Fairness and Machine Learning

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

DO NOT SHARE

Incomplete working draft

Created: Tue Sep 10 08:04:00 PDT 2019
Latest public version available at https://fairmlbook.org

Contents

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, and Transparency in socio-technical systems (FAT*) 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 FAT* 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. His research aims to make the practice of machine learning more robust, reliable, and aligned with societal values. After obtaining a PhD in Computer Science from Princeton University in 2011, Hardt was a postdoctoral scholar and research staff member at IBM Research Almaden, followed by two years as a research scientist at Google Research and Google Brain. Together with Solon Barocas, Hardt co-founded the workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML) in 2014.

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. He co-created a Massive Open Online Course as well as a textbook on Bitcoin and cryptocurrency technologies. His doctoral research showed the fundamental limits of de-identification, for which he received the Privacy Enhancing Technologies Award.

Thanks and acknowledgements

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

We are greatful 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 Andrew Brunskill, Frances Ding, Michaela Hardt, Lily Hu, Ben Hutchinson, Niki Kilbertus, Issa Kohler-Hausmann, Eric Lawrence, Zachary Lipton, Lydia T. Liu, John Miller, Smitha Milli, Shira Mitchell, Robert Netzorg, Juan Carlos Perdomo, Ludwig Schmidt, Tijana Zrnic.

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] Robyn M Dawes, David Faust, and Paul E 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 be considered to be discrimination. In turn, two key questions to ask are whether the disparities are justified and whether they are harm-

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

ful. 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.

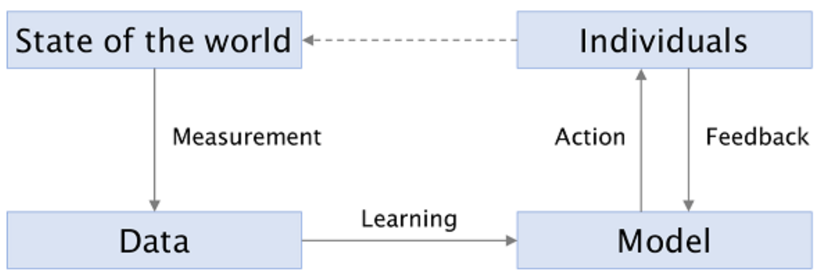


Figure 1: The machine learning loop

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]

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

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

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

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.

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

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

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.

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.

[10] Jeremy Ashkenas, 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).

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.

[11] Solon Barocas and Andrew D. Selbst, "Big Data's Disparate Impact," *California Law Review* 104 (2016).

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 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.

[12] Barocas and Selbst.

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.

[13] Lizzie 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); Rowland Manthorpe, "The Beauty.ai Robot Beauty Contest Is Back" (Wired UK. `https://www.wired.co.uk/article/robot-beauty-contest-beauty-ai`, 2017).

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

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

[15] George A 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.

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

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.

[17] Antonio Torralba and Alexei A Efros, "Unbiased Look at Dataset Bias," in *Proc. CVPR* (IEEE, 2011), 1521–8.

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

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

[19] Vence L Bonham, 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; James

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

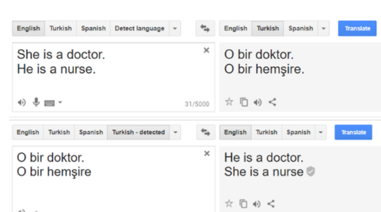
We've seen that training data reflects the disparities, distortions, and biases from the real world and the measurement process. This leads to an obvious question: when we learn a model from such data, are these disparities preserved, mitigated, or exacerbated?

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 English, the system picks the sentence that best matches the statistics

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



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]

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.

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]

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]

When we develop machine-learning models, we typically only test their overall accuracy; so a "5% error" statistic might hide the

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

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

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

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

[25] Hardt.

[26] Hardt.

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 subpopulations 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] Rich 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] Thorsten Joachims, Adith Swaminathan, and Tobias 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] Latanya Sweeney, "Discrimination in Online Ad Delivery," *Queue* 11, no. 3 (March 2013): 10:10–10:29.

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] Will Dobbie, 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).

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 Pred-

Pol, 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] Kristian Lum and William 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] Danielle 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.

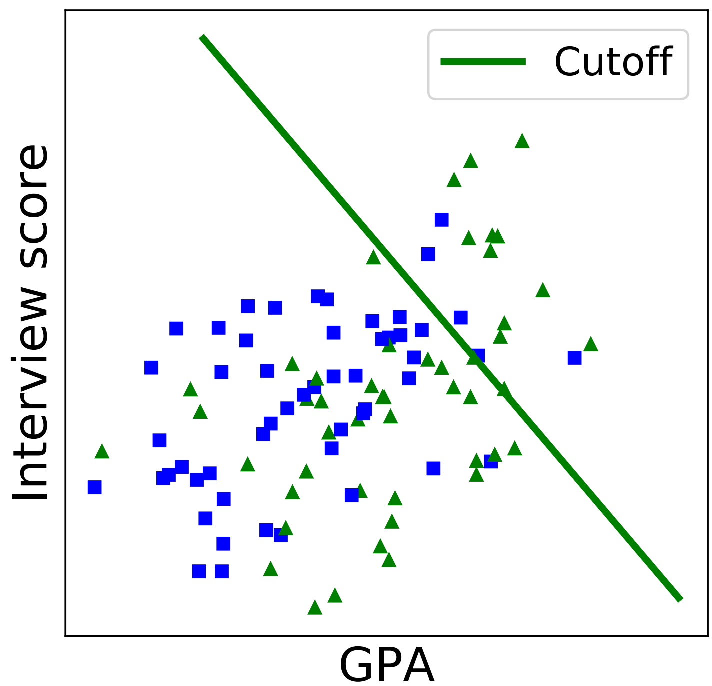


Figure 2: Toy example: a hiring classifier that predicts job performance (not shown) based on GPA and interview score, and then applies a cutoff.

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.

[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.

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]

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

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 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] David Rock and Heidi 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] Solon 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] John W. Jackson and Tyler J. 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.

[40] Clare Garvie, Alvaro Bedoya, and Jonathan 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.

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.

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] Eleanor Wiske Dillon and Jeffrey Andrew Smith, "The Determinants of Mismatch Between Students and Colleges" (National Bureau of Economic Research, 2013); Ozan Jaquette 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).

