Fairness and Machine Learning

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

Incomplete working draft

Created: Tue Jul 17 19:07:07 PDT 2018
Latest version available at http://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 grateful to our students for their active participation in pilot courses at Berkeley, Cornell, and Princeton. Thanks in particular to Claudia Roberts for lecture notes of the Princeton course.

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

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

So far, we received substantial comments and feedback on draft chapters from Michaela Hardt, and Zachary Lipton. We're greatful to Ben Hutchinson for pointing us to relevant early work on demographic fairness criteria.

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 it. 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.

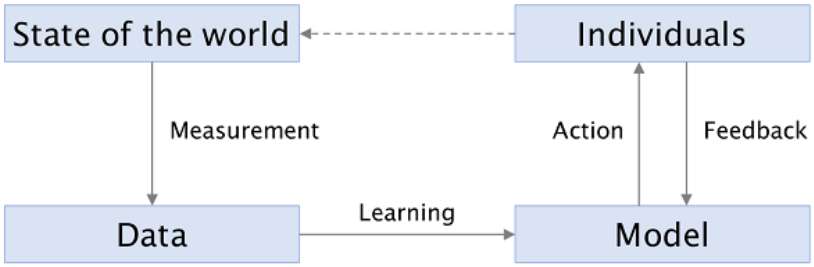


Figure 1: The machine learning loop

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 generated 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 pat-

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

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

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

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

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

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

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

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

[9] 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.

The trouble with measurement

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

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

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

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

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

[10] 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).

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

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

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

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

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

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

[12] Barocas and Selbst.

[13] 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).

[14] J. Deng et al., "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 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.

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 samples of an underlying visual world, we would expect that a computer vision system trained on one such dataset would do well on another dataset. But in reality, we observe a big drop in accuracy when we train and test on different datasets.[17] This shows that these datasets are biased relative to each other in a statistical sense, and is a good starting point for investigating whether these biases include cultural stereotypes.

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

[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.

[17] Antonio Torralba and Alexei A Efros, "Unbiased Look at Dataset Bias," in *Computer Vision and Pattern Recognition (Cvpr), 2011 Ieee Conference on* (IEEE, 2011), 1521–8.

[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 F Wilson et al., "Population Genetic

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

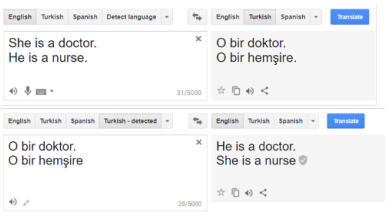
We've seen that training data reflects the disparities, distortions, and biases from the real world and the measurement process. This leads to an obvious question: when we learn a model from such data, are these disparities preserved, mitigated, or exacerbated?

Statistical methods, including machine learning, are good at calibration: ensuring that the distribution of outputs matches the prior probabilities. 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 of the training set (which is typically a large, minimally curated corpus of historical text and text found on the web).

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



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 fact that a model performs terribly for a minority group. Reporting accuracy rates by group will help alert us to problems like the above

[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.

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 *Proceedings of the 21th Acm Sigkdd International Conference on Knowledge Discovery and Data Mining* (ACM, 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 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 *Proceedings of the Tenth Acm 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, https://doi.org/10.1145/2460276.2460278.

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 can be 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 PredPol, one of the few to be published in a peer-reviewed journal.[32] By applying it to data derived from Oakland police records, they

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

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.

[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.

