, and test sets. The training set is publicly available to all participants, the leaderboard data support a leaderboard throughout the competition, and the test set is used to determine a final winner.

The outcome of the prediction challenge was disappointing. Even the winning models performed hardly better than a simple baseline their predictions didn't differ much compared to predicting the mean of each outcome.

What caused the poor performance of machine learning on the fragile families data? One obvious possibility is that none of the contestants hit upon the right machine learning techniques for this task. But the fact that 160 teams of motivated experts submitted thousands of models over the course of five months makes this highly unlikely. Besides, models from disparate model classes all

made very similar (and equally erroneous) predictions, suggesting that learning algorithms weren't the limitation.[490,491] There are a few other technical possibilities that could explain the disappointing performance, including the sample size, the study design, and the missing values.

But there is also a more fundamental reason that remains plausible. Perhaps the dynamics of life trajectories are inherently unpredictable over the six year time delay between measurement of the covariates and measurement of the outcome. This six year gap, for example, included the Great Recession, a period of economic shocks and decline between 2007 and 2009, that might have changed trajectories in unforeseeable ways.

In fact, there's an important reason why even the performance of models in the challenge, dismal as they were, may overestimate what we can expect in a real-world setting. That's because the models were allowed to peek into the future, so to speak. The training and test sets were drawn from the same distribution and, in particular, the same time period, as is the standard practice in machine learning research. Thus, the data already incorporates information about the effect of the Great Recession and other global shocks during this period. In a real application, models must be trained on data from the past whereas predictions are about the future. Thus, there is always some drift — a change in the relationship between the covariates and the outcome. This puts a further limit on model performance.

Machine learning works best in a static and stable world where the past looks like the future. Prediction alone can be a poor choice when we're anticipating dynamic changes, or when we are trying to reason about the effect that hypothetical actions would have in the real world.

[490] Salganik et al., "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration."

[491] This highlights an advantage of the benchmark dataset approach over one with less standardization: even when there is a failure to make substantial progress on prediction, we can learn something valuable from that failure.

Summary

Benchmark datasets are central to machine learning. They play many roles including enabling algorithmic innovation, measuring progress, and providing training data. Since its systematization in the late 1980s, performance evaluation on benchmarks has gradually become a ubiquitous practice because it makes it harder for researchers to cheat intentionally or unintentionally.

But an excessive focus on benchmarks brings many drawbacks. Researchers spend prodigious amounts of effort optimizing models to achieve state of the art performance. The results are often both scientifically uninteresting and of little relevance to practitioners because benchmarks omit many real-world details. The approach also amplifies the harms associated with data including downstream

harms, representational harms, and privacy violations.

As we write this book, the benchmark approach is coming under scrutiny because of these ethical concerns. While the benefits and drawbacks of benchmarks are both well known, our overarching goal in this chapter has been to provide a single framework that can help analyze both. Our position is that the core of the benchmark approach is worth preserving, but we envision a future where benchmarks play a more modest role as one of many ways to advance knowledge. To mitigate the harms associated with data, we believe that substantial changes to the practices of dataset creation, use, and governance are necessary. We have outlined a few ways to do this, adding to the emerging literature on this topic.

Chapter notes

This chapter was developed and first published by Hardt and Recht in the textbook *Patterns, Predictions, and Actions: A Story about Machine Learning*.[492] With permission from the authors, we include a large part of the original text here with only slight modifications. We removed a significant amount of material on adaptive data analysis and the problem of overfitting in machine learning benchmarks. We added new material on the roles that datasets play, as well as discussion about fairness and ethical concerns relating to datasets.

The benefits of the benchmark dataset approach are discussed in a talk by Mark Liberman, who calls it the common task method.[493]

Adaptivity in holdout reuse was the subject of[494] and much subsequent work in the area of adaptive data analysis. Similar concerns go under the name of *inference after selection* in the statistics community.

The collection and use of large ad-hoc data sets (once referred to as "big data") has been scrutinized in several important works, see, for example, boyd and Crawford,[495] as well as Tufekci.[496] More recently, Couldry and Mejias[497] use the term *data colonialism* to emphasize the processes by which data are appropriated and marginalized communities are exploited through data collection.

Olteanu et al.[498] discuss biases, methodological pitfalls, and ethical questions in the context of social data analysis. In particular, the article provides comprehensive taxonomies biases and issues that can arise in the sourcing, collection, processing, and analysis of social data.

For an introduction to measurement theory, not specific to the social sciences, see the books by Hand.[499] The comprehensive textbook by Bandalos[500] focuses on applications to the social science, including a chapter on fairness.

[492] Hardt and Recht, *Patterns, Predictions, and Actions*.

[493] Liberman, "Reproducible Research and the Common Task Method," *Simmons Foundation Lecture Https://Www. Simonsfoundation. Org/Lecture/Reproducible-Research-and-Thecommon-Task-Method* 2 (2015).

[494] Dwork et al., "The Reusable Holdout: Preserving Validity in Adaptive Data Analysis," *Science* 349, no. 6248 (2015): 636–38.

[495] boyd and Crawford, "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon," *Information, Communication & Society* 15, no. 5 (2012): 662–79.

[496] Tufekci, "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls," in *Proc. 8Th International AAAI Conference on Weblogs and Social Media*, 2014; Tufekci, "Engineering the Public: Big Data, Surveillance and Computational Politics," *First Monday*, 2014.

[497] Couldry and Mejias, "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject," *Television & New Media* 20, no. 4 (2019): 336–49.

[498] Olteanu et al., "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries," *Frontiers in Big Data* 2 (2019): 13.

[499] Hand, *Measurement Theory and Practice: The World Through Quantification* (Wiley, 2010); Hand, *Measurement: A Very Short Introduction* (Oxford University Press, 2016).

[500] Bandalos, *Measurement Theory and Applications for the Social Sciences* (Guilford Publications, 2018).

Bibliography

Agan, Amanda, and Sonja Starr. "Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment." *The Quarterly Journal of Economics* 133, no. 1 (2017): 191–235.

Akbar, Amna. "An Abolitionist Horizon for Police (Reform)." *California Law Review* 108, no. 6 (2020).

Albright, Alex. "If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions." *Harvard John M. Olin Fellow's Discussion Paper* 85 (2019).

Ali, Muhammad, 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, no. CSCW (2019): 199.

Alkhatib, Ali, and Michael Bernstein. "Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13, 2019.

Altman, Andrew. "Discrimination." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2020. https://plato.stanford.edu/archives/win2020/entries/discrimination/; Metaphysics Research Lab, Stanford University, 2020.

Amatriain, Xavier, and Justin Basilico. "Netflix Recommendations: Beyond the 5 Stars (Part 1)." *Netflix Tech Blog* 6 (2012).

Amorim, Evelin, 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)*, 229–37, 2018.

Andreou, Athanasios, Oana Goga, Krishna Gummadi, Loiseau Patrick, and Alan Mislove. "AdAnalyst." https://adanalyst.mpi-sws.org/, 2017.

Angrist, Joshua D., and Pischke Jörn-Steffen. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2009.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine Bias." *Pro Publica*, 2016.

Angwin, Julia, Madeleine Varner, and Ariana Tobin. "Facebook Enabled Advertisers to Reach 'Jew Haters'." ProPublica. https://www.%20propublica.%20org/article/facebook-enabled-advertisers-to-reach-jew-haters, 2017.

"Announcing the Contract Worker Disparity Project." Tech Equity Collaborative, 2021.

Antecol, Heather, Kelly Bedard, and Jenna Stearns. "Equal but Inequitable: Who Benefits from Gender-Neutral Tenure Clock Stopping Policies?" *American Economic Review* 108, no. 9 (2018): 2420–41.

Arp, Daniel, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. "Dos and Don'ts of Machine Learning in Computer Security," October 2020. https://arxiv.org/abs/2010.09470v1.