[43] Amit Datta, Michael Carl Tschantz, and Anupam Datta, "Automated Experiments on Ad Privacy Settings," *Proc. Privacy Enhancing Technologies (PET)* 2015, no. 1 (2015): 92–112.

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

[45] Eytan Bakshy, Solomon Messing, and Lada A Adamic, "Exposure to Ideologically Diverse News and Opinion on Facebook," *Science* 348, no. 6239 (2015): 1130–2.

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.

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

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.

[47] Kate 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] Dillon 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.

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

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

Bibliographic notes and further reading

For an introduction to statistical learning, we recommend the textbook by Hastie, Tibshirani, and Friedman.[50] It is available for download online. An excellent textbook by Wasserman[51] also provides much useful technical 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 FAT/ML workshop in 2014, a new community emerged, and the

[50] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning* (Springer, 2009).

[51] Larry 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] Cecilia Munoz, 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.

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

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

[57] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini, "Discrimination-Aware Data Mining," in *Proc. 14th SIGKDD* (ACM, 2008).

topic has since grown in popularity. Several popular-audience books have delivered critiques of algorithmic systems in modern society.[58]

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

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*.

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

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

We could interpret the body mass index as measuring risk of heart disease. Thresholding it at the value 27, we might decide that indi-

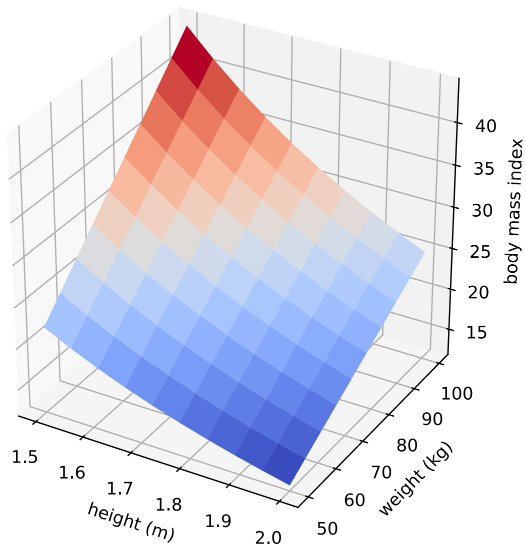


Figure 3: Plot of the body mass index.

viduals 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

From a mathematical perspective, 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.[62]

[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 $g(X)$.

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 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*).[63] Each choice of the threshold defines one binary classifier. Which threshold should we choose?

[63] 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*[64]. Note that true positive rate equals 1 - false negative rate.

[64] ROC stands for receiver operating characteristic.

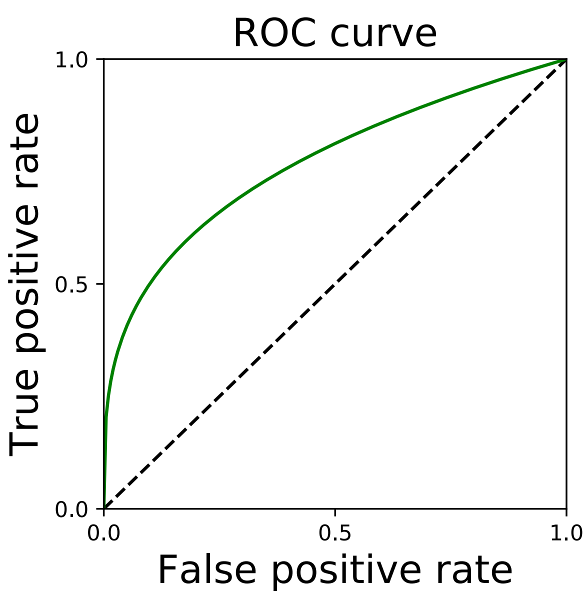


Figure 4: Example of an ROC curve. Each point on the solid curve is realized by thresholding the score function at some value. The dashed line shows the trade-offs achieved by randomly accepting an instance irrespective of its features with some probability $p \in [0, 1]$.

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[65] make *area under the curve* a tool that must be interpreted with caution.

[65] Steve Halligan, 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.

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[66]. 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.

[66] 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.[67]

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

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.

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

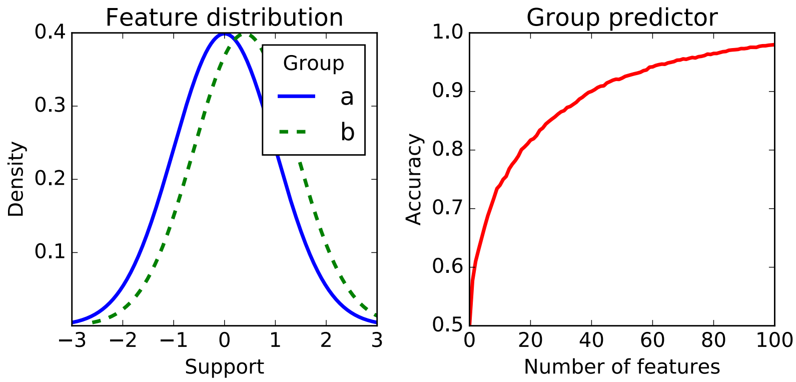


Figure 5: On the left, we see the distribution of a single feature that differs only very slightly between the two groups. In both groups the feature follows a normal distribution. Only the means are slightly different in each group. Multiple features like this can be used to build a high accuracy group membership classifier. On the right, we see how the accuracy grows as more and more features become available.

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.[68] Forcing medications to be uncorrelated with race in such cases can harm the individual.

[68] 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. Once we have acquired familiarity with the technical matter, we'll largely defer the broader debate around the purpose, scope, and meaning of these fairness criteria to 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.[69]

[69] 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.

To a first approximation, most of these criteria fall into one of three different categories defined along the lines of different (conditional) independence[70] statements between the involved random variables.

[70] 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[71] that, for $\epsilon = 0.2$, this condition relates to the *80 percent rule* in disparate impact law.

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

Yet another way to state the independence condition in full generality is to require that A and R must have zero mutual information[72] $I(A; R) = 0$. 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] 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.[73]

[73] This problem was identified and called *self-fulfilling prophecy* in, Cynthia Dwork et al., "Fairness Through Awareness," in *Proc. 3rd ITCS*, 2012, 214–26. One might object that enforcing demographic parity in this scenario might still create valuable additional training data which could then improve predictions in the future after re-training the classifier on these additional data points.