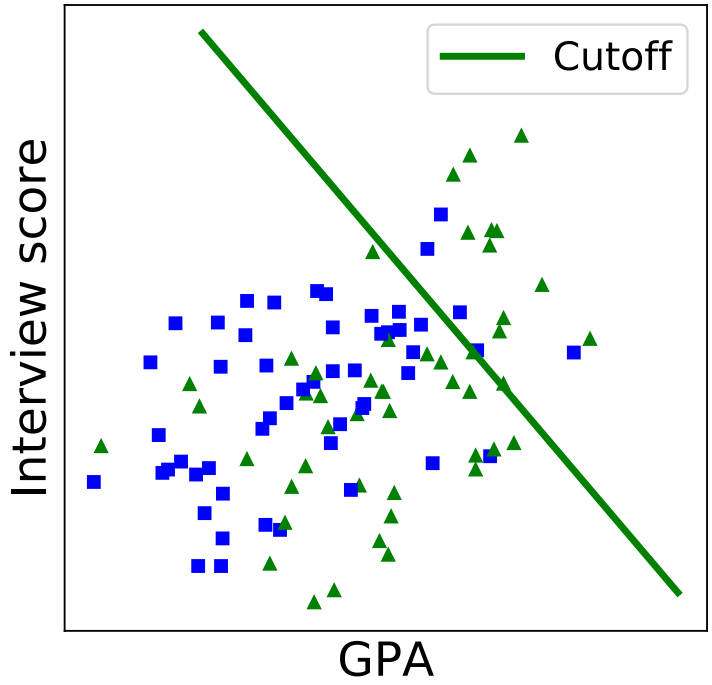


Figure 2: Toy example: a hiring classifier that predicts job performance (not shown) based on GPA and interview score, and then applies a cutoff.

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.

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.

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

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

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.[36] From this perspective, increasing the diversity of the cohort that is hired would benefit the firm and everyone in the cohort.

[36] 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.[37]

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.

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

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

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're 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?[38,39] What can go wrong with the use of DNA testing as a forensic tool? Understanding the error rate as well as the nature of errors of these technologies is critical to an informed debate.

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

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.

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

[39] 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.

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.[40] 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.[41]

A related concern arises in personalization systems: the infamous filter bubble.[42] 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[43]—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.

[40] 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).

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

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

[43] 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.[44] Should we care about such disparities in image search results? After all, these results don't affect hiring or any other consequential decisions. And what are the harms from gender stereotypes in online translation? These and other examples that are disturbing to varying degrees—such as Google's app labeling photos of black Americans as "gorillas", or offensive results in autocomplete—seem to fall into a different moral category than, say, a discriminatory system used in criminal justice, which has immediate and tangible consequences.

A recent talk lays out the differences.[45] 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.

Allocative harms have received much attention both because their

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

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

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.[46]

Second, effective interventions do exist in many machine learning

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

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,[47] expanding the tools at our disposal to reason about fairness and seek out justice.

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

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

Bibliographic notes and further reading

For an introduction to statistical learning, we recommend the textbook by Hastie, Tibshirani, and Friedman.[48] It is available for download online. An excellent textbook by Wasserman[49] 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,[50] a paper by Barocas and Selbst,[51] and a 2016 report by the White House Office of Science and Technology Policy.[52] For a broad survey of challenges raised by AI, machine learning, and algorithmic systems, see the AI Now report.[53]

An early work that investigated fairness in algorithmic systems is by Friedman and Nissenbaum in 1996.[54] Papers studying demographic disparities in classification began appearing regularly starting in 2008;[55] 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

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

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

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

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

[52] 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.

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

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

[55] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini, "Discrimination-Aware Data Mining," in *Proc. 14th Acm Sigkdd*, 2008.

topic has since grown in popularity. Several popular-audience books have delivered critiques of algorithmic systems in modern society.[56]

[56] 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).

Demographic classification criteria

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.[57] We can interpret such a classifier as a *decision rule* by equating a positive label with *acceptance* and a negative label with *rejection*.

[57] 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.[58]

[58] 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.

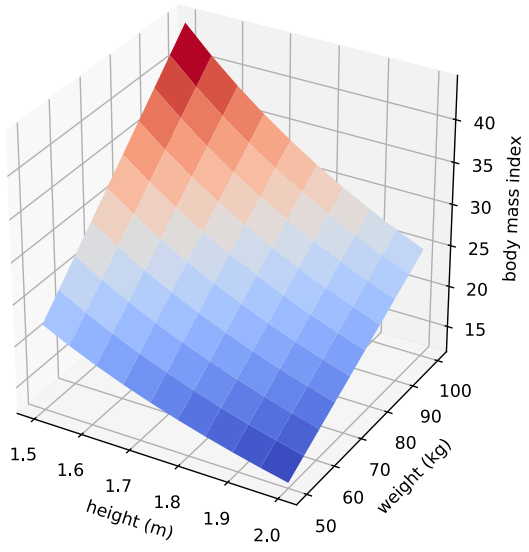
Score functions

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* or *regressor*. 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[59]. 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).