Arrow, Kenneth. "The Theory of Discrimination." *Discrimination in Labor Markets* 3, no. 10 (1973): 3–33.

Ashkenas, Jeremy, Haeyoun Park, and Adam Pearce. "Even with Affirmative Action, Blacks and Hispanics Are More Underrepresented at Top Colleges Than 35 Years Ago." https://www.nytimes.com/interactive/2017/08/24/us/affirmative-action.html, 2017.

Aström, Karl Johan, and Richard M Murray. *Feedback Systems: An Introduction for Scientists and Engineers*. Princeton university press, 2010.

Ayres, Ian. "Three Tests for Measuring Unjustified Disparate Impacts in Organ Transplantation: The Problem of" Included Variable" Bias." *Perspectives in Biology and Medicine* 48, no. 1 (2005): 68–S87.

Ayres, Ian, Mahzarin Banaji, and Christine Jolls. "Race Effects on eBay." *The RAND Journal of Economics* 46, no. 4 (2015): 891–917.

Ayres, Ian, and Peter Siegelman. "Race and Gender Discrimination in Bargaining for a New Car." *The American Economic Review*, 1995, 304–21.

Bagwell, Kyle. "The Economic Analysis of Advertising." *Handbook of Industrial Organization* 3 (2007): 1701–844.

Baker, Kevin T. "World Processors: Computer Modeling, Global Environmentalism, and the Birth of Sustainable Development." PhD thesis, Northwestern University, 2019.

Bakshy, Eytan, Solomon Messing, and Lada A Adamic. "Exposure to Ideologically Diverse News and Opinion on Facebook." *Science* 348, no. 6239 (2015): 1130–32.

Bandalos, Deborah L. *Measurement Theory and Applications for the*

Social Sciences. Guilford Publications, 2018.

Barabas, Chelsea, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. "Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment." In *Conference on Fairness, Accountability and Transparency*, 62–76. PMLR, 2018.

Barocas, Solon. "Putting Data to Work." In *Data and Discrimination: Collected Essays*, edited by Seeta Peña Gangadharan Virginia Eubanks and Solon Barocas, 59–62. New America Foundation, 2014.

Barocas, Solon, and Karen Levy. "Privacy Dependencies." *Wash. L. Rev.* 95 (2020): 555.

Barocas, Solon, and Andrew D Selbst. "Big Data's Disparate Impact." *UCLA Law Review*, 2016.

Baron, Reuben M, and David A Kenny. "The Moderator–Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51, no. 6 (1986): 1173.

Barone, Guglielmo, and Sauro Mocetti. "Intergenerational Mobility in the Very Long Run: Florence 1427-2011." *Bank of Italy Temi Di Discussione (Working Paper) No* 1060 (2016).

Bashir, Muhammad Ahmad, Sajjad Arshad, William Robertson, and Christo Wilson. "Tracing Information Flows Between Ad Exchanges Using Retargeted Ads." In *USENIX Security Symposium 16*, 481–96, 2016.

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

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23, 2021.

Benjamin, Ruha. *Race After Technology*. Polity, 2019.

Bennett, James, and Stan Lanning. "The Netflix Prize." In *Proceedings of KDD Cup and Workshop*, 2007:35. New York, NY, USA., 2007.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *ArXiv e-Prints* 1703.09207 (2017).

Berkson, Joseph. "Limitations of the Application of Fourfold Table Analysis to Hospital Data." *International Journal of Epidemiology* 43, no. 2 (2014): 511–15.

Bertrand, Marianne, and Esther Duflo. "Field Experiments on Discrimination." In *Handbook of Economic Field Experiments*, 1:309–93. Elsevier, 2017.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How

Much Should We Trust Differences-in-Differences Estimates?" *The Quarterly Journal of Economics* 119, no. 1 (2004): 249–75.

Bertrand, Marianne, and Sendhil Mullainathan. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94, no. 4 (2004): 991–1013.

Bian, Lin, Sarah-Jane Leslie, and Andrei Cimpian. "Gender Stereotypes about Intellectual Ability Emerge Early and Influence Children's Interests." *Science* 355, no. 6323 (2017): 389–91.

Bickel, Peter J, Eugene A Hammel, J William O'Connell, and others. "Sex Bias in Graduate Admissions: Data from Berkeley." *Science* 187, no. 4175 (1975): 398–404.

Billsus, Daniel, Michael J Pazzani, and others. "Learning Collaborative Information Filters." In *Icml*, 98:46–54, 1998.

Bird, Sarah, 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.

Blagus, Rok, and Lara Lusa. "Joint Use of over- and Under-Sampling Techniques and Cross-Validation for the Development and Assessment of Prediction Models." *BMC Bioinformatics* 16, no. 1 (November 2015): 363. https://doi.org/10.1186/s12859-015-0784-9.

Blank, Rebecca M. "The Effects of Double-Blind Versus Single-Blind Reviewing: Experimental Evidence from the American Economic Review." *The American Economic Review*, 1991, 1041–67.

Blum, Avrim, and Moritz Hardt. "The Ladder: A Reliable Leaderboard for Machine Learning Competitions." In *International Conference on Machine Learning*, 1006–14. PMLR, 2015.

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

Bongers, Stephan, Jonas Peters, Bernhard Schölkopf, and Joris M. Mooij. "Theoretical Aspects of Cyclic Structural Causal Models." *arXiv.org Preprint* arXiv:1611.06221v2 (2018).

Bonham, Vence L, Shawneequa L Callier, and Charmaine D Royal. "Will Precision Medicine Move Us Beyond Race?" *The New England Journal of Medicine* 374, no. 21 (2016): 2003.

Bouk, Dan. *How Our Days Became Numbered: Risk and the Rise of the Statistical Individual*. University of Chicago Press, 2015.

Bouthillier, Xavier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, et al. "Accounting for Variance in Machine Learning Benchmarks." *Proceedings of Machine Learning and Systems* 3 (2021).

Bowker, Geoffrey C., and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. MIT Press, 2000.

Bowles, Samuel, and Rajiv Sethi. "Social Segregation and the Dynamics of Group Inequality," 2006.

Boyd, D. "White Flight in Networked Publics: How Race and Class Shaped American Teen Engagement with MySpace and Facebook. Nakamura l, Chow-White PA, Eds. Race After the Internet." *Race After the Internet*, 2012, 203–22.

boyd, danah, and Kate Crawford. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15, no. 5 (2012): 662–79.

Branwen, Gwern. "The Neural Net Tank Urban Legend," 2011.

Brayne, Sarah. *Predict and Surveil: Data, Discretion, and the Future of Policing*. Oxford University Press, USA, 2020.

Breiman, Leo, and others. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *Statistical Science* 16, no. 3 (2001): 199–231.

Bromley, J, and E Sackinger. "Neural-Network and k-Nearest-Neighbor Classifiers." *Rapport Technique*, 1991, 11359–910819.

Bullard, Robert Doyle, Glenn Steve Johnson, and Angel O Torres. *Highway Robbery: Transportation Racism & New Routes to Equity*. South End Press, 2004.

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Conference on Fairness, Accountability and Transparency*, 77–91, 2018.

Buranyi, Stephen. "How to Persuade a Robot That You Should Get the Job." Guardian, 2018.

Calders, Toon, Faisal Kamiran, and Mykola Pechenizkiy. "Building Classifiers with Independency Constraints." In *In Proc. IEEE ICDMW*, 13–18, 2009.

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan. "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases." *Science* 356, no. 6334 (2017): 183–86.

Campolo, Alex, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. "AI Now 2017 Report." *AI Now Institute at New York University*, 2017.

Carlini, Nicholas, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks." In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, 267–84, 2019.

Cartwright, Nancy. *Hunting Causes and Using Them, Too*. Cambridge University Press, 2006.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm,

and Noemie Elhadad. "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission." In *Proc. 21St ACM SIGKDD*, 1721–30, 2015.