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[74].

[74] 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.

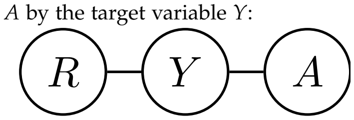
Separation

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* $R \perp A \mid Y$.[75]

[75] 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.[76]

[76] 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.[77]

[77] 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.

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 given an instance randomly picks and applies the first classifier with probabil-

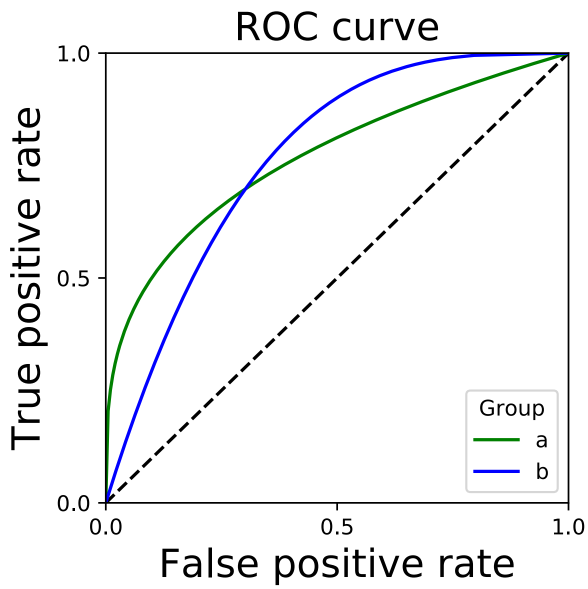


Figure 6: ROC curve by group.

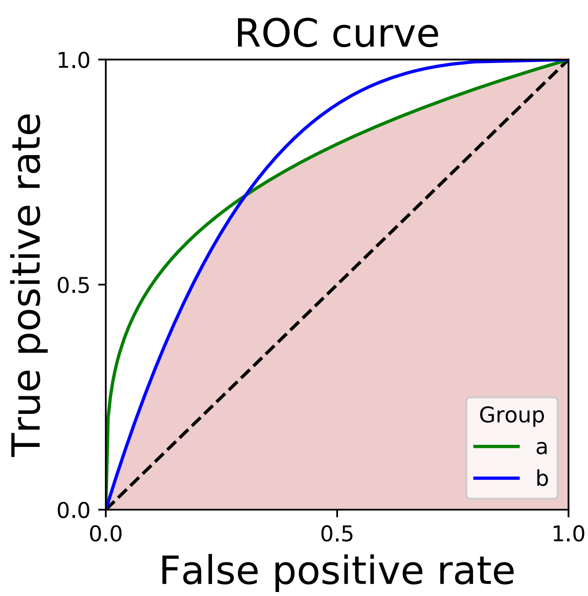


Figure 7: Intersection of area under the curves.

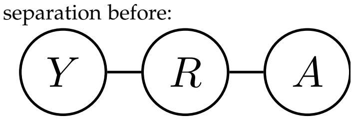
ity $1 - p$, and the second with probability p. This classifier achieves true and false positive rate p thus giving us one point on the dashed line in the plot. In the same manner, we could have picked any other pair of classifiers and randomized between them. We can fill out the entire shaded region 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* $Y \perp A \mid R$.[78]

[78] 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.

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 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 $R = \mathbb{E}[Y \mid X]$.[79]

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[80]

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

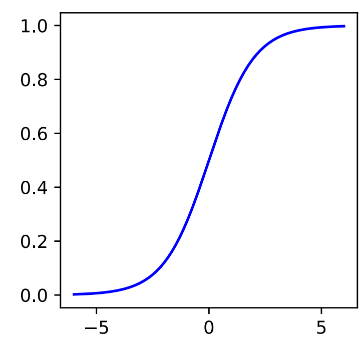
fits the target variable Y with respect to the so-called *log loss*

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

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

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

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



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.

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.

$\square$

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.[81]

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[82]. 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.

[81] Source

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

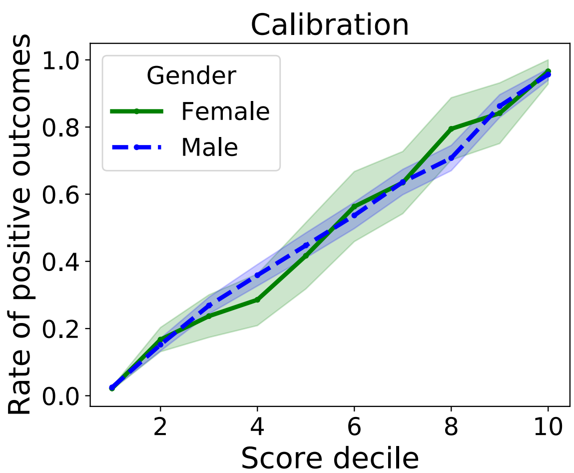


Figure 8: Calibration by gender on UCI adult data. A straight diagonal line would correspond to perfect calibration.

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 data has only 34 instances with the 'Race' attribute set to 'Black'. As a result, the error bars[83] 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

[83] The shaded region in the plot indicates a 95% confidence interval for a binomial model.

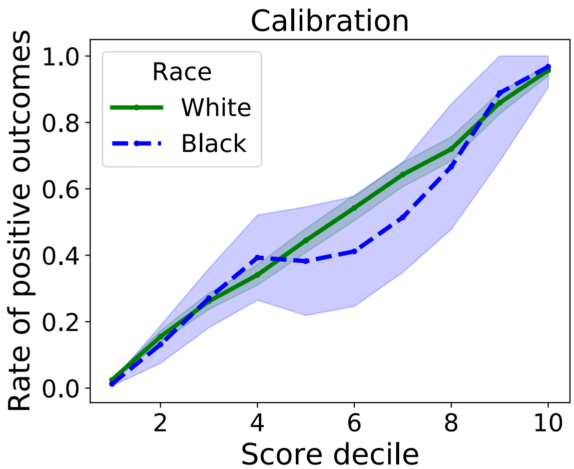


Figure 9: Calibration by race on UCI adult data.