[59] 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.[60]

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

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

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

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

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 and therefore more conservative 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*[62].

[62] ROC stands for receiver operating characteristic.

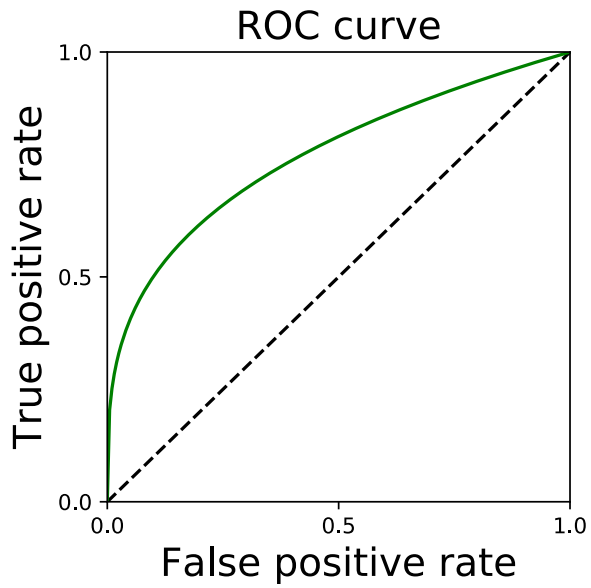


Figure 4: Example of an ROC curve. Each point on the solid curve is realized by thresholding the score function at some value. The dashed line shows the trade-offs achieved by randomly accepting an instance irrespective of its features with some probability $p \in [0, 1]$.

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

[63] 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[64]. 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.

[64] 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 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.[65]

[65] 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 slightly correlated features can be used to build high accuracy classifiers for the sensitive 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 ancestry, which in turn correlates with income in some countries

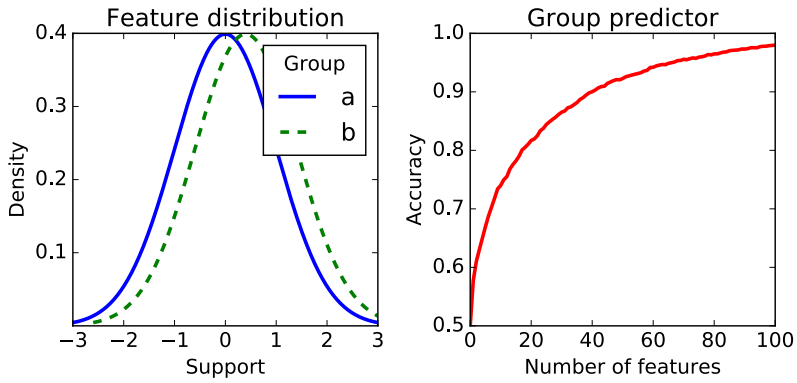


Figure 5: On the left, we see the distribution of a single feature that differs only very slightly between the two groups. In both groups the feature follows a normal distribution. Only the means are slightly different in each group. Multiple features like this can be used to build a high accuracy group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

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

[66] 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.[67]

To a first approximation, most of these criteria fall into one of three different categories defined along the lines of different (conditional)

[67] 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.

independence[68] statements between the involved random variables.

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.

[68] Learn more about conditional independence 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 variants and relaxations, 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[69] that, for $\epsilon = 0.2$, this condition relates to the *80 percent rule* in disparate impact law.

Yet another way to state the independence condition in full generality is to require that A and R must have zero mutual information[70] $I(A; R) = 0$. The characterization in terms of mutual information leads to useful relaxations of the constraint. For example, we could require $I(A; R) \leq \epsilon$.

[69] Michael Feldman et al., "Certifying and Removing Disparate Impact," in *Proc.* 21*th ACM SIGKDD*, 2015.

[70] 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.[71]

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.

[71] 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.

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[72].

[72] 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.

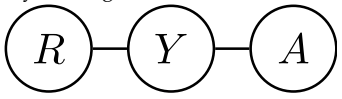
Separation

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* $R \perp A \mid Y$.[73]

[73] 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.[74]

[74] 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.