Chaaban, Ibrahim, and Michael R Scheessele. "Human Performance on the USPS Database." *Report, Indiana University South Bend*, 2007.

Chaney, Allison JB, 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*, 224–32. ACM, 2018.

Chen, Le, Alan Mislove, and Christo Wilson. "Peeking Beneath the Hood of Uber." In *Proceedings of the 2015 Internet Measurement Conference*, 495–508. ACM, 2015.

Chetty, Raj, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. "Race and Economic Opportunity in the United States: An Intergenerational Perspective." *The Quarterly Journal of Economics* 135, no. 2 (2020): 711–83.

Chiappa, Silvia. "Path-Specific Counterfactual Fairness." In *Proc. 33Rd AAAI*, 33:7801–8, 2019.

Chiappa, Silvia, and William S. Isaac. "A Causal Bayesian Networks Viewpoint on Fairness." *Arxiv.org* arXiv:1907.06430 (2019).

Chohlas-Wood, Alex, Joe Nudell, Zhiyuan Jerry Lin, Julian Nyarko, and Sharad Goel. "Blind Justice: Algorithmically Masking Race in Charging Decisions." Technical report, 2020.

Chouldechova, Alexandra. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." In *Proc. 3Rd FATML*, 2016.

Chouldechova, Alexandra, 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*, 134–48, 2018.

Christodoulou, Evangelia, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, Jan Y. Verbakel, and Ben Van Calster. "A Systematic Review Shows No Performance Benefit of Machine Learning over Logistic Regression for Clinical Prediction Models." *Journal of Clinical Epidemiology* 110 (June 2019): 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004.

Church, Kenneth Ward. "Emerging Trends: A Tribute to Charles Wayne." *Natural Language Engineering* 24, no. 1 (2018): 155–60.

Citron, Danielle Keats. "Technological Due Process." *Wash. UL Rev.* 85 (2007): 1249.

Cleary, T Anne. "Test Bias: Prediction of Grades of Negro and White

Students in Integrated Colleges." *Journal of Educational Measurement* 5, no. 2 (1968): 115–24.

———. "Test Bias: Validity of the Scholastic Aptitude Test for Negro and White Students in Integrated Colleges." *ETS Research Bulletin Series* 1966, no. 2 (1966): i–23.

Coltrane, Scott, and Melinda Messineo. "The Perpetuation of Subtle Prejudice: Race and Gender Imagery in 1990s Television Advertising." *Sex Roles* 42, no. 5–6 (2000): 363–89.

Cook, Cody, Rebecca Diamond, Jonathan Hall, John A List, and Paul Oyer. "The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers." National Bureau of Economic Research, 2018.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. "Algorithmic Decision Making and the Cost of Fairness." *arXiv Preprint arXiv:1701.08230*, 2017.

Cortes, Corinna, and Vladimir Vapnik. "Support-Vector Networks." *Machine Learning* 20, no. 3 (1995): 273–97.

Couldry, Nick, and Ulises A. Mejias. "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject." *Television & New Media* 20, no. 4 (2019): 336–49.

Cowgill, Bo, Fabrizio Dell'Acqua, Samuel Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau. "Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing Ai Ethics." In *Proceedings of the 21st ACM Conference on Economics and Computation*, 679–81, 2020.

Crawford, Kate. *The Atlas of AI*. Yale University Press, 2021.

———. "The Hidden Biases in Big Data." *Harvard Business Review* 1 (2013).

———. "The Trouble with Bias." NIPS Keynote https://www.youtube.com/watch?v=fMym_BKWQzk, 2017.

Crawford, Kate, and Trevor Paglen. "Excavating AI: The Politics of Training Sets for Machine Learning." *Excavating AI (Www.excavating.ai)*, 2019.

Creel, Kathleen, and Deborah Hellman. "The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems." *Virginia Public Law and Legal Theory Research Paper*, no. 2021–13 (2021).

D'Onfro, Jillian. "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-w
html, 2019.

Danesi, Marcel. *Dictionary of Media and Communications*. Routledge, 2014.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. "Extraneous Factors in Judicial Decisions." *Proceedings of the National*

Academy of Sciences 108, no. 17 (2011): 6889–92.

Darlington, Richard B. "Another Look at 'Cultural Fairness'." *Journal of Educational Measurement* 8, no. 2 (1971): 71–82.

Dastin, Jeffrey. "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women." *Reuters*, 2018.

Datta, Amit, Michael Carl Tschantz, and Anupam Datta. "Automated Experiments on Ad Privacy Settings." *Proceedings on Privacy Enhancing Technologies* 2015, no. 1 (2015): 92–112.

Davidai, Shai, and Thomas Gilovich. "Building a More Mobile America—One Income Quintile at a Time." *Perspectives on Psychological Science* 10, no. 1 (2015): 60–71.

Dawes, Robyn M, David Faust, and Paul E Meehl. "Clinical Versus Actuarial Judgment." *Science* 243, no. 4899 (1989): 1668–74.

De-Arteaga, Maria, 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*, 120–28. ACM, 2019.

Deaton, Angus, and Nancy Cartwright. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210 (2018): 2–21.

DeCoste, Dennis, and Bernhard Schölkopf. "Training Invariant Support Vector Machines." *Machine Learning* 46, no. 1 (2002): 161–90.

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41, no. 6 (1990): 391–407.

Dell, Melissa. "The Persistent Effects of Peru's Mining Mita." *Econometrica* 78, no. 6 (2010): 1863–903.

DeMasi, Orianna, Konrad Kording, and Benjamin Recht. "Meaningless Comparisons Lead to False Optimism in Medical Machine Learning." *PLoS ONE* 12, no. 9 (September 2017). https://doi.org/10.1371/journal.pone.0184604.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "ImageNet: A Large-Scale Hierarchical Image Database." In *Proc. CVPR*, 2009.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A Large-Scale Hierarchical Image Database." In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–55. Ieee, 2009.

Dieterich, William, Christina Mendoza, and Tim Brennan. "COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity," 2016. https://www.documentcloud.org/documents/

2998391-ProPublica-Commentary-Final-070616.html.

Dillon, Eleanor Wiske, and Jeffrey Andrew Smith. "The Determinants of Mismatch Between Students and Colleges." National Bureau of Economic Research, 2013.

Dobbie, Will, Jacob Goldin, and Crystal Yang. "The Effects of Pre-Trial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges." National Bureau of Economic Research, 2016.

Duda, Richard O, Peter E Hart, and David G Stork. *Pattern Classification and Scene Analysis*. Vol. 3. Wiley New York, 1973.

Dwork, Cynthia, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. "The Reusable Holdout: Preserving Validity in Adaptive Data Analysis." *Science* 349, no. 6248 (2015): 636–38.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness Through Awareness." In *Proc. 3Rd ITCS*, 214–26, 2012.

Dwork, Cynthia, Aaron Roth, and others. "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends in Theoretical Computer Science* 9, no. 3–4 (2014): 211–407.

Dwork, Cynthia, Adam Smith, Thomas Steinke, and Jonathan Ullman. "Exposed! A Survey of Attacks on Private Data." *Annual Review of Statistics and Its Application* 4 (2017): 61–84.

Dynarski, Susan, CJ Libassi, Katherine Michelmore, and Stephanie Owen. "Closing the Gap: The Effect of a Targeted, Tuition-Free Promise on College Choices of High-Achieving, Low-Income Students." National Bureau of Economic Research, 2018.

Edelman, Benjamin, Michael Luca, and Dan Svirsky. "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment." *American Economic Journal: Applied Economics* 9, no. 2 (2017): 1–22.

Edelman, Lauren B. "Law at Work: The Endogenous Construction of Civil Rights." In *Handbook of Employment Discrimination Research*, 337–52. Springer, 2005.

Einhorn, Hillel J, and Alan R Bass. "Methodological Considerations Relevant to Discrimination in Employment Testing." *Psychological Bulletin* 75, no. 4 (1971): 261.