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[84] 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.

□

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

[84] See Theorem 17.2 in Wasserman, *All of Statistics*

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 $\mathrm{FPR} > 0$, and the same positive true positive rate $\mathrm{TPR} > 0$. 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

$$\mathrm{PPV_a} = \frac{\mathrm{TPR} p_a}{\mathrm{TPR} p_a + \mathrm{FPR}(1 - p_a)} \,.$$

From the expression we can see that $\mathrm{PPV}_0 = \mathrm{PPV}_1$ only if $\mathrm{TPR} = 0$ or $\mathrm{FPR} = 0$. The latter is ruled out by assumption. So it must be that $\mathrm{TPR} = 0$. 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

$$\mathrm{NPV_a} = \frac{(1 - \mathrm{FPR})(1 - p_a)}{(1 - \mathrm{TPR}) p_a + (1 - \mathrm{FPR})(1 - p_a)} \,.$$

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.[85]

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.[86] Informally, a criterion is observational if we can express it using probability statements involving the random variables at hand.

[85] 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.

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

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

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*.[87] The main fact we need is that a node is conditionally independent of any node that is not a direct ancestor given its parents.

[87] Learn more about graphical models here.

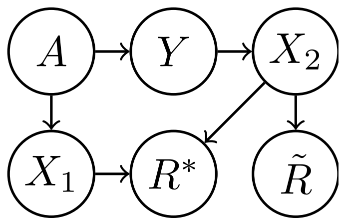


Figure 10: Directed graphical model for the variables in Scenario I

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.[88] 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`.

[88] 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.

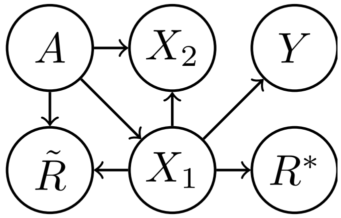


Figure 11: Directed graphical model for the variables in Scenario II

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.[89]

[89] This may not be true in reality. It's an assumption we make in this example.

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[90]. 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.[91]

[90] 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).

[91] 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.

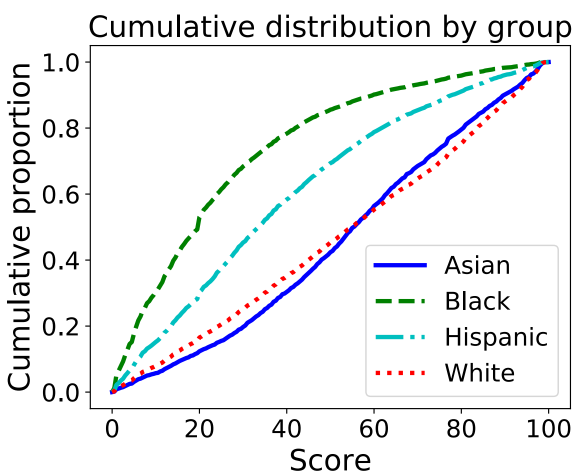


Figure 12: Cumulative density of scores by group.

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

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.[92]

[92] 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.

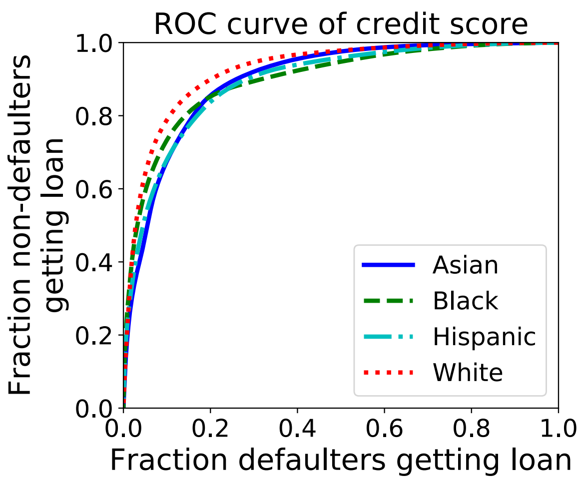


Figure 13: ROC curve of credit score by group.

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*.

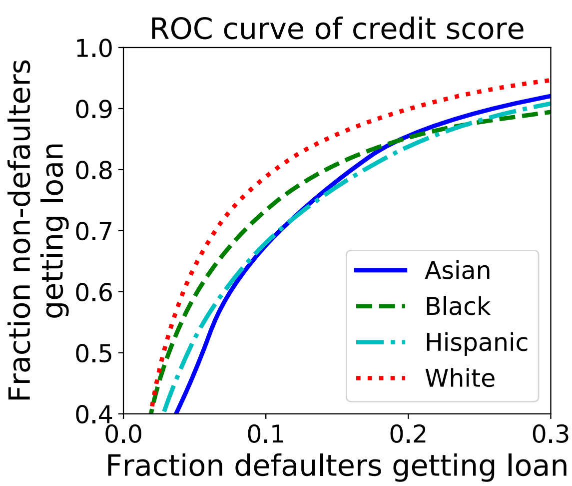


Figure 14: ROC curve of credit score by group zoomed in on region of large differences.

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 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[93], which reports that 'single threshold' achieves higher profit than 'separation', which in turn achieves higher profit than 'independence'.

[93] Moritz Hardt, Eric Price, and Nati Srebro, "Equality of Opportunity in Supervised Learning," in *Proc. 29th 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

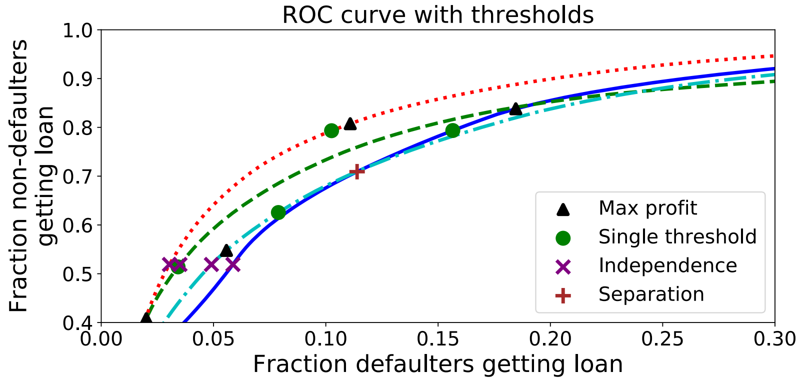


Figure 15: ROC curves with thresholds induced by different criteria.