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

Achieving separation

As was the case with independence, we can achieve separation by post-processing a given score function without the need for retraining.[75]

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-

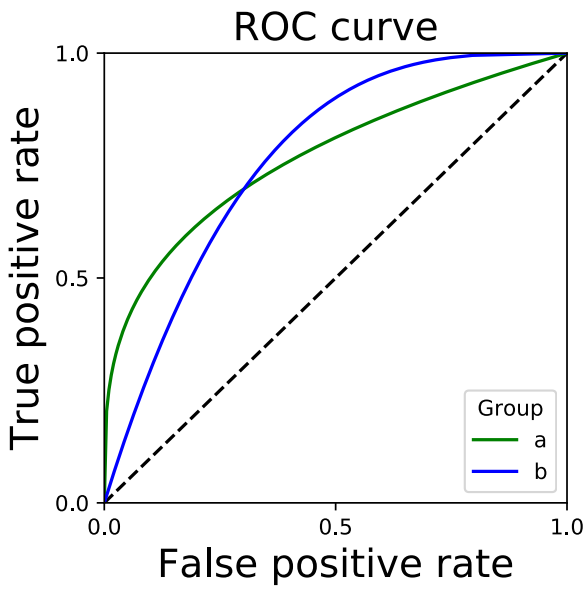


Figure 6: ROC curve by group.

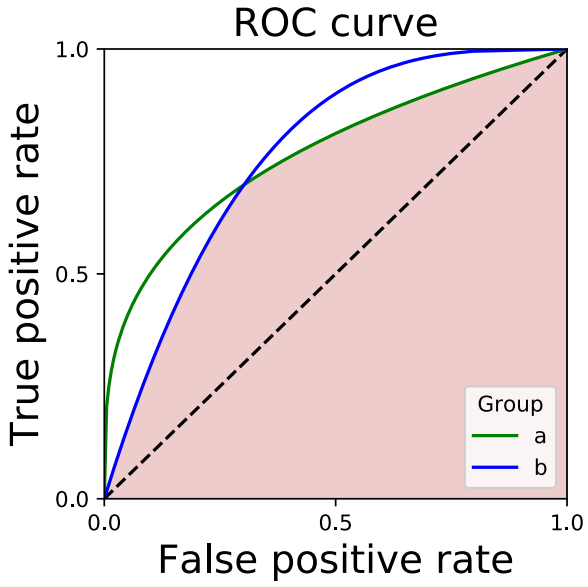


Figure 7: Intersection of area under the curves.

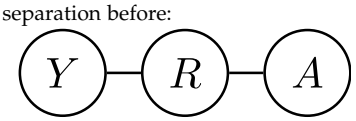
ity $1 - p$, and the second with probability p. This classifier achieves true and false positive rate p thus giving us one point on the dashed line in the plot. In the same manner, we could have picked any other pair of classifiers and randomized between them. We can fill out the entire shaded region 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* $Y \perp A \mid R$.[76]

[76] 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 $R = \mathbb{E}[Y \mid X]$.[77]

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[78]

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

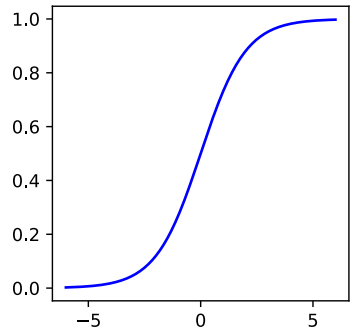
fits the target variable Y with respect to the so-called *log loss*

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

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

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

[78] 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.

□

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.

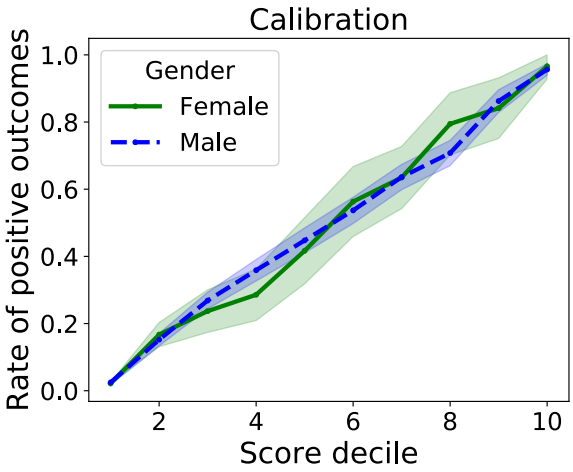
Calibration as a consequence of unconstrained learning

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.[79]

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[80]. We make no effort to either tune or calibrate the model.