Ensign, Danielle, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. "Runaway Feedback Loops in Predictive Policing." *arXiv Preprint arXiv:1706.09847*, 2017.

Eren, Ozkan, and Naci Mocan. "Emotional Judges and Unlucky Juveniles." *American Economic Journal: Applied Economics* 10, no. 3 (2018): 171–205.

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile,*

Police, and Punish the Poor. St. Martin's Press, 2018.

Evans, Richard. "RA Fisher and the Science of Hatred." *The New Statesman*, 2020.

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. "Certifying and Removing Disparate Impact." In *Proc. 21St SIGKDD*. ACM, 2015.

Fellner, Jamie. "Race, Drugs, and Law Enforcement in the United States." *Stan. L. & Pol'y Rev.* 20 (2009): 257.

Fields, Karen E., and Barbara J. Fields. *Racecraft: The Soul of Inequality in American Life*. Verso, 2014.

Fiesler, Casey, Natalie Garrett, and Nathan Beard. "What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis." In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 289–95, 2020.

Fishbane, Alissa, Aurelie Ouss, and Anuj K Shah. "Behavioral Nudges Reduce Failure to Appear for Court." *Science* 370, no. 6517 (2020).

Fisher, Ronald A. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7, no. 2 (1936): 179–88.

Forrester, Jay W. "Counterintuitive Behavior of Social Systems." *Technological Forecasting and Social Change* 3 (1971): 1–22.

———. "System Dynamics, Systems Thinking, and Soft OR." *System Dynamics Review* 10, no. 2–3 (1994): 245–56.

———. "Urban Dynamics." *IMR; Industrial Management Review (Pre-1986)* 11, no. 3 (1970): 67.

Frankel, Marvin E. "Criminal Sentences: Law Without Order," 1973.

Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. "Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures." In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 1322–33, 2015.

Freeman, Jonathan B, Andrew M Penner, Aliya Saperstein, Matthias Scheutz, and Nalini Ambady. "Looking the Part: Social Status Cues Shape Race Perception." *PloS One* 6, no. 9 (2011): e25107.

Friedman, Batya, and Helen Nissenbaum. "Bias in Computer Systems." *ACM Transactions on Information Systems (TOIS)* 14, no. 3 (1996): 330–47.

Frucci, Adam. "HP Face-Tracking Webcams Don't Recognize Black People." https://gizmodo.com/hp-face-tracking-webcams-dont-recognize-black-people-5431190, 2009.

Funk, Simon. "Try This at Home." *Http://Sifter.org/~Simon/Journal/2006*, 2006.

Gandy, Oscar H. *Coming to Terms with Chance: Engaging Rational Discrimination and Cumulative Disadvantage*. Routledge, 2016.

Gao, Leo, Stella Biderman, Sid Black, Laurence Golding, Travis

Hoppe, Charles Foster, Jason Phang, et al. "The Pile: An 800gb Dataset of Diverse Text for Language Modeling." *arXiv Preprint arXiv:2101.00027*, 2020.

Garofolo, John S, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. "DARPA TIMIT Acoustic-Phonetic Continous Speech Corpus CD-ROM. NIST Speech Disc 1-1.1." *NASA STI/Recon Technical Report n* 93 (1993): 27403.

Garvie, Clare, Alvaro Bedoya, and Jonathan Frankle. "The Perpetual Line-up." *Georgetown Law: Center on Privacy and Technology.*, 2016.

Gates, Susan Wharton, Vanessa Gail Perry, and Peter M Zorn. "Automated Underwriting in Mortgage Lending: Good News for the Underserved?" *Housing Policy Debate* 13, no. 2 (2002): 369–91.

Ge, Yanbo, Christopher R Knittel, Don MacKenzie, and Stephen Zoepf. "Racial and Gender Discrimination in Transportation Network Companies." National Bureau of Economic Research, 2016.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. "Datasheets for Datasets." *arXiv Preprint arXiv:1803.09010*, 2018.

Gillespie, Tarleton. "Content Moderation, AI, and the Question of Scale." *Big Data & Society* 7, no. 2 (2020): 2053951720943234.

———. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, 2018.

———. "The Politics of 'Platforms'." *New Media & Society* 12, no. 3 (2010): 347–64.

Glasgow, Joshua, Sally Haslanger, Chike Jeffers, and Quayshawn Spencer. "What Is Race?: Four Philosophical Views," 2019.

Glymour, Clark. "Comment: Statistics and Metaphysics." *Journal of the American Statistical Association* 81, no. 396 (1986): 964–66.

Glymour, M Maria. "Using Causal Diagrams to Understand Common Problems in Social Epidemiology." *Methods in Social Epidemiology*, 2006, 393–428.

Golebiewski, M, and D Boyd. "Data Voids: Where Missing Data Can Easily Be Exploited." *Data & Society* 29 (2018).

Gonen, Hila, and Yoav Goldberg. "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings but Do Not Remove Them." *arXiv Preprint arXiv:1903.03862*, 2019.

Google re:work team. "Guide: Hire by Committee." https://rework.withgoogle.com/print/guides/6053596147744768/, 2021.

Gray, Mary L, and Siddharth Suri. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Eamon Dolan Books, 2019.

Green, Lisa J. *African American English: A Linguistic Introduction*. Cambridge University Press, 2002.

Greene, Daniel, Anna Lauren Hoffmann, and Luke Stark. "Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning." In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

Greiner, D. James, and Donald B. Rubin. "Causal Effects of Perceived Immutable Characteristics." *The Review of Economics and Statistics* 93, no. 3 (2011): 775–85.

Grother, Patrick J. "NIST Special Database 19." *Handprinted Forms and Characters Database, National Institute of Standards and Technology*, 1995, 10.

Hacking, Ian. "Making up People." *London Review of Books* 28, no. 16 (2006).

———. *The Social Construction of What?* Harvard University Press, 2000.

Hall, Anna B. AND Cook, Joshua D. AND O'Connell. "Predictors of Student Productivity in Biomedical Graduate School Applications." *PLOS ONE* 12, no. 1 (January 2017): 1–14.

Halligan, Steve, Douglas G. Altman, and Susan Mallett. "Disadvantages of Using the Area Under the Receiver Operating Characteristic Curve to Assess Imaging Tests: A Discussion and Proposal for an Alternative Approach." *European Radiology* 25, no. 4 (April 2015): 932–39.

Hand, David J. *Measurement: A Very Short Introduction*. Oxford University Press, 2016.

Hand, David J. *Measurement Theory and Practice: The World Through Quantification*. Wiley, 2010.

Hanna, Rema N, and Leigh L Linden. "Discrimination in Grading." *American Economic Journal: Economic Policy* 4, no. 4 (2012): 146–68.

Hannak, Aniko, 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*, 527–38. ACM, 2013.

Hardt, Moritz. "How Big Data Is Unfair." https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de, 2014.

Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of Opportunity in Supervised Learning." In *Proc. 29Th NIPS*, 3315–23, 2016.

Hardt, Moritz, and Benjamin Recht. *Patterns, Predictions, and Actions: A Story about Machine Learning*. https://mlstory.org, 2021. https://arxiv.org/abs/2102.05242.

Harvey, Adam, and Jules LaPlace. "MegaPixels: Origins, Ethics, and

Privacy Implications of Publicly Available Face Recognition Image Datasets," 2019. https://megapixels.cc/.

Harwell, Drew. "Doorbell-Camera Firm Ring Has Partnered with 400 Police Forces, Extending Surveillance Concerns." *Washington Post*, 2019.

Haslanger, Sally. *Resisting Reality: Social Construction and Social Critique*. Oxford University Press, 2012.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2017.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on Imagenet Classification." In *Proceedings of the IEEE International Conference on Computer Vision*, 1026–34, 2015.

Hellman, Deborah. "Sex, Causation, and Algorithms: How Equal Protection Prohibits Compounding Prior Injustice." *Washington University Law Review* 98, no. 2 (2020): 481–523.