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.[94]

[94] 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.

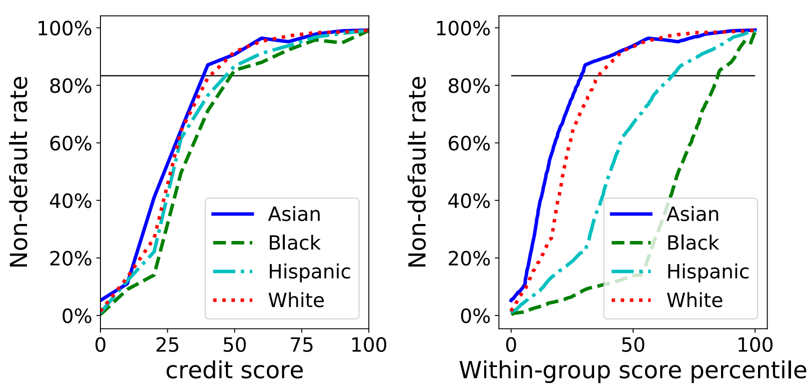


Figure 16: Calibration values of credit score by group.

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[95], we'll begin to scratch at the surface of the complex criminal justice domain. Our starting point is an investigation carried out by Propublica[96] 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.

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

[96] Julia Angwin et al., "Machine Bias," *ProPublica*, May 2016, `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentenci`

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.

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.[97]

[97] Lydia T. 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.[98] 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.

[98] Sam 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.[99] An important fairness criterion is due to Cleary[100] 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[101] considered equality of precision values, which is a relaxation of sufficiency as we saw earlier. Thorndike[102] 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[103] stated four different criteria in terms of succinct

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

[100] T Anne 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; T Anne Cleary, "Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges," *Journal of Educational Measurement* 5, no. 2 (1968): 115–24.

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

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

[103] Richard B 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[104] 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[105]. Feldman et al.[106] studied a relaxation of demographic parity in the context of disparate impact law. Zemel et al.[107] adopted the mutual information viewpoint and proposed a heuristic pre-processing approach for minimizing mutual information. Dwork et al.[108] argued that the independence criterion was inadequate as a fairness constraint.

The separation criterion appeared under the name *equalized odds*[109], alongside the relaxation to equal false negative rates, called *equality of opportunity*. These criteria also appeared in an independent work[110] under different names. Woodworth et al.[111] studied a relaxation of separation stated in terms of correlation coefficients. This relaxation corresponds to the third criterion studied by Darlington[112].

ProPublica[113] 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[114] to ProPublica's article. Similar arguments were made quickly after the publication of ProPublica's article by bloggers including Abe Gong.[115] There has been extensive scholarship on the actuarial risk assessment in criminal justice that long predates the ProPublica debate; Berk et al.[116] provide a survey with commentary.

Variants of the trade-off between separation and sufficiency were shown by Chouldechova[117] and Kleinberg et al.[118] 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[119] 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[120], except for minor changes in the choice of graphical models and their interpretation.

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

105 Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy, "Building Classifiers with Independency Constraints," in *In Proc. IEEE ICDMW*, 2009, 13–18; Faisal Kamiran and Toon Calders, "Classifying Without Discriminating," in *Proc. 2nd International Conference on Computer, Control and Communication*, 2009.

106 Feldman et al., "Certifying and Removing Disparate Impact."

107 Richard S. Zemel et al., "Learning Fair Representations," in *Proc. 30th ICML*, 2013.

108 Dwork et al., "Fairness Through Awareness."

109 Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning."

110 Muhammad Bilal Zafar et al., "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification Without Disparate Mistreatment," in *Proc. 26th WWW*, 2017.

111 Blake E. Woodworth et al., "Learning Non-Discriminatory Predictors," in *Proc. 30th COLT*, 2017, 1920–53.

112 Darlington, "Another Look at 'Cultural Fairness'."

113 Angwin et al., "Machine Bias."

114 William Dieterich, 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.

115 See this and subsequent posts.

116 Richard Berk et al., "Fairness in Criminal Justice Risk Assessments: The State of the Art," *ArXiv E-Prints* 1703.09207 (2017).

117 Alexandra Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," in *Proc. 3rd FATML*, 2016.

118 Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," *Proc. 8th ITCS*, 2017.

119 Geoff Pleiss et al., "On Fairness and Calibration," in *Proc. 30th NIPS*, 2017.

120 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)

Legal background and normative questions

Coming up soon!

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 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.[121] Historical data show that 12763 applicants were considered for admission to one of 101 departments and inter-departmental majors. Of the 4321

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

women who applied roughly 35 percent were admitted, while 44 percent of the 8442 men who applied were admitted. Standard statistical significance tests suggest hat 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 admissions for 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.[122]

[122] 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*[123], 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.

[123] For clarifications regarding the popular interpretation of Simpson's original article Edward H Simpson, "The Interpretation of Interaction in Contingency Tables," *Journal of the Royal Statistical Society: Series B (Methodological)* 13, no. 2 (1951): 238–41, see Miguel A Hernán, 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`. and Judea Pearl, *Causality* (Cambridge University Press, 2009).

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

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.*[124]

[124] 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.[125] 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.

[125] The example has been heavily discussed in various other writings, such as Pearl's recent discussion Judea Pearl and Dana 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[126] random variables, i.e., biased coin flips: $U_1 \sim \mathrm{B}(1/2), U_2 \sim \mathrm{B}(1/3), U_3 \sim \mathrm{B}(1/3)$.

[126] 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 $1 - p$.

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

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 elementary 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 to 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.

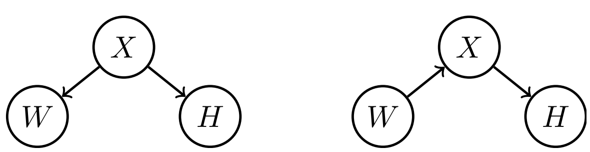


Figure 17: Causal diagrams for the heart disease examples.

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.

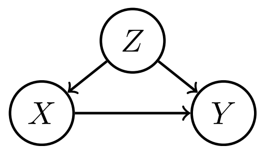


Figure 18: Example of a fork.