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



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

The deviation we see in the mid deciles may be due to the scarcity of the test data in the corresponding group and deciles. For example, the 6th decile, corresponding to the score range $(0.5, 0.6]$, on the test data has only 34 instances with the 'Race' attribute set to 'Black'. As a result, the error bars[81] 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

[79] Source

[80] 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.

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

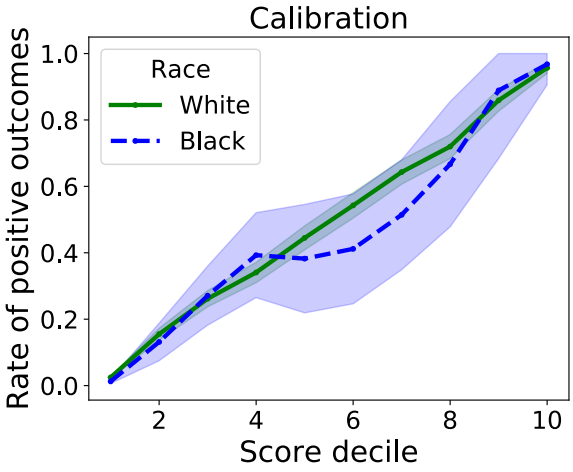


Figure 9: Calibration by race on UCI adult data.

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 satified 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[82] 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.

□

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

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

In case the classifier is also binary, we can weaken the assumption to require only that the classifier is imperfect in the sense of making

at least one false positive prediction. What's appealing about the resulting claim is that its proof essentially only uses a well-known relationship between true positive rate (recall) and positive predictive value (precision). This trade-off is often called *precision-recall trade-off*.

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

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

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

Now suppose that separation holds. Since the classifier is imperfect this means that all groups have the same non-zero false positive rate $\mathrm{FPR} > 0$, 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.[83]

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

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

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

[83] 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.

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

- Y indicates whether the user is a software engineer

We can summarize the conditional independence relationships between the variables in a *directed graphical model*.[85] The main fact we need is that a node is conditionally independent of any node that is not a direct ancestor given its parents.

[85] Learn more about graphical models here.

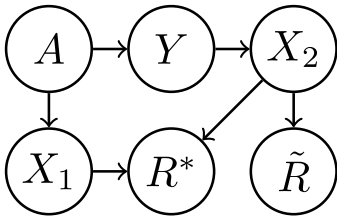


Figure 10: Directed graphical model for the variables in Scenario I

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.[86] 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`.

[86] 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.

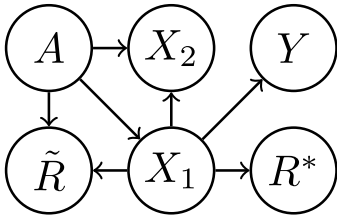


Figure 11: Directed graphical model for the variables in Scenario II

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.[87]

[87] 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 secenario 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[88]. 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.[89]

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

[88] 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).

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

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.

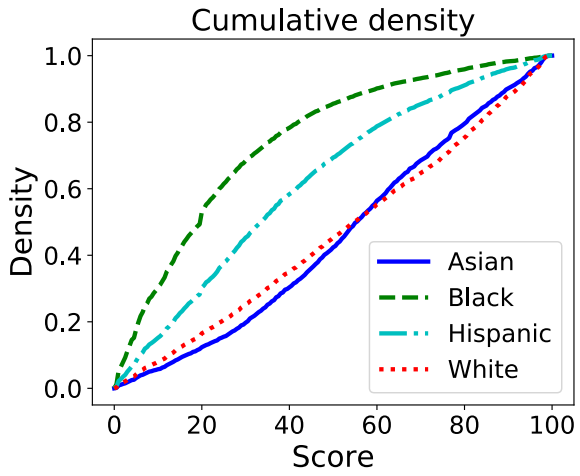


Figure 12: Cumulative density of scores by group.