———. "Two Concepts of Discrimination." *Va. L. Rev.* 102 (2016): 895.

Herlocker, Jonathan L, Joseph A Konstan, Loren G Terveen, and John T Riedl. "Evaluating Collaborative Filtering Recommender Systems." *ACM Transactions on Information Systems (TOIS)* 22, no. 1 (2004): 5–53.

Hern, Alex. "Flickr Faces Complaints over 'Offensive'auto-Tagging for Photos." *The Guardian* 20 (2015).

Hernán, Miguel A, David Clayton, and Niels Keiding. "The Simpson's paradox unraveled." *International Journal of Epidemiology* 40, no. 3 (March 2011): 780–85. https://doi.org/10.1093/ije/dyr041.

Hernán, Miguel, and James Robins. *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming, 2019.

Hill, Kashmir. "Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match." *New York Times*, 2020.

Hirschman, Daniel, and Isaac Ariail Reed. "Formation Stories and Causality in Sociology." *Sociological Theory* 32, no. 4 (2014): 259–82.

Hoffmann, Anna Lauren. "Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse." *Information, Communication & Society* 22, no. 7 (2019): 900–915.

Holland, Paul W. "Statistics and Causal Inference." *Journal of the American Statistical Association (JASA)* 81 (1986): 945–70.

Huang, Gary B., Manu Ramesh, Tamara Berg, and Erik Learned-Miller. "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments." University of Massachusetts, Amherst, 2007.

Humphrey, Linda L., Benjamin K. S. Chan, and Harold C. Sox. "Postmenopausal Hormone Replacement Therapy and the Primary Prevention of Cardiovascular Disease." *Annals of Internal Medicine* 137, no. 4 (August 2002): 273–84.

Huq, Aziz Z. "Racial Equity in Algorithmic Criminal Justice." *Duke LJ* 68 (2018): 1043.

Hutson, Jevan A, 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, no. CSCW (2018): 73.

Imbens, Guido W., and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Ingold, David, and Spencer Soper. "Amazon Doesn't Consider the Race of Its Customers. Should It?" https://www.bloomberg.com/graphics/2016-amazon-same-day/, 2016.

Ivanescu, A. E., P. Li, B. George, A. W. Brown, S. W. Keith, D. Raju, and D. B. Allison. "The Importance of Prediction Model Validation and Assessment in Obesity and Nutrition Research." *International Journal of Obesity* 40, no. 6 (June 2016): 887–94. https://doi.org/10.1038/ijo.2015.214.

Jackson, John W., and Tyler J. VanderWeele. "Decomposition Analysis to Identify Intervention Targets for Reducing Disparities." *Epidemiology*, 2018, 825–35.

Jacobs, Abigail Z, and Hanna Wallach. "Measurement and Fairness." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–85, 2021.

Jaquette, Ozan, and Karina Salazar. "Opinion | Colleges Recruit at Richer, Whiter High Schools - the New York Times." https://www.nytimes.com/interactive/2018/04/13/opinion/college-recruitment-rich-white.html, 2018.

Jo, Eun Seo, and Timnit Gebru. "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning." In *Proc. 3Rd Conference on Fairness, Accountability, and Transparency (FAccT)*, 306–16, 2020.

Joachims, Thorsten, Adith Swaminathan, and Tobias Schnabel. "Unbiased Learning-to-Rank with Biased Feedback." In *Proc. 10Th International Conference on Web Search and Data Mining*, 781–89. ACM, 2017.

Jobin, Anna, Marcello Ienca, and Effy Vayena. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 1, no. 9 (2019): 389–99.

Johnson, Melvin. "A Scalable Approach to Reducing Gender Bias in Google Translate." *Google Blog*, 2020.

Judd, Sarah. "Activities for Building Understanding: How Ai4all

Teaches AI to Diverse High School Students." In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education*, 633–34, 2020.

Kaggle. "The Hewlett Foundation: Automated Essay Scoring." https://www.kaggle.com/c/asap-aes, 2012.

Kahneman, Daniel, AM Rosenfield, L Gandhi, and T Blaser. "Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Makinghttps." *Harvard Business Review*, 2016.

Kalantari, Nima Khademi, and Ravi Ramamoorthi. "Deep High Dynamic Range Imaging of Dynamic Scenes." *ACM Trans. Graph* 36, no. 4 (2017): 144.

Kalluri, Ria. "The Values of Machine Learning." NeurIPS Queer in AI workshop, 2019.

Kamiran, Faisal, and Toon Calders. "Classifying Without Discriminating." In *Proc. 2Nd International Conference on Computer, Control and Communication*, 2009.

Kang, Sonia K, Katherine A DeCelles, András Tilcsik, and Sora Jun. "Whitened Resumes: Race and Self-Presentation in the Labor Market." *Administrative Science Quarterly* 61, no. 3 (2016): 469–502.

Karabel, Jerome. *The Chosen: The Hidden History of Admission and Exclusion at Harvard, Yale, and Princeton*. Houghton Mifflin Harcourt, 2005.

Kasy, Maximilian, and Rediet Abebe. "Fairness, Equality, and Power in Algorithmic Decision-Making." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 576–86, 2021.

Kaufman, Liad, Dani Lischinski, and Michael Werman. "Content-Aware Automatic Photo Enhancement." In *Computer Graphics Forum*, 31:2528–40. 8. Wiley Online Library, 2012.

Kaufman, Shachar, Saharon Rosset, Claudia Perlich, and Ori Stitelman. "Leakage in Data Mining: Formulation, Detection, and Avoidance." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6, no. 4 (2012): 1–21.

Kay, Matthew, Cynthia Matuszek, and Sean A Munson. "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations." In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3819–28. ACM, 2015.

Kilbertus, Niki, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. "Avoiding Discrimination Through Causal Reasoning." In *Proc. 30Th NIPS*, 656–66, 2017.

Kiritchenko, Svetlana, and Saif Mohammad. "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems." In *Proceedings of the Seventh Joint Conference on Lexical and Computa-*

tional Semantics, 43–53. New Orleans, Louisiana: Association for Computational Linguistics, 2018. https://doi.org/10.18653/v1/S18-2005.

Kiritchenko, Svetlana, and Saif M Mohammad. "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems." *arXiv Preprint arXiv:1805.04508*, 2018.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *arXiv Preprint arXiv:1609.05807*, 2016.

Klonick, Kate. "The New Governors: The People, Rules, and Processes Governing Online Speech." *Harv. L. Rev.* 131 (2017): 1598.

Kochhar, Rakesh, and Anthony Cilluffo. "Key Findings on the Rise in Income Inequality Within America's Racial and Ethnic Groups." *Pew Research Center*, 2018.

Koenecke, Allison, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. "Racial Disparities in Automated Speech Recognition." *Proceedings of the National Academy of Sciences* 117, no. 14 (2020): 7684–89.

Koh, Pang Wei, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, et al. "Wilds: A Benchmark of in-the-Wild Distribution Shifts." *arXiv Preprint arXiv:2012.07421*, 2020.

Kohler-Hausmann, Issa. "Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination." *Nw. UL Rev.* 113 (2018): 1163.

———. "Eddie Murphy and the Dangers of Counterfactual Causal Thinking about Detecting Racial Discrimination." *SSRN*, 2019.

Koren, Yehuda, Robert Bell, and Chris Volinsky. "Matrix Factorization Techniques for Recommender Systems." *Computer* 42, no. 8 (2009): 30–37.

Krieger, Nancy. "Discrimination and Health Inequities." *International Journal of Health Services* 44, no. 4 (2014): 643–710.

———. "Epidemiology and the People's Health: Theory and Context," 2011.

———. "On the Causal Interpretation of Race." *Epidemiology* 25, no. 6 (2014): 937.

Kuczmarski, James. "Reducing Gender Bias in Google Translate." *Google Blog* 6 (2018).