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.[127]

[127] Linda L. Humphrey, 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.

Mediators

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

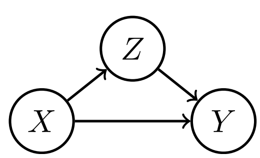


Figure 19: Example of a chain.

In this case, the path $X \rightarrow Z \rightarrow Y$ 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 discriminatino 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*.

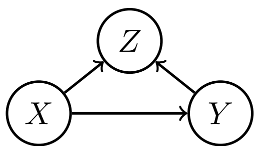


Figure 20: Example of a collider.

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.[128]

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[129] 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.

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

[129] Abigail M. AND Petrie Moneta-Koehler Liane AND Brown, "The Limitations of the Gre in Predicting Success in Biomedical Graduate School," *PLOS ONE* 12, no. 1 (January 2017): 1–17; Anna B. AND Cook Hall Joshua D. AND O'Connell, "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.

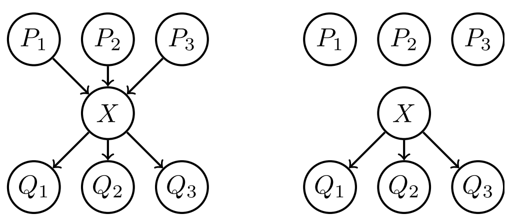


Figure 21: Graph before and after substitution.

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:

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

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 that 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 that happens when we control for the variable Z in the three causal graphs we discussed above.

- Controlling on a confounding variable Z in a fork $X \leftarrow Z \rightarrow Y$ will deconfound the effect of X on Z.
- 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[130]. 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:

[130] Pearl, *Causality*.

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

Intuitively, backdoor paths allow information flow from X to Y in a way that is not causal, i.e., does not correspond to a path with only forward edges. 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 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.[131]

[131] 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.

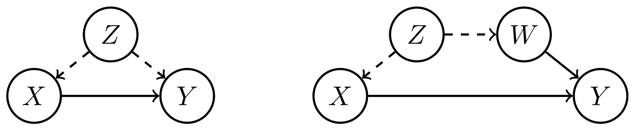


Figure 22: Two cases of unobserved confounding.

The above figure shows to 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 control 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[132].

[132] Angus Deaton and Nancy 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.[133]

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

This means we need to include either the arrow $A \to Z$ or $Z \to A$. 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 *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.

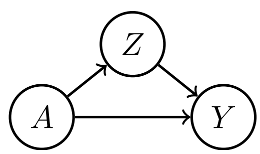


Figure 23: Possible causal graph for the UC Berkeley graduate admissions scenario.

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[134]:

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.

[134] 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

Department choice is now a collider between A and R. Conditioning on a collider opens the backdoor path $A \to Z \leftarrow R \to Y$. In

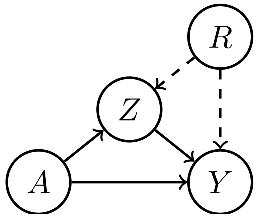


Figure 24: Alternative causal graph for the UC Berkeley graduate admissions scenario showing influence of residence.

this graph, conditioning on department choice does *not* give us the desired direct effect.[135]

[135] 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*.

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.

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.

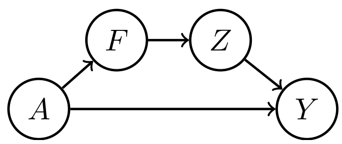


Figure 25: Alternative causal graph for the UC Berkeley graduate admissions scenario where department preferences are shaped by fear of discrimination.

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.*[136] 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). 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.

136 Griggs v. Duke Power Co., 401 U.S. 424 (1971)

Glymour[137] 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.

[137] M Maria 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 an indirect effects. Estimating an indirect effect size 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[138], such as poverty and inequality.

[138] Nancy 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.

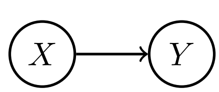


Figure 26: Causal diagram for our traffic scenario.

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:[139]

[139] 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.[140]

[140] 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 units can vary. In our traffic example, the noise variables dictates 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.[141]

[141] Jonas Peters, Dominik Janzing, and Bernhard 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.[142]

[142] 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 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 is independent of treatment given some deconfounding variables Z, i.e., $T \perp (Y_0, Y_1) \mid Z$. More generally, 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[143]. 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.

[143] Reuben M Baron 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.

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 case of a basic three variable case of a mediator, although the ideas extent to more complicated structures.

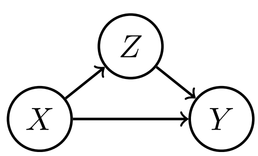


Figure 27: Causal graph with mediator Z.

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:

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

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.[144]

[144] 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):

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

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*[145], 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.

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[146]:

> 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

[145] This criterion was introduced as *counterfactual fairness*, Matt J. Kusner et al., "Counterfactual Fairness," in *Proc. 30th NIPS*, 2017, 4069–79.

[146] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell, *Causal Inference in Statistics: A Primer* (Wiley, 2016), 114.

> 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).)

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)[147] allege discrimination in Harvard undergraduate admissions against Asian-Americans.

[147] 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%.[148]

[148] 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.[149] 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 $\delta > 0$.

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.[150]

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 counterfactual. 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 it is that is required in creating valid causal models and counterfactuals. We turn to this important topic next.

[149] 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.

[150] Defendant's expert report of David Card, Professor of Economics at the University of California, Berkeley.

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.[151]

[151] 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*[152] 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.

[152] To borrow an expression from Clark 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 reliable 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:

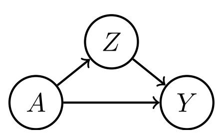


Figure 28: Religion as a root node.

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.[153] From this perspective, religious affiliation becomes an ancestor of level of education and so the graph might look like this:

[153] 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.".

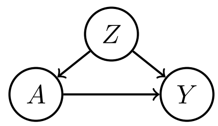


Figure 29: Religion as ancestor.