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.[90]

[90] 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.

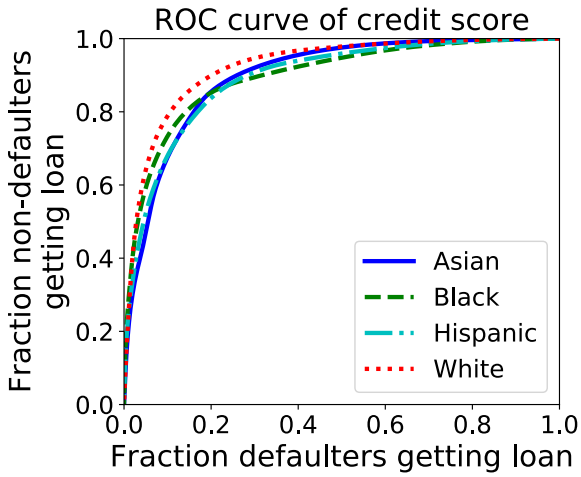


Figure 13: ROC curve of credit score by group.

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

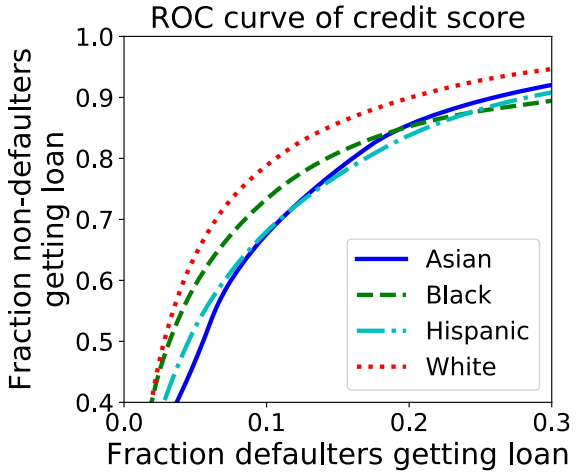


Figure 14: ROC curve of credit score by group zoomed in on region of large differences.

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

[91] Moritz Hardt, Eric Price, and Nati Srebro, "Equality of Opportunity in Supervised Learning," in *Proc. 29th NIPS*, 2016.

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 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.

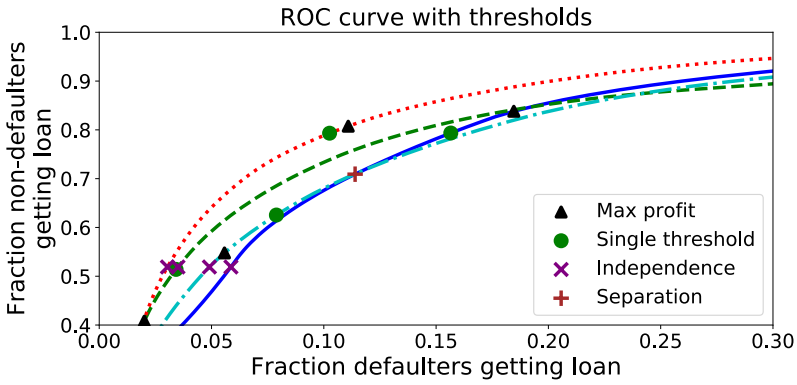


Figure 15: ROC curves with thresholds induced by different criteria.

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.[92]

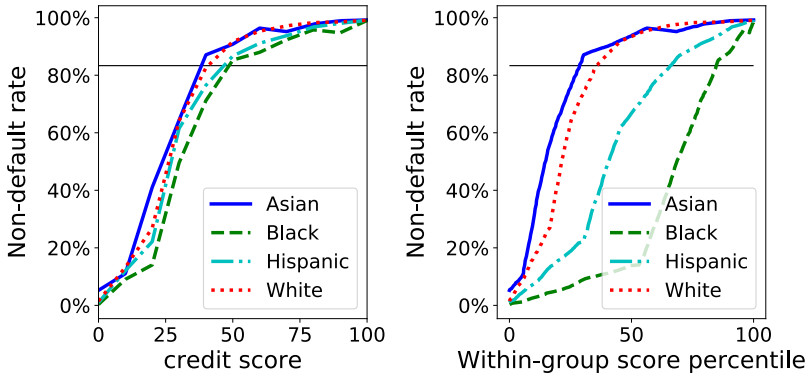


Figure 16: Calibration values of credit score by group.