Kumar, Neeraj, Alexander Berg, Peter N Belhumeur, and Shree Nayar. "Describable Visual Attributes for Face Verification and Image Search." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, no. 10 (2011): 1962–77.

Kusner, Matt J., Joshua R. Loftus, Chris Russell, and Ricardo Silva.

"Counterfactual Fairness." In *Proc. 30Th NIPS*, 4069–79, 2017.

Lakens, Daniel. "Impossibly Hungry Judges." https://daniellakens.blogspot.com/2017/07/impossibly-hungry-judges.html, 2017.

Lakkaraju, Himabindu, 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*, 275–84. ACM, 2017.

Lambrecht, Anja, and Catherine Tucker. "Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads." *Management Science*, 2019.

Langley, Pat. "Machine Learning as an Experimental Science." Springer, 1988.

———. "The Changing Science of Machine Learning." Springer, 2011.

LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-Based Learning Applied to Document Recognition." *Proceedings of the IEEE* 86, no. 11 (1998): 2278–2324.

Lee, Min Kyung, 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*, 1603–12. ACM, 2015.

Leslie, Sarah-Jane, Andrei Cimpian, Meredith Meyer, and Edward Freeland. "Expectations of Brilliance Underlie Gender Distributions Across Academic Disciplines." *Science* 347, no. 6219 (2015): 262–65.

Levy, Karen, and Solon Barocas. "Designing Against Discrimination in Online Markets." *Berkeley Tech. LJ* 32 (2017): 1183.

Lewis, Mary A. "A Comparison of Three Models for Determining Test Fairness." Federal Aviation Administration Washington DC Office of Aviation Medicine, 1978.

Li, Xiaochang, and Mara Mills. "Vocal Features: From Voice Identification to Speech Recognition by Machine." *Technology and Culture* 60, no. 2 (2019): S129–60.

Liberman, Marc. "Reproducible Research and the Common Task Method." *Simmons Foundation Lecture Https://Www. Simonsfoundation. Org/Lecture/Reproducible-Research-and-Thecommon-Task-Method* 2 (2015).

Liberman, Mark. "Fred Jelinek." *Computational Linguistics* 36, no. 4 (2010): 595–99.

Liberman, Mark, and Charles Wayne. "Human Language Technology." *AI Magazine* 41, no. 2 (2020).

Lily Hu. "Direct Effects." https://phenomenalworld.org/analysis/direct-effects, 2020.

Lipsky, Michael. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Service*. Russell Sage Foundation, 2010.

Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. "Delayed Impact of Fair Machine Learning." In *Proc. 35Th ICML*, 3156–64, 2018.

Liu, Zicheng, Cha Zhang, and Zhengyou Zhang. "Learning-Based Perceptual Image Quality Improvement for Video Conferencing." In *Multimedia and Expo, 2007 IEEE International Conference on*, 1035–38. IEEE, 2007.

Louçã, Francisco. "Emancipation Through Interaction–How Eugenics and Statistics Converged and Diverged." *Journal of the History of Biology* 42, no. 4 (2009): 649–84.

Loury, Glenn C. "A Dynamic Theory of Racial Income Differences." Discussion paper, 1976.

Lum, Kristian, and William Isaac. "To Predict and Serve?" *Significance* 13, no. 5 (2016): 14–19.

Lundberg, Shelly, and Richard Startz. "On the Persistence of Racial Inequality." *Journal of Labor Economics* 16, no. 2 (1998): 292–323.

Mallon, Ron. "'Race': Normative, Not Metaphysical or Semantic." *Ethics* 116, no. 3 (2006): 525–51.

———. *The Construction of Human Kinds*. Oxford University Press, 2018.

Manthorpe, Rowland. "The Beauty.AI Robot Beauty Contest Is Back." Wired UK. https://www.wired.co.uk/article/robot-beauty-contest-beauty-ai, 2017.

Martin, C Dianne, Chuck Huff, Donald Gotterbarn, and Keith Miller. "Implementing a Tenth Strand in the CS Curriculum." *Communications of the ACM* 39, no. 12 (1996): 75–84.

Martineau, Paris. "Cities Examine Proper—and Improper—Uses of Facial Recognition | WIRED." https://www.wired.com/story/cities-examine-proper-improper-facial-recognition/, 2019.

Massey, Douglas S, Jonathan Rothwell, and Thurston Domina. "The Changing Bases of Segregation in the United States." *The Annals of the American Academy of Political and Social Science* 626, no. 1 (2009): 74–90.

McEntegart, Jane. "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.

Meadows, Donella H., Jorgan Randers, and Dennis Meadows. *The Limits to Growth: The 30-Year Update*. Routledge, 2012.

Mehrotra, Rishabh, 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*,

626–33, 2017.

Merill, Jeremy. "Does Facebook Still Sell Discriminatory Ads?" The Markup, 2020.

Miller, George A. "WordNet: A Lexical Database for English." *Communications of the ACM* 38, no. 11 (1995): 39–41.

———. *WordNet: An Electronic Lexical Database*. MIT press, 1998.

Mitchell, Tom M. *The Need for Biases in Learning Generalizations*. Department of Computer Science, Laboratory for Computer Science Research . . . , 1980.

Moneta-Koehler, Abigail M. AND Petrie, Liane AND Brown. "The Limitations of the GRE in Predicting Success in Biomedical Graduate School." *PLOS ONE* 12, no. 1 (January 2017): 1–17.

Morgan, Stephen L., and Christopher Winship. *Counterfactuals and Causal Inference*. Cambridge University Press, 2014.

Mouw, Ted. "Are Black Workers Missing the Connection? The Effect of Spatial Distance and Employee Referrals on Interfirm Racial Segregation." *Demography* 39, no. 3 (2002): 507–28.

Munoz, Cecilia, Megan Smith, and D Patil. "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights." *Executive Office of the President. The White House*, 2016.

Muthukumar, Vidya, 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.

Myrdal, Gunnar. *An American Dilemma: The Negro Problem and Modern Democracy, Volume 2*. Routledge, 2017.

Nabi, Razieh, and Ilya Shpitser. "Fair Inference on Outcomes." In *Proc. 32Nd AAAI*, 1931–40, 2018.

Nanayakkara, Priyanka, Jessica Hullman, and Nicholas Diakopoulos. "Unpacking the Expressed Consequences of AI Research in Broader Impact Statements." *arXiv Preprint arXiv:2105.04760*, 2021.

Narayanan, Arvind, and Vitaly Shmatikov. "Robust de-Anonymization of Large Sparse Datasets." In *2008 IEEE Symposium on Security and Privacy (Sp 2008)*, 111–25. IEEE, 2008.

Natasha Tiku. "Google's Approach to Historically Black Schools Helps Explain Why There Are Few Black Engineers in Big Tech." Washington Post, 2021.

Neckerman, Kathryn M, and Joleen Kirschenman. "Hiring Strategies, Racial Bias, and Inner-City Workers." *Social Problems* 38, no. 4 (1991): 433–47.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. nyu Press, 2018.

Norton, Helen. "The Supreme Court's Post-Racial Turn Towards a

Zero-Sum Understanding of Equality." *Wm. & Mary L. Rev.* 52 (2010): 197.

O'Neil, Cathy. "How Algorithms Rule Our Working Lives." *The Guardian* 16 (2016).

———. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, 2016.

O'TOOLE, ALICE J, KENNETH DEFFENBACHER, Hervé Abdi, and JAMES C BARTLETT. "Simulating the 'Other-Race Effect'as a Problem in Perceptual Learning." *Connection Science* 3, no. 2 (1991): 163–78.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366, no. 6464 (2019): 447–53.

Ojala, Markus, and Gemma C Garriga. "Permutation Tests for Studying Classifier Performance." *Journal of Machine Learning Research* 11, no. Jun (2010): 1833–63.

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries." *Frontiers in Big Data* 2 (2019): 13.

Pager, Devah. "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, no. 1 (2007): 104–33.