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.[154]

[154] 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.[155]

[155] Ron 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.[156]

[156] For an entry point to this topic, see for example Geoffrey C. Bowker and Susan Leigh Star, *Sorting Things Out: Classification and Its Consequences* (MIT Press, 2000); Karen E. Fields and Barbara J. Fields, *Racecraft: The Soul of Inequality in American Life* (Verso, 2014); Ruha 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*.[157]

[157] 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.[158]

[158] For a recent entry point to this literature, see Ron Mallon, *The Construction of Human Kinds* (Oxford University Press, 2018); Joshua 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.[159]

[159] For a deeper discussion of modularity in causality, see Nancy 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.[160]

[160] 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.[161]

[161] 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."[162] As such, social categories are moving targets that need constant revision.[163]

[162] Ian Hacking, "Making up People," *London Review of Books* 28, no. 16 (2006).

[163] 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 longstanding 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.[164]

[164] Paul W. 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.[165]

[165] 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."[166]

[166] 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[167], as do social scientists when reasoning about inequality in poverty, crime, and education.

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*.[168] What this boils down 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.

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.[169]

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.[170]

In opposition to Holland's view, other scholars, including Pearl, argue that causation does not require manipulability but rather an

[167] Jackson and VanderWeele, "Decomposition Analysis to Identify Intervention Targets for Reducing Disparities"; Tyler J. VanderWeele and Whitney R. Robinson, "On Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables," *Epidemiology*, 2014.

[168] D. James Greiner and Donald B. Rubin, "Causal Effects of Perceived Immutable Characteristics," *The Review of Economics and Statistics* 93, no. 3 (2011): 775–85.

[169] For a deeper discussion of this point and an insightful critique of counterfactual arguments about race more broadly, see Issa Kohler-Hausmann, "Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination," *SSRN*, 2019.

[170] Note the term *racecraft* coined in Fields and Fields, *Racecraft*.

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 and scope of the model.[171] We will return to questions of validity in our next chapter on measurement that provides a helpful complementary perspective.

[171] On this point of purpose and scope, see Cartwright's essay in Isabelle F. Peschard and Bas C. 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[172],

[172] Pearl, Glymour, and Jewell, *Causal Inference in Statistics*.

or the more comprehensive text[173]. 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[174]. All of these texts touch on the topic of discrimination just cited. In these books, Pearl takes the position that discrimination corresponds to the direct effect of the sensitive category on a decision. The technically-minded reader will enjoy complementing Pearl's book with the recent open access text by Peters, Janzing, and Scholkopf[175] that is available online. The text emphasizes two variable causal models and applications to machine learning. See[176] for a general introduction based on causal graphs with an emphasis on *graph discovery*, i.e., inferring causal graphs from observational data.

[173] Pearl, *Causality*.

[174] Pearl and Mackenzie, *The Book of Why*.

[175] Peters, Janzing, and Schölkopf, *Elements of Causal Inference*.

[176] Peter Spirtes et al., *Causation, Prediction, and Search* (MIT press, 2000).

Morgan and Winship[177] focus on applications in the social sciences. Imbens and Rubin[178] give a comprehensive overview of the technical repertoire of causal inference in the potential outcomes model. Angrist and Pischke[179] focus on causal inference and potential outcomes in econometrics.

Hernan and Robins[180] give another detailed introduction to causal inference that draws on the authors' experience in epidemiology.

Pearl[181] 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.

[177] Stephen L. Morgan and Christopher Winship, *Counterfactuals and Causal Inference* (Cambridge University Press, 2014).

[178] Guido W. Imbens and Donald B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences* (Cambridge University Press, 2015).

[179] Joshua D. Angrist and Pischke Jörn-Steffen, *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton University Press, 2009).

[180] Miguel Hernán and James Robins, *Causal Inference* (Boca Raton: Chapman & Hall/CRC, forthcoming, 2019).

[181] 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.[182] previously also considered discrimination analysis via path-specific causal effects. Kusner et al.[183] introduced a notion of *counterfactual fairness*. The authors extend this line of thought in another work[184]. Chiappa introduces a path-specific notion of counterfactual fairness[185].

Kilbertus et al.[186] 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[187] conceptualize discrimination as the influence of the sensitive attribute on the outcome along certain *disallowed* causal paths. Chiappa and Isaac[188] give a tutorial on causality and fairness with an emphasis on the COMPAS debate.

[182] Lu Zhang, Yongkai Wu, and Xintao Wu, "A Causal Framework for Discovering and Removing Direct and Indirect Discrimination," in *Proc.* 26*th IJCAI*, 2017, 3929–35.

[183] Kusner et al., "Counterfactual Fairness."

[184] Chris Russell et al., "When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness," in *Proc.* 30*th NIPS*, 2017, 6417–26.

[185] Silvia Chiappa, "Path-Specific Counterfactual Fairness," in *Proc.* 33*rd Aaai*, vol. 33, 2019, 7801–8.

[186] Niki Kilbertus et al., "Avoiding Discrimination Through Causal Reasoning," in *Proc.* 30*th NIPS*, 2017, 656–66.

[187] Razieh Nabi and Ilya Shpitser, "Fair Inference on Outcomes," in *Proc.* 32*nd AAAI*, 2018, 1931–40.

[188] Silvia Chiappa and William S. Isaac, "A Causal Bayesian Networks Viewpoint on Fairness," *Arxiv.org* arXiv:1907.06430 (2019).

Philosophical debate

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[189] is a good starting point in this direction.

[189] Cartwright, *Hunting Causes and Using Them, Too*.

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 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 naturally 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 a 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. This is traditionally the focus of control theory, an area with a long history and vast technical repertoire.[190]

While much of control theory focuses on physical systems, concepts from control theory of also influenced policy and decision making in other domains. A well-known example is the area of *system dynamics* pioneered by Forrester[191] in the 60s and 70s that lead to some politically influential works such as *Limits to Growth*[192]. But see Baker's thesis[193] for a history of system modeling and its pitfalls.

[190] Karl Johan Aström and Richard M Murray, *Feedback Systems: An Introduction for Scientists and Engineers* (Princeton university press, 2010).

[191] Jay W Forrester, "Urban Dynamics," *IMR; Industrial Management Review (Pre-1986)* 11, no. 3 (1970): 67; Jay W Forrester, "Counterintuitive Behavior of Social Systems," *Technological Forecasting and Social Change* 3 (1971): 1–22; Jay W Forrester, "System Dynamics, Systems Thinking, and Soft or," *System Dynamics Review* 10, nos. 2-3 (1994): 245–56.