[92] 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.

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[93], we'll begin to scratch at the surface of the complex criminal justice domain. Our starting point is an investigation carried out by Propublica[94] of a proprietary risk score, called COMPAS score. These scores are intended to assess the risk that a defendant will re-offend, a task often called *recidivism prediction*. Within the academic community, the ProPublica article drew much attention to the trade-off between separation and sufficiency that we saw earlier.

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

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

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

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.

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.[95]

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 dif-

[95] Lydia T. Liu et al., "Delayed Impact of Fair Machine Learning," *arXiv* abs:1803.04383 (2018).

ferent 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.[96] 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.

[96] 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.[97] An important fairness criterion is due to Cleary[98] 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[99] considered equality of precision values, which is a relaxation of sufficiency as we saw earlier. Thorndike[100] 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[101] stated four different criteria in terms of succinct 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[102] reviewed three fairness criteria including equal precision and equal true/false positive rates.

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

[98] 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.

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

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

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

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

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

The separation criterion appeared under the name *equalized odds*[107], alongside the relaxation to equal false negative rates, called *equality of opportunity.* These criteria also appeared in an independent work[108] under different names. Woodworth et al.[109] studied a relaxation of separation stated in terms of correlation coefficients. This relaxation corresponds to the third criterion studied by.[110]

ProPublica[111] 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 to ProPublica's article. Similar arguments were made quickly after the publication of ProPublica's article by bloggers including Abe Gong.[112] There has been extensive scholarship on the actuarial risk assessment in criminal justice that long predates the ProPublica debate; Berk et al.[113] provide a survey with commentary.

Variants of the trade-off between separation and sufficiency were shown by Chouldechova[114] and Kleinberg et al.[115] 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[116] 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[117], except for minor changes in the choice of graphical models and their interpretation.

[103] 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 *In Proc. 2Nd International Conference on Computer, Control and Communication*, 2009.

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

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

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

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

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

[109] Blake E. Woodworth et al., "Learning Non-Discriminatory Predictors," in *Proc. 30th Colt*, 2017, 1920–53.

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

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

[112] See this and subsequent posts.

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

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

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

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

[117] 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.

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)
Overall accuracy equality	Independence	Relaxation	Berk 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)

Bibliography

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.

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 PeÃśa 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).

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).

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.

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.

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 *Proceedings of the 21th Acm Sigkdd International Conference on Knowledge Discovery and*

Data Mining, 1721–30. ACM, 2015.

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." *Proceedings on Privacy Enhancing Technologies* 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.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "ImageNet: A Large-Scale Hierarchical Image Database." In *CVPR09*, 2009.

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. 21th ACM SIGKDD*, 2015.

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.

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. 29th NIPS*, 2016.

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

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.

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 *Proceedings of the Tenth Acm 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 *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 *Proceedings of the 33rd Annual Acm Conference on Human Factors in Computing Systems*, 3819–28. ACM, 2015.

Kleinberg, Jon M., Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *Proc. 8th ITCS*, 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." *arXiv* abs:1803.04383 (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.

Manthorpe, Rowland. "The Beauty.ai Robot Beauty Contest Is Back." Wired UK. https://www.wired.co.uk/article/robot-beauty-contest-beauty-ai, 2017.

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

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.

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.

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

Plaugic, Lizzie. "FaceApp's Creator Apologizes for the App's Skin-Lightening 'Hot' Filter." The Verge. https://www.theverge.com/2017/4/25/15419522/faceapp-hot-filter-racist-apology,

2017.

Pleiss, Geoff, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. "On Fairness and Calibration." In *Proc. 30th NIPS*, 2017.

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.

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. https://doi.org/10.1145/2460276.2460278.

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 *Computer Vision and Pattern Recognition (Cvpr), 2011 Ieee Conference on*, 1521–8. IEEE, 2011.

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 GÃşmez 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.

List of Figures