Pager, Devah, and Hana Shepherd. "The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets." *Annu. Rev. Sociol* 34 (2008): 181–209.

Palamar, Joseph J, Shelby Davies, Danielle C Ompad, Charles M Cleland, and Michael Weitzman. "Powder Cocaine and Crack Use in the United States: An Examination of Risk for Arrest and Socioeconomic Disparities in Use." *Drug and Alcohol Dependence* 149 (2015): 108–16.

Paluck, Elizabeth Levy, and Donald P Green. "Prejudice Reduction: What Works? A Review and Assessment of Research and Practice." *Annual Review of Psychology* 60 (2009): 339–67.

Paluck, Elizabeth Levy, Seth A Green, and Donald P Green. "The Contact Hypothesis Re-Evaluated." *Behavioural Public Policy* 3, no. 2 (2019): 129–58.

Paluck, Elizabeth Levy, Roni Porat, Chelsey S Clark, and Donald P Green. "Prejudice Reduction: Progress and Challenges." *Annual Review of Psychology* 72 (2020).

Pariser, Eli. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin UK, 2011.

Pasquale, Frank. *The Black Box Society: The Secret Algorithms That*

Control Money and Information. Harvard University Press, 2015.

Passi, Samir, and Solon Barocas. "Problem Formulation and Fairness." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 39–48. ACM, 2019.

Paullada, Amandalynne, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. "Data and Its (Dis) Contents: A Survey of Dataset Development and Use in Machine Learning Research." *arXiv Preprint arXiv:2012.05345*, 2020.

Pearl, Judea. *Causality*. Cambridge University Press, 2009.

Pearl, Judea, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016.

Pearl, Judea, and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

Pedreschi, Dino, Salvatore Ruggieri, and Franco Turini. "Discrimination-Aware Data Mining." In *Proc. 14Th ACM SIGKDD*, 560–68. ACM, 2008.

Pendall, Rolf, and Carl Hedman. "Worlds Apart: Inequality Between America's Most and Least Affluent Neighborhoods." *Urban Institute*, 2015.

Peschard, Isabelle F., and Bas C. Van Fraassen. *The Experimental Side of Modeling*. University of Minnesota Press, 2018.

Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference*. MIT Press, 2017.

Phelps, Edmund S. "The Statistical Theory of Racism and Sexism." *The American Economic Review* 62, no. 4 (1972): 659–61.

Pischke, Jorn-Steffen. "Empirical Methods in Applied Economics: Lecture Notes," 2005.

Plaugic, Lizzie. "FaceApp's Creator Apologizes for the App's Skin-Lightening 'Hot' Filter." The Verge. https://www.theverge.com/2017/4/25/15419522/faceapp-hot-filter-racist-apology, 2017.

Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. "On Fairness and Calibration." In *Proc. 30Th NIPS*, 2017.

Posselt, Julie R. *Inside Graduate Admissions*. Harvard University Press, 2016.

Prabhu, Vinay Uday, and Abeba Birhane. "Large Image Datasets: A Pyrrhic Win for Computer Vision?" *arXiv Preprint arXiv:2006.16923*, 2020.

"Pre-Employment Inquiries and Marital Status or Number of Children." U.S. Equal Employment Opportunity Commission, 2021.

Quillian, Lincoln, 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, no. 41 (2017): 10870–75.

Radin, Joanna. "'Digital Natives': How Medical and Indigenous Histories Matter for Big Data." *Osiris* 32, no. 1 (2017): 43–64.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. "Mitigating Bias in Algorithmic Employment Screening: Evaluating Claims and Practices." *arXiv Preprint arXiv:1906.09208*, 2019.

Ramineni, Chaitanya, 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, no. 1 (2018): 1–31.

Raso, Jennifer. "Displacement as Regulation: New Regulatory Technologies and Front-Line Decision-Making in Ontario Works." *Canadian Journal of Law and Society* 32, no. 1 (2017): 75–95.

Recht, Benjamin, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. "Do ImageNet Classifiers Generalize to ImageNet?" In *Proc. 36th ICML*, 2019.

Reichman, Nancy E, Julien O Teitler, Irwin Garfinkel, and Sara S McLanahan. "Fragile Families: Sample and Design." *Children and Youth Services Review* 23, no. 4–5 (2001): 303–26.

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability." https://ainowinstitute.org/aiareport2018.pdf, 2018.

Rivera, Lauren A. *Pedigree: How Elite Students Get Elite Jobs*. Princeton University Press, 2016.

Robertson, Ronald E, 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, no. CSCW (2018): 148.

Rock, David, and Heidi Grant. "Why Diverse Teams Are Smarter." Harvard Business Review. https://hbr.org/2016/11/why-diverse-teams-are-smarter, 2016.

Rosenblat, Alex, Karen EC Levy, Solon Barocas, and Tim Hwang. "Discriminating Tastes: Customer Ratings as Vehicles for Bias." *Data & Society*, 2016, 1–21.

Rosenblatt, Frank. "Perceptron Simulation Experiments." *Proceedings of the IRE* 48, no. 3 (1960): 301–9.

Roth, Alvin E. "The Origins, History, and Design of the Resident Match." *Jama* 289, no. 7 (2003): 909–12.

Roth, Lorna. "Looking at Shirley, the Ultimate Norm: Colour Balance, Image Technologies, and Cognitive Equity." *Canadian Journal of Communication* 34, no. 1 (2009): 111.

Rothstein, Richard. *The Color of Law: A Forgotten History of How Our Government Segregated America*. Liveright Publishing, 2017.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, et al. "Imagenet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115, no. 3 (2015): 211–52.

Russell, Chris, Matt J. Kusner, Joshua R. Loftus, and Ricardo Silva. "When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness." In *Proc. 30Th NIPS*, 6417–26, 2017.

Saitta, Lorenza, and Filippo Neri. "Learning in the 'Real World'." *Machine Learning* 30, no. 2 (1998): 133–63.

Salganik, Matthew. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press, 2019.

Salganik, Matthew J, Ian Lundberg, Alexander T Kindel, Caitlin E Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M Altschul, et al. "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration." *Proceedings of the National Academy of Sciences* 117, no. 15 (2020): 8398–8403.

Salzberg, Steven L. "On Comparing Classifiers: A Critique of Current Research and Methods." *Data Mining and Knowledge Discovery* 1, no. 1 (1999): 1–12.

Sandvig, C., 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.

Sap, Maarten, 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*, 1668–78, 2019.

Semega, Jessica L, Kayla R Fontenot, and Melissa A Kollar. "Income and Poverty in the United States: 2016." *Current Population Reports*, no. P60–259 (2017).

Shankar, Shreya, 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.

Shankar, Vaishaal, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. "Evaluating Machine Accuracy on Imagenet." In *International Conference on Machine Learning*, 8634–44. PMLR, 2020.

Simoiu, Camelia, Sam Corbett-Davies, and Sharad Goel. "The Problem of Infra-Marginality in Outcome Tests for Discrimination." *The Annals of Applied Statistics* 11, no. 3 (2017): 1193–1216.

Simonite, Tom. "When It Comes to Gorillas, Google Photos Remains Blind." *Wired, January* 13 (2018).

Simpson, Edward H. "The Interpretation of Interaction in Contingency Tables." *Journal of the Royal Statistical Society: Series B (Methodological)* 13, no. 2 (1951): 238–41.

Singer-Vine, Jeremy, Jennifer Valentino-DeVries, and Ashkan Soltani. "How the Journal Tested Prices and Deals Online." Wall Street Journal. http://blogs.%20wsj.%20com/digits/2012/12/23/how-the-journal-tested-prices-and-deals-online, 2012.

Sloane, Mona, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. "Participation Is Not a Design Fix for Machine Learning." *arXiv Preprint arXiv:2007.02423*, 2020.