[192] Dennis Meadows and Jorgan Randers, *The Limits to Growth: The 30-Year Update* (Routledge, 2012).

[193] Kevin T. Baker, "World Processors: Computer Modeling, Global Environmentalism, and the Birth of Sustainable Development" (Northwestern University, 2019).

Bibliography

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." *ProPublica*, May 2016. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

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.

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): 1130–2.

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 Andrew D. Selbst. "Big Data's Disparate Impact." *California Law Review* 104 (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.

Benjamin, Ruha. *Race After Technology*. Polity, 2019.

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.

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.

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.

Bowker, Geoffrey C., and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. MIT Press, 2000.

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.

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.

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).

Chouldechova, Alexandra. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." In *Proc. 3rd FATML*, 2016.

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.

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.

Crawford, Kate. "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.

Danesi, Marcel. *Dictionary of Media and Communications*. Routledge, 2014.

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

Datta, Amit, Michael Carl Tschantz, and Anupam Datta. "Automated Experiments on Ad Privacy Settings." *Proc. Privacy Enhancing Technologies (PET)* 2015, no. 1 (2015): 92–112.

Dawes, Robyn M, David Faust, and Paul E Meehl. "Clinical Versus Actuarial Judgment." *Science* 243, no. 4899 (1989): 1668–74.

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

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.

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.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness Through Awareness." In *Proc. 3rd ITCS*, 214–26, 2012.

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.

Eubanks, Virginia. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018.

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. "Certifying and Removing Disparate Impact." In *Proc. 21st SIGKDD*. ACM, 2015.

Fields, Karen E., and Barbara J. Fields. *Racecraft: The Soul of Inequality in American Life*. Verso, 2014.

Forrester, Jay W. "Counterintuitive Behavior of Social Systems." *Technological Forecasting and Social Change* 3 (1971): 1–22.

———. "System Dynamics, Systems Thinking, and Soft or." *Sys-*

tem Dynamics Review 10, nos. 2-3 (1994): 245–56.

———. "Urban Dynamics." *IMR; Industrial Management Review (Pre-1986)* 11, no. 3 (1970): 67.

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

Garvie, Clare, Alvaro Bedoya, and Jonathan Frankle. "The Perpetual Line-up." *Georgetown Law: Center on Privacy and Technology.*, 2016.

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.

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.

Hacking, Ian. "Making up People." *London Review of Books* 28, no. 16 (2006).

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.

Hanna, Rema N, and Leigh L Linden. "Discrimination in Grading." *American Economic Journal: Economic Policy* 4, no. 4 (2012): 146–68.

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.* 29*th NIPS*, 3315–23, 2016.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.

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.

Holland, Paul W. "Statistics and Causal Inference." *Journal of the*

American Statistical Association (JASA) 81 (1986): 945–70.

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.

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.

Jackson, John W., and Tyler J. VanderWeele. "Decomposition Analysis to Identify Intervention Targets for Reducing Disparities." *Epidemiology*, 2018, 825–35.

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.

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.

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

Kalantari, Nima Khademi, and Ravi Ramamoorthi. "Deep High Dynamic Range Imaging of Dynamic Scenes." *ACM Trans. Graph* 36, no. 4 (2017): 144.

Kamiran, Faisal, and Toon Calders. "Classifying Without Discriminating." In *Proc. 2nd International Conference on Computer, Control and Communication*, 2009.

Kaufman, Liad, Dani Lischinski, and Michael Werman. "Content-Aware Automatic Photo Enhancement." In *Computer Graphics Forum*, 31:2528–40. 8. Wiley Online Library, 2012.

Kay, Matthew, Cynthia Matuszek, and Sean A Munson. "Unequal Representation and Gender Stereotypes in Image Search Results for Occupations." In *Proc. 33rd Conference on Human Factors in Computing Systems*, 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.

Kleinberg, Jon M., Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *Proc. 8th ITCS*, 2017.

Kohler-Hausmann, Issa. "Eddie Murphy and the Dangers of

Counterfactual Causal Thinking About Detecting Racial Discrimination." *SSRN*, 2019.

Krieger, Nancy. "Epidemiology and the People's Health: Theory and Context," 2011.

Kusner, Matt J., Joshua R. Loftus, Chris Russell, and Ricardo Silva. "Counterfactual Fairness." In *Proc.* 30*th NIPS*, 4069–79, 2017.

Lewis, Mary A. "A Comparison of Three Models for Determining Test Fairness." Federal Aviation Administration Washington DC Office of Aviation Medicine, 1978.

Liu, Lydia T., Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. "Delayed Impact of Fair Machine Learning." In *Proc.* 35*th 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*, 1035–8. IEEE, 2007.

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

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.

Meadows, Dennis, and Jorgan Randers. *The Limits to Growth: The 30-Year Update*. Routledge, 2012.

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

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.

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.

Nabi, Razieh, and Ilya Shpitser. "Fair Inference on Outcomes." In *Proc.* 32*nd AAAI*, 1931–40, 2018.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, 2016.

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.

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.

Pedreshi, Dino, Salvatore Ruggieri, and Franco Turini. "Discrimination-Aware Data Mining." In *Proc.* 14*th SIGKDD*. ACM, 2008.

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.

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.* 30*th NIPS*, 2017.

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.

Rock, David, and Heidi Grant. "Why Diverse Teams Are Smarter." Harvard Business Review. https://hbr.org/2016/11/why-diverse-teams-are-smarter, 2016.

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

Russell, Chris, Matt J. Kusner, Joshua R. Loftus, and Ricardo Silva. "When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness." In *Proc.* 30*th NIPS*, 6417–26, 2017.

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.

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.

Sweeney, Latanya. "Discrimination in Online Ad Delivery." *Queue*

11, no. 3 (March 2013): 10:10–10:29.

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.

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

Torralba, Antonio, and Alexei A Efros. "Unbiased Look at Dataset Bias." In *Proc. CVPR*, 1521–8. IEEE, 2011.

VanderWeele, Tyler J., and Whitney R. Robinson. "On Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables." *Epidemiology*, 2014.

Wasserman, Larry. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2010.

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.

Woodworth, Blake E., Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. "Learning Non-Discriminatory Predictors." In *Proc. 30th COLT*, 1920–53, 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