Small, Mario L, and Devah Pager. "Sociological Perspectives on Racial Discrimination." *Journal of Economic Perspectives* 34, no. 2 (2020): 49–67.

Solaiman, Irene, 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.

Spirtes, Peter, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, Prediction, and Search*. MIT press, 2000.

Sprietsma, Maresa. "Discrimination in Grading: Experimental Evidence from Primary School Teachers." *Empirical Economics* 45, no. 1 (2013): 523–38.

Squires, Gregory D, and others. *Capital and Communities in Black and White: The Intersections of Race, Class, and Uneven Development*. Suny Press, 1994.

Steed, Ryan, and Aylin Caliskan. "Image Representations Learned with Unsupervised Pre-Training Contain Human-Like Biases." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 701–13, 2021.

Stevenson, Megan. "Assessing Risk Assessment in Action." *Minn. L. Rev.* 103 (2018): 303.

Stevenson, Megan T, and Jennifer L Doleac. "Algorithmic Risk Assessment in the Hands of Humans." *Available at SSRN 3489440*, 2019.

Stevenson, Megan T, and Sandra G Mayson. "Pretrial Detention and the Value of Liberty." *Virginia Public Law and Legal Theory Research Paper*, no. 2021–14 (2021).

Susser, Daniel, Beate Roessler, and Helen Nissenbaum. "Online Manipulation: Hidden Influences in a Digital World." *Available at SSRN 3306006*, 2018.

Sweeney, Latanya. "Discrimination in Online Ad Delivery." *Queue* 11, no. 3 (March 2013): 10:10–29.

Tatman, Rachael. "Gender and Dialect Bias in YouTube's Automatic

Captions." In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 53–59. Valencia, Spain: Association for Computational Linguistics, 2017. https://doi.org/10.18653/v1/W17-1606.

The Federal Reserve Board. "Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit." https://www.federalreserve.gov/boarddocs/rptcongress/creditscore/, 2007.

Thebault-Spieker, Jacob, 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, no. 3 (2017): 21.

Thorndike, Robert L. "Concepts of Culture-Fairness." *Journal of Educational Measurement* 8, no. 2 (1971): 63–70.

Tjaden, Jasper Dag, Carsten Schwemmer, and Menusch Khadjavi. "Ride with Me—Ethnic Discrimination, Social Markets, and the Sharing Economy." *European Sociological Review* 34, no. 4 (2018): 418–32.

Torralba, Antonio, and Alexei A Efros. "Unbiased Look at Dataset Bias." In *Proc. CVPR*, 1521–28. IEEE, 2011.

Tripodi, Francesca. "Searching for Alternative Facts: Analyzing Scriptural Inference in Conservative News Practices." *Data & Society*, 2018.

Tufekci, Zeynep. "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls." In *Proc. 8Th International AAAI Conference on Weblogs and Social Media*, 2014.

———. "Engineering the Public: Big Data, Surveillance and Computational Politics." *First Monday*, 2014.

Turow, Joseph, 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.

"US prisoners released early by software bug." BBC news, 2015.

Valencia Caicedo, Felipe. "The Mission: Human Capital Transmission, Economic Persistence, and Culture in South America." *The Quarterly Journal of Economics* 134, no. 1 (2019): 507–56.

Valentino-Devries, Jennifer, Jeremy Singer-Vine, and Ashkan Soltani. "Websites Vary Prices, Deals Based on Users' Information." *Wall Street Journal* 10 (2012): 60–68.

VanderWeele, Tyler J., and Whitney R. Robinson. "On Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables." *Epidemiology*, 2014.

Vandewiele, Gilles, Isabelle Dehaene, György Kovács, Lucas Sterckx, Olivier Janssens, Femke Ongenae, Femke De Backere, et al.

"Overly Optimistic Prediction Results on Imbalanced Data: A Case Study of Flaws and Benefits When Applying over-Sampling." *Artificial Intelligence in Medicine* 111 (January 2021): 101987. https://doi.org/10.1016/j.artmed.2020.101987.

Venkatadri, Giridhari, Elena Lucherini, Piotr Sapiezynski, and Alan Mislove. "Investigating Sources of PII Used in Facebook's Targeted Advertising." *Proceedings on Privacy Enhancing Technologies* 2019, no. 1 (2019): 227–44.

Vries, Terrance de, 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*, 52–59, 2019.

Wang, Angelina, Arvind Narayanan, and Olga Russakovsky. "REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets." In *European Conference on Computer Vision*, 733–51. Springer, 2020.

Wasserman, Larry. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2010.

Weinshall-Margel, Keren, and John Shapard. "Overlooked Factors in the Analysis of Parole Decisions." *Proceedings of the National Academy of Sciences* 108, no. 42 (2011): E833–33.

West, Candace, and Don H Zimmerman. "Doing Gender." *Gender & Society* 1, no. 2 (1987): 125–51.

Whelan, Robert, and Hugh Garavan. "When Optimism Hurts: Inflated Predictions in Psychiatric Neuroimaging." *Biological Psychiatry*, Mechanisms of Aging and Cognition, 75, no. 9 (May 2014): 746–48. https://doi.org/10.1016/j.biopsych.2013.05.014.

"Whistleblowers: Software Bug Keeping Hundreds of Inmates in Arizona Prisons Beyond Release Dates." KJZZ, 2021.

Whittaker, Meredith, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Mysers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. *AI Now Report 2018*. AI Now Institute at New York University New York, 2018.

Wienk, Ronald E, Clifford E. Reid, John C. Simonson, and Frederick J. Eggers. "Measuring Racial Discrimination in American Housing Markets: The Housing Market Practices Survey." 1979.

Wikipedia contributors. "Certified Software Development Professional — Wikipedia, the Free Encyclopedia," 2021. https://en.wikipedia.org/w/index.php?title=Certified_Software_Development_Professional&oldid=1021729741.

———. "Same-Sex Marriage — Wikipedia, the Free Encyclopedia," 2021. https://en.wikipedia.org/w/index.php?title=Same-sex_marriage&oldid=1027216965.

Williams, Wendy M, 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, no. 17 (2015): 5360–65.

Wilson, Benjamin, Judy Hoffman, and Jamie Morgenstern. "Predictive Inequity in Object Detection." *arXiv Preprint arXiv:1902.11097*, 2019.

Wilson, James F, Michael E Weale, Alice C Smith, Fiona Gratrix, Benjamin Fletcher, Mark G Thomas, Neil Bradman, and David B Goldstein. "Population Genetic Structure of Variable Drug Response." *Nature Genetics* 29, no. 3 (2001): 265.

Winner, Langdon. *Do Artifacts Have Politics?* Routledge, 2017.

Woodworth, Blake E., Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. "Learning Non-Discriminatory Predictors." In *Proc. 30Th COLT*, 1920–53, 2017.

Wu, Tim. *The Master Switch: The Rise and Fall of Information Empires*. Vintage, 2010.

Yadav, Chhavi, and Léon Bottou. "Cold Case: The Lost Mnist Digits." *arXiv Preprint arXiv:1905.10498*, 2019.

Yang, Kaiyu, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. "Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the Imagenet Hierarchy." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 547–58, 2020.

Yao, Sirui, and Bert Huang. "Beyond Parity: Fairness Objectives for Collaborative Filtering." In *Advances in Neural Information Processing Systems*, 2921–30, 2017.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gómez Rodriguez, and Krishna P. Gummadi. "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment." In *Proc. 26Th WWW*, 2017.

Zemel, Richard S., Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. "Learning Fair Representations." In *Proc. 30Th ICML*, 2013.

Zhang, Junzhe, and Elias Bareinboim. "Fairness in Decision-Making — the Causal Explanation Formula." In *Proc. 32Nd AAAI*, 2018.

Zhang, Lu, Yongkai Wu, and Xintao Wu. "A Causal Framework for Discovering and Removing Direct and Indirect Discrimination." In *Proc. 26Th IJCAI*, 3929–35, 2017.

List of Figures

List of Tables